A Corpus of Preposition Supersenses

Nathan Schneider
University of Edinburgh /
Georgetown University
nschneid@inf.ed.ac.uk

Jena D. Hwang
IHMC
jhwang@ihmc.us

Vivek Srikumar
University of Utah
svivek@cs.utah.edu

Meredith Green **Abhijit Suresh** **Kathryn Conger** **Tim O'Gorman** **Martha Palmer**
University of Colorado at Boulder
{laura.green,abhijit.suresh,kathryn.conger,timothy.ogorman,martha.palmer}@colorado.edu

Abstract

We present the first corpus annotated with **preposition supersenses**, unlexicalized categories for semantic functions that can be marked by English prepositions (Schneider et al., 2015). The preposition supersenses are organized hierarchically and designed to facilitate comprehensive manual annotation. Our dataset is publicly released on the web.[1]

1 Introduction

English prepositions exhibit stunning frequency and wicked polysemy. In the 450M-word COCA corpus (Davies, 2010), 11 prepositions are more frequent than the most frequent noun.[2] In the corpus presented in this paper, prepositions account for 8.5% of tokens (the top 11 prepositions comprise >6% of all tokens). Far from being vacuous grammatical formalities, prepositions serve as essential linkers of meaning, and the few extremely frequent ones are exploited for many different functions (figure 1). For all their importance, however, prepositions have received relatively little attention in computational semantics, and the community has not yet arrived at a comprehensive and reliable scheme for annotating the semantics of prepositions in context (§2). We believe that such annotation of preposition functions is needed if preposition sense disambiguation systems are to be useful for downstream tasks—e.g., translation[3] or semantic parsing (cf. Dahlmeier et al., 2009; Srikumar and Roth, 2011).

This paper describes a new corpus, fully annotated with **preposition supersenses** (hierarchically organized unlexicalized classes primarily reflecting thematic roles; Schneider et al., 2015). Whereas fine-grained sense annotation for individual prepositions is difficult and limited by the coverage and quality of a lexicon, preposition supersense annotation offers a practical alternative (§2). We comprehensively annotate English preposition tokens in a corpus of web reviews (§3). It is the first English corpus with semantic annotations of prepositions that are both *comprehensive* (describing all preposition types and tokens) and *double-annotated* (to attenuate subjectivity in the annotation scheme and measure inter-annotator agreement). The corpus gives us an empirical distribution of preposition supersenses, and the annotation process has helped us improve upon the supersense hierarchy. Additionally, we examine the correspondences between our annotations and role labels from PropBank (§4). For some labels, clean correspondences between the two independent annotations speak to the validity of our hierarchy and annotation, but this analysis also reveals mismatches deserving of further examination. The corpus is publicly released (footnote 1).

(1) I have been going **to**/DESTINATION the Wildwood_,_NJ **for**/DURATION over 30 years **for**/PURPOSE summer~vacations

(2) It is close **to**/LOCATION bus_lines **for**/DESTINATION Opera_Plaza

(3) I was looking~**to**/`i bring a customer **to**/DESTINATION their lot **to**/PURPOSE buy a car

Figure 1: Preposition supersenses illustrating the polysemy of **to** and **for**. Both can mark a DESTINATION or PURPOSE, while there are other functions that do not overlap. The syntactic complement use of infinitival **to** is tagged as `i. The **over** token in (1) receives the label APPROXIMATOR. See §3.1 for details.

2 Background and Motivation

Theoretical linguists have puzzled over questions such as how individual prepositions can acquire such a broad range of meanings and to what extent those meanings are systematically related (e.g.,

[1] STREUSLE 3.0, available at http://www.cs.cmu.edu/~ark/LexSem/

[2] http://www.wordfrequency.info/free.asp?s=y

[3] This work focuses on English, but adposition and case systems vary considerably across languages, challenging second language learners and machine translation systems (Chodorow et al., 2007; Shilon et al., 2012; Hashemi and Hwa, 2014).

Brugman, 1981; Lakoff, 1987; Tyler and Evans, 2003; O'Dowd, 1998; Saint-Dizier and Ide, 2006; Lindstromberg, 2010). Prepositional polysemy has also been recognized as a challenge for AI (Herskovits, 1986) and natural language processing, motivating semantic disambiguation systems (O'Hara and Wiebe, 2003; Ye and Baldwin, 2007; Hovy et al., 2010; Srikumar and Roth, 2013b). Training and evaluating these requires semantically annotated corpus data. Below, we comment briefly on existing resources and why (in our view) a new resource is needed to "road-test" an alternative, hopefully more scalable, semantic representation for prepositions.

2.1 Existing Preposition Corpora

Beginning with the seminal resources from The Preposition Project (TPP; Litkowski and Hargraves, 2005), the computational study of preposition semantics has been fundamentally grounded in corpus-based lexicography centered around individual preposition types. Most previous datasets of English preposition semantics at the token level (Litkowski and Hargraves, 2005, 2007; Dahlmeier et al., 2009; Tratz and Hovy, 2009; Srikumar and Roth, 2013a) only cover high-frequency prepositions (the 34 represented in the SemEval-2007 shared task based on TPP, or a subset thereof).[4]

We sought a scheme that would facilitate *comprehensive* semantic annotation of all preposition tokens in a corpus, covering the full range of usages possible for all English preposition types. The recent TPP PDEP corpus (Litkowski, 2014, 2015) comes closer to this goal, as it consists of randomly sampled tokens for over 300 types. However, since sentences were sampled separately for each preposition, there is only one annotated preposition token per sentence. By contrast, we will fully annotate documents for all preposition tokens. No inter-annotator agreement figures have been reported for the PDEP data to indicate its quality, or the overall difficulty of token annotation with TPP senses across a broad range of prepositions.

2.2 Supersenses

From the literature on other kinds of supersenses, there is reason to believe that token annotation with **preposition supersenses** (Schneider et al., 2015) will be more scalable and useful than senses. The term **supersense** has been applied to lexical semantic classes that label a large number of word types (i.e., they are unlexicalized). The best-known supersense scheme draws on two inventories—one for nouns and one for verbs—which originated as a high-level partitioning of senses in WordNet (Miller et al., 1990). A scheme for adjectives has been proposed as well (Tsvetkov et al., 2014).

One argument advanced in favor of supersenses is that they provide a coarse level of generalization for essential contextual distinctions—such as artifact vs. person for *chair*, or temporal vs. locative **in**—without being so fine-grained that systems cannot learn them (Ciaramita and Altun, 2006). A similar argument applies for *human* learning as pertains to rapid, cost-effective, and open-vocabulary annotation of corpora: an inventory of dozens of categories (with mnemonic names) can be learned and applied to unlimited vocabulary without having to refer to dictionary definitions (Schneider et al., 2012). Like with WordNet for nouns and verbs, the same argument holds for prepositions: TPP-style sense annotation requires familiarity with a different set of (often highly nuanced) distinctions for each preposition type. For example, **in** has 15 different TPP senses, among them **in 10(7a)** 'indicating the key in which a piece of music is written: *Mozart's Piano Concerto in E flat*'.

Supersenses have been exploited for a variety of tasks (e.g., Agirre et al., 2008; Tsvetkov et al., 2013, 2015), and full-sentence noun and verb taggers have been built for several languages (Segond et al., 1997; Johannsen et al., 2014; Picca et al., 2008; Martínez Alonso et al., 2015; Schneider et al., 2013, 2016). They are typically implemented as sequence taggers. In the present work, we extend a corpus that has already been hand-annotated with noun and verb supersenses, thus raising the possibility of systems that can learn all three kinds of supersenses jointly (cf. Srikumar and Roth, 2011).

Though they go by other names, the TPP "classes" (Litkowski, 2015),[5] the "clusters" of Tratz and Hovy (2011), and the "relations" of Srikumar and Roth (2013a) similarly label coarse-grained semantic functions of English prepositions; notably, they group senses from a lexicon rather than directly annotating tokens, and restrict each sense

[4] A further limitation of the SemEval-2007 dataset is the way in which it was sampled: illustrative tokens from a corpus were manually selected by a lexicographer. As Litkowski (2014) showed, a disambiguation system trained on this dataset will therefore be biased and perform poorly on an ecologically valid sample of tokens.

[5] `http://www.clres.com/db/classes/ClassAnalysis.php`

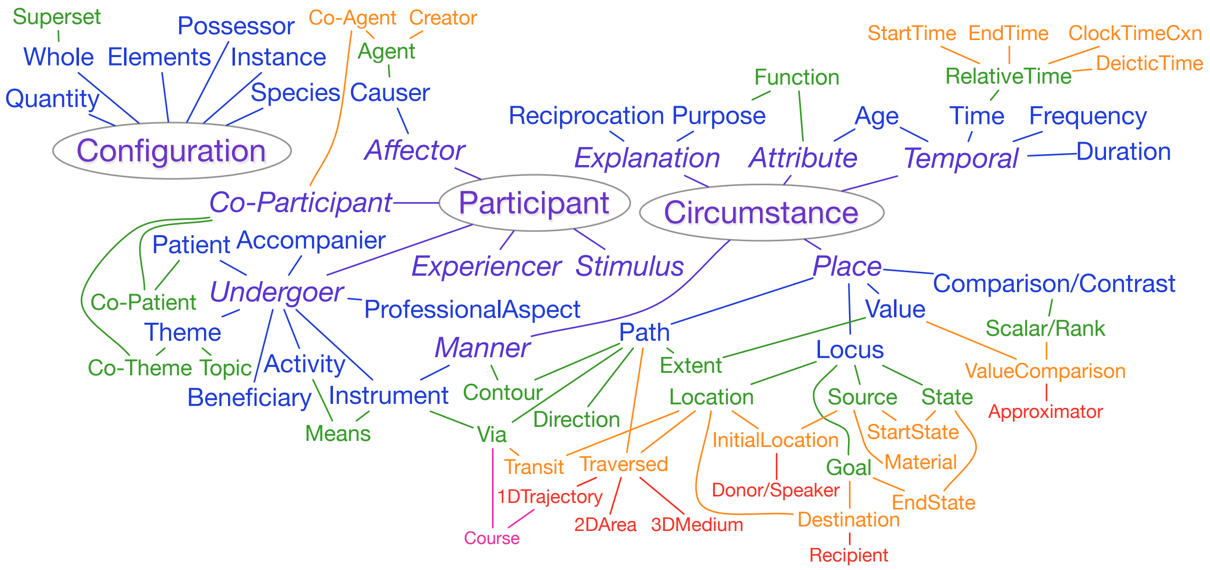


Figure 2: Supersense hierarchy used in this work (adapted from Schneider et al., 2015). Circled nodes are roots (the most abstract categories); subcategories are shown above and below. Each node's color and formatting reflect its depth.

to (at most) 1 grouping. Schneider et al. (2015) used the Srikumar and Roth (2013a) "relation" categories as a starting point in creating the preposition supersense inventory, but removed the assumption that each TPP sense could only belong to 1 category. Müller et al.'s (2012) semantic class inventory targets German prepositions.

2.3 PrepWiki

Schneider et al.'s (2015) preposition supersense scheme is described in detail in a lexical resource, PrepWiki,[6] which records associations between supersenses and preposition types. Hereafter, we adopt the term **usage** for a pairing of a preposition type and a supersense label (e.g., **at**/Time). Usages are organized in PrepWiki via (lexicalized) **senses** from the TPP lexicon. The mapping is many-to-many, as senses and supersenses capture different generalizations. (TPP senses, being lexicalized, are more numerous and generally finer-grained, but in some cases lump together functions that receive different supersenses, as in the sense **for 2(2)** 'affecting, with regard to, or in respect of'.) Thus, for a given preposition, a sense may be mapped to multiple usages, and vice versa.

2.4 The Supersense Hierarchy

Unlike the noun, verb, and adjective supersense schemes mentioned in §2.2, the preposition supersense inventory is hierarchical (as are Litkowski's (2015) and Müller et al.'s (2012) inventories). The hierarchy, depicted in figure 2, encodes inheritance: characteristics of higher-level categories are asserted to apply to their descendants. Multiple inheritance is used for cases of overlap: e.g., Destination inherits from both Location (because a destination is a point in physical space) and Goal (it is the endpoint of a concrete or abstract path).

The structure of the hierarchy was modeled after VerbNet's hierarchy of thematic roles (Bonial et al., 2011; Hwang, 2014). But there are many additional categories: some are refinements of the VerbNet roles (e.g., subclasses of Time), while others have no VerbNet counterpart because they do not pertain to core roles of verbs. The Configuration subhierarchy, used for **of** and other prepositions when they relate two nominals, is a good example.

The hierarchical structure will be useful for comparing against other annotation schemes which operate at different levels of granularity, as we do in §4 below. We expect that it will also help supervised classifiers to learn better generalizations when faced with sparse training data.

3 Corpus Annotation

3.1 Annotating Preposition Supersenses

Source data. We fully annotated the Reviews section of the English Web Treebank (Bies et al., 2012), chosen because it had previously been annotated for multiword expressions, noun and verb supersenses (Schneider et al., 2014; Schneider and Smith, 2015), and PropBank predicate-argument structures (§4). The corpus comprises 55,579 tokens organized into 3,812 sentences and 723 documents with gold tokenization and PTB-style POS

[6] http://tiny.cc/prepwiki

tags.

Identifying preposition tokens. TPP, and therefore PrepWiki, contains senses for canonical prepositions, i.e., those used transitively in the [$_{\text{PP}}$ P NP] construction. Taking inspiration from Pullum and Huddleston (2002), PrepWiki further assigns supersenses to spatiotemporal particle uses of **out**, **up**, **away**, **together**, etc., and subordinating uses of **as**, **after**, **in**, **with**, etc. (including infinitival **to** and infinitival-subject **for**, as in *It took over 1.5 hours* ***for*** *our food* ***to*** *come out*).[7]

Non-supersense labels. These are used where the preposition serves a special syntactic function not captured by the supersense inventory. The most frequent is `i, which applies only to infinitival **to** tokens that are not Purpose or Function adjuncts.[8] The label `d applies to discourse expressions like ***On*** *the other hand*; the unqualified backtick (`) applies to miscellaneous cases such as infinitival-subject **for** and both prepositions in the **as**-**as** comparative construction (***as*** *wet* ***as*** *water*; ***as*** *much cake* ***as*** *you want*).[9]

Multiword expressions. Figure 3 shows how prepositions can interact with multiword expressions (MWEs). An MWE may function holistically as a preposition: PrepWiki treats these as multiword prepositions. An idiomatic phrase may be headed by a preposition, in which case we assign it a preposition supersense or tag it as a discourse expression (`d: see the previous paragraph). Finally, a preposition may be embedded within an MWE (but not its head): we do not use a preposition supersense in this case, though the MWE as a whole may already be tagged with a verb supersense.

Heuristics. The annotation tool uses heuristics to detect candidate preposition tokens in each sentence given its POS tagging and MWE annotation. A *single-word expression* is included if: (a) it is tagged as a verb particle (RP) or infinitival **to** (TO), or, (b) it is tagged as a transitive preposition or subordinator (IN) or adverb (RB), and it is listed in PrepWiki (or the spelling variants list). A strong *MWE* instance is included if: (a) the MWE begins with a word that matches the single-word criteria (idiomatic PP), or, (b) the MWE is listed in PrepWiki (multiword preposition).

(4) **Because_of**/Explanation the ants I dropped them **to**/EndState a 3_star .

(5) I was told **to**/`i take my coffee **to**_go/Manner if I wanted **to**/`i finish it .

(6) **With**/Attribute higher **than**/Scalar/Rank average prices **to**_boot/`d !

(7) I worked~**with**/ProfessionalAspect Sam_Mones who took_ great _care_**of** me .

Figure 3: Prepositions involved in multiword expressions. (4) Multiword preposition **because of** (others include **in front of**, **due to**, **apart from**, and **other than**). (5) PP idiom: the preposition supersense applies to the MWE as a whole. (6) Discourse PP idiom: instead of a supersense, expressions serving a discourse function are tagged as `d. (7) Preposition within a multiword expression: the expression is headed by a verb, so it receives a verb supersense (not shown) rather than a preposition supersense.

Annotation task. Annotators proceeded sentence by sentence, working in a custom web interface (figure 4). For each token matched by the above heuristics, annotators filled in a text box with the contextually appropriate label. A dropdown menu showed the list of preposition supersenses and non-supersense labels, starting with labels known to be associated with the preposition being annotated. Hovering over a menu item would show example sentences to illustrate the usage in question, as well as a brief definition of the supersense. This preposition-specific rendering of the dropdown menu—supported by data from PrepWiki—was crucial to reducing the overhead of annotation (and annotator training) by focusing the annotator's attention on the relevant categories/usages. New examples were added to PrepWiki as annotators spotted coverage gaps. The tool also showed the multiword expression annotation of the sentence, which could be modified if necessary to fit PrepWiki's conventions for multiword prepositions.

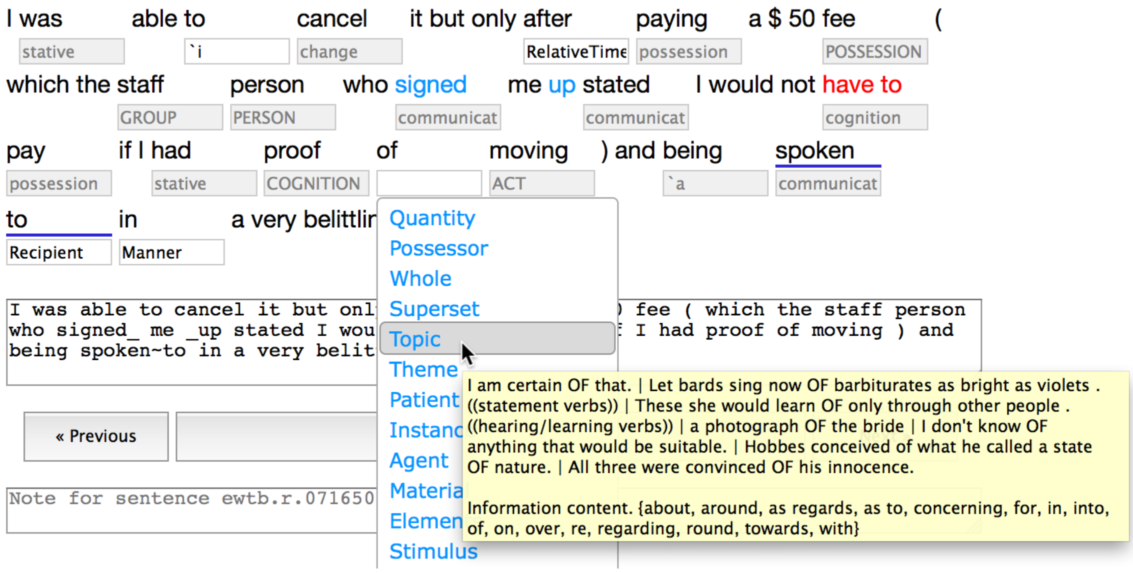
3.2 Quality Control

Annotators. Annotators were selected from undergraduate and graduate linguistics students at the University of Colorado at Boulder. All annotators had prior experience with semantic role labeling. Every sentence was independently annotated by two annotators, and disagreements were subse-

[7] PrepWiki does not include subordinators/complementizers that cannot take NP complements: *that*, *because*, *while*, *if*, etc.

[8] Because the word **to** is ambiguous between infinitival and prepositional usages, and because infinitivals, like PPs, can serve as Purpose or Function modifiers, we allow infinitival **to** to be so marked. E.g., *a shoulder* ***to*** *cry on* would qualify as Function. By contrast, *I want/love/try* ***to*** *eat cookies* and ***To*** *love is* ***to*** *suffer* would qualify as `i. See figure 1 for examples from the corpus.

[9] Annotators used additional non-supersense labels to mark tokens that were incorrectly flagged as prepositions by our heuristics: e.g., *price was way to high* was marked as an adverb. We ignore these tokens for purposes of this paper.



I was able to cancel it but onl… fee (which the staff person who signed_ me _up stated I wou… I had proof of moving) and being spoken~to in a very belit…

Note for sentence ewtb.r.071650

I am certain OF that. | Let bards sing now OF barbiturates as bright as violets . ((statement verbs)) | These she would learn OF only through other people . ((hearing/learning verbs)) | a photograph OF the bride | I don't know OF anything that would be suitable. | Hobbes conceived of what he called a state OF nature. | All three were convinced OF his innocence.

Information content. {about, around, as regards, as to, concerning, for, in, into, of, on, over, re, regarding, round, towards, with}

Figure 4: Supersense annotation interface, developed in-house. The main thing to note is that preposition, noun, and verb supersenses are stored in text boxes below the sentence. A dropdown menu displays the full list of preposition supersenses, starting with those with PrepWiki mappings to the preposition in question. Hovering the mouse over a menu item displays a tooltip with PrepWiki examples of the usage (if applicable) and a general definition of the supersense.

quently adjudicated by a third, “expert” annotator. There were two expert annotators, both authors of this paper.

Training. 200 sentences were set aside for training annotators. Annotators were first shown how to use the preposition annotation tool and instructed on the supersense distinctions for this task. Annotators then completed a training set of 100 sentences. An adjudicator evaluated the annotator’s annotations, providing feedback and assigning another 50–100 training instances if necessary.

Inter-annotator agreement (IAA) measures are useful in quantifying annotation “reliability”, i.e., indicating how trustworthy and reproducible the process is (given guidelines, training, tools, etc.). Specifically, IAA scores can be used as a diagnostic for the reliability of (i) individual annotators (to identify those who need additional training/guidance); (ii) the annotation scheme and guidelines (to identify problematic phenomena requiring further documentation or changes to the scheme); (iii) the final dataset (as an indicator of what could reasonably be expected of an automatic system).

Individual annotators. The main annotation was divided into 34 batches of 100 sentences. Each batch took on the order of an hour for an annotator to complete. We monitored original annotators’ IAA throughout the annotation process as a diagnostic for when to intervene in giving further guidance. Original IAA for most of these batches fell between 60% and 78%, depending on factors such as the identities of the annotators and when the annotation took place (annotator experience and PrepWiki documentation improved over time).[10] These rates show that it was not an easy annotation task, though many of the disagreements were over slight distinctions in the hierarchy (such as PURPOSE vs. FUNCTION).

Guidelines. Though Schneider et al. (2015) conducted pilot annotation in constructing the supersense inventory, our annotators found a few details of the scheme to be confusing. Informed by their difficulties and disagreements, we therefore made several minor improvements to the preposition supersense categories and hierarchy structure. For example, the supersense categories for partitive constructions proved persistently problematic, so we adjusted their boundaries and names. We also improved the high-level organization of the original hierarchy, clarified some supersense descriptions, and removed the miscellaneous OTHER supersense.

Revisions. The changes to categories/guidelines noted in the previous paragraph required a small-scale post hoc revision to the annotations by the expert annotators. Some additional post hoc revisions were performed to improve consistency, e.g., some anomalous multiword expression annotations

[10] The agreement rate among tokens where both annotators assigned a preposition supersense was between 82% and 87% for 4 batches; 72% and 78% for 11 batches; 60% and 70% for 17 batches; and below 60% for 2 batches. This measure did not award credit for agreement on non-supersense labels and ignored some cases of disagreement on the MWE analysis.

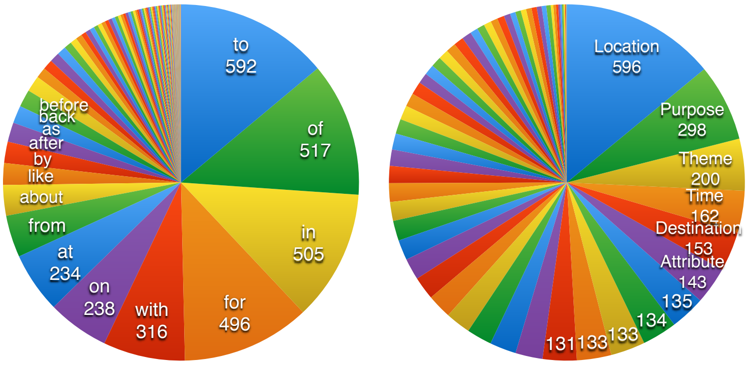


Figure 5: Distributions of preposition types and supersenses for the 4,250 supersense-tagged preposition tokens in the corpus. Observe that just 9 prepositions account for 75% of tokens, whereas the head of the supersense distribution is much smaller.

involving prepositions were fixed.[11]

Expert IAA. We also measured IAA on a sample independently annotated from scratch by both experts.[12] Applying this procedure to 203 sentences annotated late in the process (using the measure described in footnote 10) gives an agreement rate of 276/313 = 88%.[13] Because every sentence in the rest of the corpus was adjudicated by one of these two experts, the expert IAA is a rough estimate of the dataset's adjudication reliability—i.e., the expected proportion of tokens that would have been labeled the same way if adjudicated by the other expert. While it is difficult to put an exact quality figure on a dataset that was developed over a period of time and with the involvement of many individuals, the fact that the expert-to-expert agreement approaches 90% despite the large number of labels suggests that the data can serve as a reliable resource for training and benchmarking disambiguation systems.

3.3 Resulting Corpus

4,250 tokens in the corpus have preposition supersenses. 114 prepositions and 63 supersenses are attested.[14] Their distributions appear in figure 5. Over 75% of tokens belong to the top 10 preposition types, while the supersense distribution is closer to uniform. 1,170 tokens are labeled as LOCATION, PATH, or a subtype thereof: these can roughly be described as spatial. 528 come from the TEMPORAL subtree of the hierarchy, and 452 from the CONFIGURATION subtree. Thus, fully half the tokens (2,100) mark non-spatiotemporal participants and circumstances.

Of the 4,250 tokens, 582 are MWEs (multiword prepositions and/or PP idioms). A further 588 preposition tokens (not included in the 4,250) have non-supersense labels: 484 `i, 83 `d, and 21 `.

3.4 Splits

To facilitate future experimentation on a standard benchmark, we partitioned our data into training and test sets. We randomly sampled 447 *sentences* (4,073 total tokens and 950 (19.6%) preposition instances) for a held-out test set, leaving 3,888 preposition instances for training.[15] The sampling was stratified by preposition supersense to encourage a reasonable balance for the rare labels; e.g., supersenses that occur twice are split so that one instance is assigned to the training set and one to the test set.[16] 61 preposition supersenses are attested in the training data, while 14 are unattested.

4 Inter-annotation Evaluation with PropBank

The REVIEWS corpus that we annotated with preposition supersenses had been independently

[11] In particular, many of the borderline prepositional verbs were revised according to the guidelines outlined at `https://github.com/nschneid/nanni/wiki/Prepositional-Verb-Annotation-Guidelines`.

[12] These sentences were then jointly adjudicated by the experts to arrive at a final version.

[13] For completeness, Cohen's κ = .878. It is almost as high as raw agreement because the expected agreement rate is very low, but keep in mind that κ's model of chance agreement does not take into account preposition types or the fact that, for a given type, a relatively small subset of labels were suggested to the annotator. On the 4 most frequent prepositions in the sample, *per-preposition* κ is .84 for **for**, 1.0 for **to**, .59 for **of**, and .73 for **in**.

[14] For the purpose of counting prepositions by type, we split up supersense-tagged PP idioms such as those shown in (5) and (6) by taking the longest prefix of words that has a PrepWiki entry to be the preposition.

[15] These figures include tokens with non-supersense labels (§3.1); the supersense-labeled prepositions amount to 3,397 training and 853 test instances.

[16] The sampling algorithm considered supersenses in increasing order of frequency: for each supersense ℓ having n_ℓ instances, enough sentences were assigned to the test set to fill a minimum quota of $\lceil .195 n_\ell \rceil$ tokens for that supersense (and remaining unassigned sentences containing that supersense were placed in the training set). Relative to the training set, the test set is skewed slightly in favor of rarer supersenses. A small number of annotation errors were corrected after determining the splits. Entire sentences were sampled to facilitate future studies involving joint prediction over the full sentence.

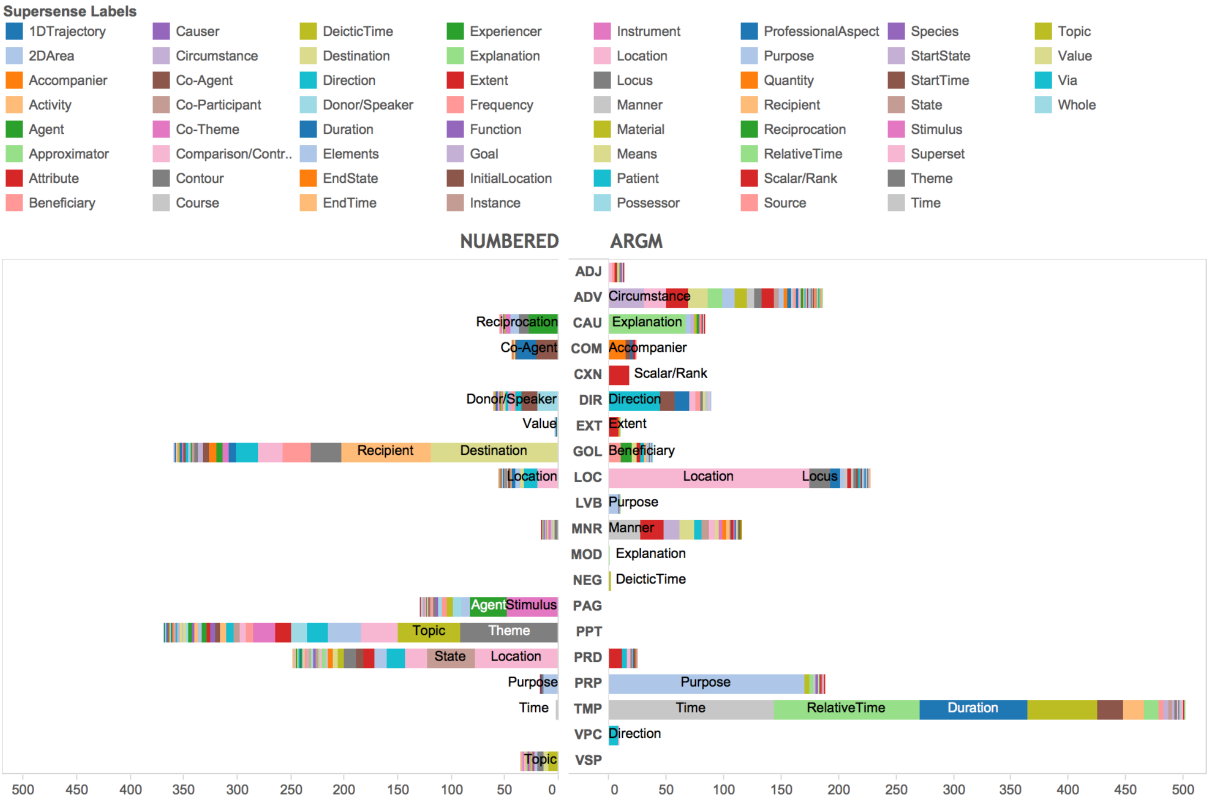


Figure 6: PropBank function tags on PP arguments and counts of their observed token correspondences with preposition supersenses. For each function tag, counts are split into numbered (core) arguments, left, and `ArgM` (modifier/non-core) arguments, right.

annotated with PropBank (Palmer et al., 2005; Bonial et al., 2014) predicate-argument structures. As a majority of preposition usages mark a semantic role, this affords us the opportunity to empirically compare the two annotation schemes as applied to the dataset—assessing not just inter-*annotator* agreement, but also inter-*annotation* agreement. (Our annotators did not have access to the PropBank annotations.) Others have conducted similar token-level analyses to compare different semantic representations (e.g., Fellbaum and Baker, 2013).

The supersense inventory is finer-grained than the PropBank function tags, ruling out a one-to-one correspondence. However, if the two sets of categories are both linguistically valid and correctly applied, then we expect that a label from either scheme will be predictive of the other scheme's label(s). Thus, we investigate the kinds and causes of divergence to see whether they reveal theoretical or practical problems with either scheme.

4.1 Function Tags in PropBank

In comparing our supersense annotation to the PropBank annotation of prepositional phrases, we focus on the mapping of the supersenses to PropBank's **function tags** marking location (`LOC`), extent (`EXT`), cause (`CAU`), temporal (`TMP`), and manner (`MNR`), among others.

Originally associated with modifier (`ArgM`) labels, function tags were recently added to all PropBank numbered arguments in an effort to address the performance problems in SRL systems caused by the higher-numbered arguments (Bonial et al., 2016).[17] In addition to the 13 existing function tags, three tags were introduced specifically for numbered roles: Proto-Agent (`PAG`), Proto-Patient (`PPT`), and Verb-Specific (`VSP`). These three tags are used, respectively, for `Arg0`, `Arg1`, and other arguments that simply do not have an appropriate function tag because they are unique to the lemma in question. Each of the numbered arguments has thus been annotated with a function tag. Unlike modifiers, where the function tag is annotated at the token level, function tags on the numbered arguments were assigned at the type level (in verbs' frameset definitions) by selecting the function tag most applicable to existing annotations.

Example (8) shows a sentence annotated for the predicate *going*; function tags appear in each argu-

[17] While automatic SRL performance is quite good for the detection of `Arg0` and `Arg1`, the performance on identification of higher-numbered arguments, 2–6, is relatively poor due to the variety of semantic roles they are associated with, depending on which relation is being considered.

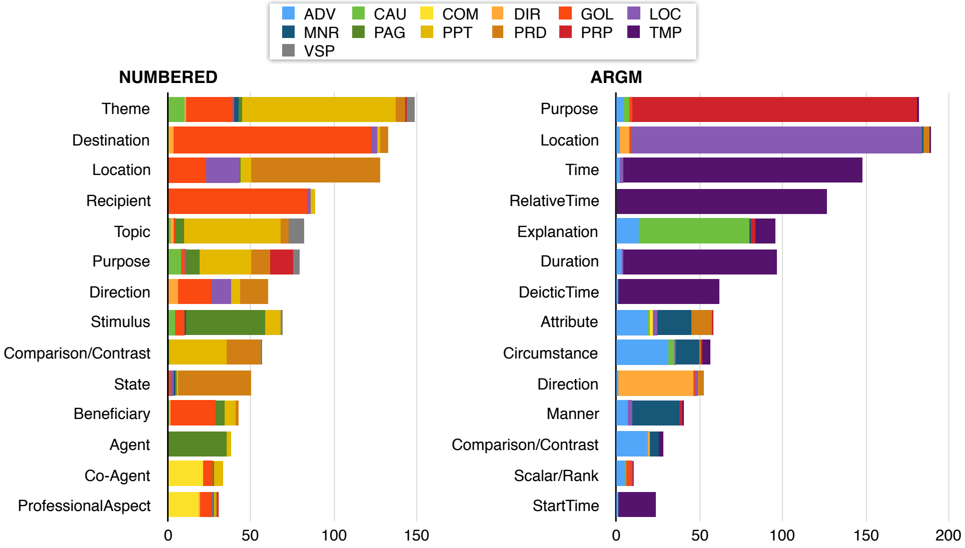


Figure 7: Distribution of PropBank function tags for the most frequent mapped supersenses. Counts are split into numbered (core) arguments, left, and ArgM (modifier/non-core) arguments, right.

Function tags mapped to fewer than 20 supersense-tagged prepositions overall are not displayed. (This accounts for why the bars are not strictly decreasing in width.) Numbered arguments tagged with VPC are mapped to DIRECTION in 8 instances. ArgM-LVB is mapped to PURPOSE in 8 instances, while ArgM-CXN is the dominant function tag mapped to SCALAR/RANK (18 instances).

ment name, following a hyphen:

(8) **I**$_{\texttt{Arg1-PPT}}$ have been **going**$_{\texttt{rel}}$ **[to the Wildwood, NJ]**$_{\texttt{Arg4-GOL}}$ **[for over 30 years]**$_{\texttt{ArgM-TMP}}$ **[for summer vacations]**$_{\texttt{ArgM-PRP}}$.

Of interest to this study are the three labels assigned to the prepositional phrases—Arg4-GOL, ArgM-TMP, and ArgM-PRP—and their corresponding supersense labels in (1). If the supersense annotation is valid, we should see a consistent correspondence between these PropBank function tags and semantically equivalent supersenses DESTINATION, DURATION, and PURPOSE, respectively, or their semantic relatives in the hierarchy.

Of the 4,250 supersense-annotated preposition tokens in the REVIEWS corpus (see §3.1), we were able to map 2,973 to arguments in the PropBank annotation—1,435 numbered arguments and 1,538 ArgM arguments.[18] Most of the remaining prepositions belong to non-predicative NPs and multiword expressions, which PropBank does not annotate.

4.2 Supersense and PropBank function tag correspondence

Figures 6 and 7 show the distribution of correspondences between the PropBank function tags and the supersense labels. Figure 6 visualizes all the mapped tokens, organized by function tag; figure 7 visualizes the function tag distributions for the most frequent supersenses that could be mapped.

Modifiers. We find that the supersense hierarchy captures some of the same generalizations as PropBank's coarser-grained distinctions. Most notably, the PropBank ArgM labels (visualized in the right-hand sides of figures 6 and 7) correspond relatively cleanly to the supersense labels: PropBank's TMP maps exclusively to the TEMPORAL branch of the hierarchy; and PRP, CAU, and to a slightly lesser extent LOC, map cleanly to their supersense counterparts PURPOSE, EXPLANATION, and LOCUS (and its subcategory LOCATION). The supersenses ATTRIBUTE, CIRCUMSTANCE, MANNER and the function tags ADV, MNR, PRD, and GOL stand out as warranting further scrutiny as applied to ArgMs.

Numbered arguments. The situation for numbered arguments is considerably messier. Note, for example, that in the left portion of figure 7, only a few of the supersenses map consistently to a single function tag: DESTINATION and RECIPIENT to GOL, STATE to PRD, and AGENT to PAG. The mappings for THEME, LOCATION, PURPOSE, and DIRECTION are extremely inconsistent. In part this is because PropBank captures predicate-centric, sometimes orthogonal distinctions: e.g., the copula is tagged as **be.01**, and its complement is always PRD—whether the PP describes a location (*It is in the box*), state (*We are in danger*), time (*That was 4 years ago*), etc. Other verbs, like *stay* and *find*, similarly have an argument tagged

[18] To perform the mapping, we first converted the gold PropBank annotations into a dependency representation using ClearNLP (`https://github.com/clir/clearnlp`; Choi and Palmer, 2012) and then heuristically postprocessed the output for special cases such as infinitival **to** marked as PURPOSE.

as PRD because that argument's function is to elaborate some other argument. Of course, *that* they elaborate some other argument is different from *how* (with respect to location, state, time, or other function conveyed by the preposition).

Because Arg0 and Arg1 had been consistently assigned to the verb's proto-agent (PAG) and proto-patient (PPT), respectively, we expected PAG to correspond cleanly to the Affector subhierarchy, and PPT to the Undergoer subhierarchy. We find that to a large extent, Arg0 does correspond to the Affector subhierarchy, which includes Agent and Causer. However, Arg0 also maps to other supersenses such as Stimulus (an entity that prompts sensory input), Topic (an Undergoer), and Purpose (a Circumstance). The source of the difference is partly due to a systematic disagreement on the status of a semantic label. Consider the following two PropBank frames:

amuse.01	**see.01**
Arg0-PAG: causer of mirth	Arg0-PAG: viewer
Arg1-PPT: mirthful entity	Arg1-PPT: thing viewed
Arg2-MNR: instrument	Arg2-PRD: attribute of Arg1
"Mary was amused **by** John"	"Mary was seen **by** John"

The preposition **by** for verbs *amuse* and *see* would carry the supersense labels of Stimulus (entity triggering amusement) and Experiencer (entity experiencing the sight), respectively. But PropBank's choice is verb-specific, assigning PAG based on which argument displays volitional involvement in the event or is causing an event or a state change in another participant (Bonial et al., 2012). Experiencer and Stimulus are known to compete over Dowty's Proto-Agent status, so this type of mismatch is not surprising (Dowty, 1991).

Arg1 is similarly muddled. Setting aside the expected mappings to Theme and Topic—both of which are undergoers—Arg1 overlaps with Stimulus (for the same reasons as cited above) and, also, to a wide range of semantics including Purpose, Attribute, and Comparison/Contrast.

Post hoc analysis. Well after the original annotation and adjudication, we undertook a post hoc review of the supersense-annotated tokens that were also PropBank-annotated to determine how much noise was present in the correspondences. We created a sample of 224 such tokens, stratified to cover a variety of correspondences (most supersenses were allotted 4 samples each, and for each supersense, function tags were diversified to the extent possible). Each token in the sample was reviewed independently by 4 annotators (all authors of this paper). Two annotators passed judgment on the gold supersense annotations; there were just 6 tokens for which they both said the supersense was clearly incorrect. The other two annotators (who have PropBank expertise) checked the gold PropBank annotations, agreeing that 5 of the tokens were clearly incorrect.

This analysis tells us that obvious errors with both types of annotation are indeed present in the corpus (11 tokens in the sample), adding some noise to the supersense–function tag correspondences. However, the outright errors are probably dwarfed by difficult/borderline cases for which the annotations are not entirely consistent throughout the corpus. For example, ***on*** *time* (i.e., 'not late') is variously annotated as State, Manner, and Time. Inconsistency detection methods (e.g., Hollenstein et al., 2016) may help identify these—though it remains to be seen whether methods developed for nouns and verbs would succeed on function words so polysemous as prepositions.

Summary. The (mostly) clean correspondences of the supersenses to the independently annotated PropBank *modifier* labels speak to the linguistic validity of our supersense hierarchy. On the other hand, the confusion evident for the supersense labels corresponding to PropBank's *numbered* arguments suggests further analysis and refinement is necessary for both annotation schemes. Some of these issues—especially correspondences between labels with unrelated semantics that occur in no more than a few tokens—are due to erroneous supersense or PropBank annotations. However, other categorizations are pervasively inconsistent between the two schemes, warranting a closer examination.

5 Conclusion

We have introduced a new lexical semantics corpus that disambiguates prepositions with hierarchical supersenses. Because it is comprehensively annotated over full documents (English web reviews), it offers insights into the semantic distribution of prepositions within that genre. Moreover, the same corpus has independently been annotated with PropBank predicate-argument structures, which facilitates analysis of correspondences and further refinement of both schemes and datasets. We expect that comprehensively annotated preposition supersense data will facilitate the development of automatic preposition disambiguation systems.

Acknowledgments

We thank our annotators—Evan Coles-Harris, Audrey Farber, Nicole Gordiyenko, Megan Hutto, Celeste Smitz, and Tim Watervoort—as well as Ken Litkowski, Michael Ellsworth, Orin Hargraves, and Susan Brown for helpful discussions. This research was supported in part by a Google research grant for Q/A PropBank Annotation.

References

Eneko Agirre, Timothy Baldwin, and David Martinez. 2008. Improving parsing and PP attachment performance with sense information. In *Proc. of ACL-HLT*, pages 317–325. Columbus, Ohio, USA.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English Web Treebank. Technical Report LDC2012T13, Linguistic Data Consortium, Philadelphia, PA. URL http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2012T13.

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D. Hwang, and Martha Palmer. 2014. PropBank: semantics of new predicate types. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 3013–3019. Reykjavík, Iceland.

Claire Bonial, Kathryn Conger, Jena D. Hwang, Aous Mansouri, Yahya Aseri, Julia Bonn, Timothy O'Gorman, and Martha Palmer. 2016. Current directions in English and Arabic PropBank. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*. Springer, New York.

Claire Bonial, William Corvey, Martha Palmer, Volha V. Petukhova, and Harry Bunt. 2011. A hierarchical unification of LIRICS and VerbNet semantic roles. In *Fifth IEEE International Conference on Semantic Computing*, pages 483–489. Palo Alto, CA, USA.

Claire Bonial, Jena D. Hwang, Julia Bonn, Kathryn Conger, Olga Babko-Malaya, and Martha Palmer. 2012. English propbank annotation guidelines. *Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder*.

Claudia Brugman. 1981. *The story of 'over': polysemy, semantics and the structure of the lexicon*. MA thesis, University of California, Berkeley, Berkeley, CA. Published New York: Garland, 1981.

Martin Chodorow, Joel R. Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proc. of the Fourth ACL-SIGSEM Workshop on Prepositions*, pages 25–30. Prague, Czech Republic.

Jinho D. Choi and Martha Palmer. 2012. Guidelines for the CLEAR style constituent to dependency conversion. Technical Report 01-12, Institute of Cognitive Science, University of Colorado at Boulder, Boulder, Colorado, USA. URL http://www.mathcs.emory.edu/~choi/doc/clear-dependency-2012.pdf.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP*, pages 594–602. Sydney, Australia.

Daniel Dahlmeier, Hwee Tou Ng, and Tanja Schultz. 2009. Joint learning of preposition senses and semantic roles of prepositional phrases. In *Proc. of EMNLP*, pages 450–458. Suntec, Singapore.

Mark Davies. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, 25(4):447–464.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, pages 547–619.

Christiane Fellbaum and Collin F. Baker. 2013. Comparing and harmonizing different verb classifications in light of a semantic annotation task. *Linguistics*, 51(4):707–728.

Homa B. Hashemi and Rebecca Hwa. 2014. A comparison of MT errors and ESL errors. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 2696–2700. Reykjavík, Iceland.

Annette Herskovits. 1986. *Language and spatial cognition: an interdisciplinary study of the prepositions in English*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, UK.

Nora Hollenstein, Nathan Schneider, and Bonnie Webber. 2016. Inconsistency detection in semantic annotation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 3986–3990. Portorož, Slovenia.

Dirk Hovy, Stephen Tratz, and Eduard Hovy. 2010. What's in a preposition? Dimensions of sense disambiguation for an interesting word class. In *Coling 2010: Posters*, pages 454–462. Beijing, China.

Jena D. Hwang. 2014. *Identification and representation of caused motion constructions*. Ph.D. dissertation, University of Colorado, Boulder, Colorado.

Anders Johannsen, Dirk Hovy, Héctor Martínez Alonso, Barbara Plank, and Anders Søgaard. 2014. More or less supervised supersense tagging of Twitter. In *Proc. of *SEM*, pages 1–11. Dublin, Ireland.

George Lakoff. 1987. *Women, fire, and dangerous things: what categories reveal about the mind*. University of Chicago Press, Chicago.

Seth Lindstromberg. 2010. *English Prepositions Explained*. John Benjamins, Amsterdam, revised edition.

Ken Litkowski. 2014. Pattern Dictionary of English Prepositions. In *Proc. of ACL*, pages 1274–1283. Baltimore, Maryland, USA.

Ken Litkowski. 2015. Notes on barbecued opakapaka: ontology in preposition patterns. Technical Report 15-01, CL Research, Damascus, MD. URL http://www.clres.com/online-papers/PDEPOntology.pdf.

Ken Litkowski and Orin Hargraves. 2005. The Preposition Project. In *Proc. of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 171–179. Colchester, Essex, UK.

Ken Litkowski and Orin Hargraves. 2007. SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions. In *Proc. of SemEval*, pages 24–29. Prague, Czech Republic.

Héctor Martínez Alonso, Anders Johannsen, Sussi Olsen, Sanni Nimb, Nicolai Hartvig Sørensen, Anna Braasch, Anders Søgaard, and Bolette Sandford Pedersen. 2015. Supersense tagging for Danish. In Beáta Megyesi, editor, *Proc. of NODALIDA*, pages 21–29. Vilnius, Lithuania.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Five Papers on WordNet. Technical Report 43, Princeton University, Princeton, NJ.

Antje Müller, Claudia Roch, Tobias Stadtfeld, and Tibor Kiss. 2012. The annotation of preposition senses in German. In Britta Stolterfoht and Sam Featherston, editors, *Empirical Approaches to Linguistic Theory: Studies in Meaning and Structure*, Studies in Generative Grammar, pages 63–82.

Walter de Gruyter, Berlin.

Elizabeth M. O'Dowd. 1998. *Prepositions and particles in English: a discourse-functional account*. Oxford University Press, New York.

Tom O'Hara and Janyce Wiebe. 2003. Preposition semantic classification via Treebank and FrameNet. In Walter Daelemans and Miles Osborne, editors, *Proc. of CoNLL*, pages 79–86. Edmonton, Canada.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: an annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Davide Picca, Alfio Massimiliano Gliozzo, and Massimiliano Ciaramita. 2008. 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.

Geoffrey K. Pullum and Rodney Huddleston. 2002. Prepositions and preposition phrases. In Rodney Huddleston and Geoffrey K. Pullum, editors, *The Cambridge Grammar of the English Language*, pages 579–611. Cambridge University Press, Cambridge, UK.

Patrick Saint-Dizier and Nancy Ide, editors. 2006. *Syntax and Semantics of Prepositions*, volume 29 of *Text, Speech and Language Technology*. Springer, Dordrecht, The Netherlands.

Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval-2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM). In *Proc. of SemEval*. San Diego, California, USA.

Nathan Schneider, Behrang Mohit, Chris Dyer, Kemal Oflazer, and Noah A. Smith. 2013. Supersense tagging for Arabic: the MT-in-the-middle attack. In *Proc. of NAACL-HLT*, pages 661–667. Atlanta, Georgia, USA.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A. Smith. 2012. Coarse lexical semantic annotation with supersenses: an Arabic case study. In *Proc. of ACL*, pages 253–258. Jeju Island, Korea.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 455–461. Reykjavík, Iceland.

Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proc. of NAACL-HLT*, pages 1537–1547. Denver, Colorado.

Nathan Schneider, Vivek Srikumar, Jena D. Hwang, and Martha Palmer. 2015. A hierarchy with, of, and for preposition supersenses. In *Proc. of The 9th Linguistic Annotation Workshop*, pages 112–123. Denver, Colorado, USA.

Frédérique Segond, Anne Schiller, Gregory Grefenstette, and Jean-Pierre Chanod. 1997. An experiment in semantic tagging using hidden Markov model tagging. In Piek Vossen, Geert Adriaens, Nicoletta Calzolari, Antonio Sanfilippo, and Yorick Wilks, editors, *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications: ACL/EACL-97 Workshop Proceedings*, pages 78–81. Madrid, Spain.

Reshef Shilon, Hanna Fadida, and Shuly Wintner. 2012. Incorporating linguistic knowledge in statistical machine translation: translating prepositions. In *Proc. of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 106–114. Avignon, France.

Vivek Srikumar and Dan Roth. 2011. A joint model for extended semantic role labeling. In *Proc. of EMNLP*, pages 129–139. Edinburgh, Scotland, UK.

Vivek Srikumar and Dan Roth. 2013a. An inventory of preposition relations. Technical Report arXiv:1305.5785. URL `http://arxiv.org/abs/1305.5785`.

Vivek Srikumar and Dan Roth. 2013b. Modeling semantic relations expressed by prepositions. *Transactions of the Association for Computational Linguistics*, 1:231–242.

Stephen Tratz and Dirk Hovy. 2009. Disambiguation of preposition sense using linguistically motivated features. In *Proc. of NAACL-HLT Student Research Workshop and Doctoral Consortium*, pages 96–100. Boulder, Colorado.

Stephen Tratz and Eduard Hovy. 2011. A fast, accurate, non-projective, semantically-enriched parser. In *Proc. of EMNLP*, pages 1257–1268. Edinburgh, Scotland, UK.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proc. of EMNLP*. Lisbon, Portugal.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proc. of the First Workshop on Metaphor in NLP*, pages 45–51. Atlanta, Georgia, USA.

Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui, and Chris Dyer. 2014. Augmenting English adjective senses with supersenses. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proc. of LREC*, pages 4359–4365. Reykjavík, Iceland.

Andrea Tyler and Vyvyan Evans. 2003. *The Semantics of English Prepositions: Spatial Scenes, Embodied Meaning and Cognition*. Cambridge University Press, Cambridge, UK.

Patrick Ye and Timothy Baldwin. 2007. MELB-YB: Preposition sense disambiguation using rich semantic features. In *Proc. of SemEval*, pages 241–244. Prague, Czech Republic.