Comprehensive Supersense Disambiguation of English Prepositions and Possessives

Nathan Schneider[*]
Georgetown University

Jena D. Hwang
IHMC

Vivek Srikumar
University of Utah

Jakob Prange
Austin Blodgett
Georgetown University

Sarah R. Moeller
University of Colorado Boulder

Aviram Stern
Adi Bitan
Omri Abend
Hebrew University of Jerusalem

Abstract

Semantic relations are often signaled with prepositional or possessive marking—but extreme polysemy bedevils their analysis and automatic interpretation. We introduce a new annotation scheme, corpus, and task for the disambiguation of prepositions and possessives in English. Unlike previous approaches, our annotations are comprehensive with respect to types and tokens of these markers; use broadly applicable supersense classes rather than fine-grained dictionary definitions; unite prepositions and possessives under the same class inventory; and distinguish between a marker's *lexical* contribution and the *role* it marks in the context of a predicate or scene. Strong interannotator agreement rates, as well as encouraging disambiguation results with established supervised methods, speak to the viability of the scheme and task.

1 Introduction

Grammar, as per a common metaphor, gives speakers of a language a shared toolbox to construct and deconstruct meaningful and fluent utterances. Being highly analytic, English relies heavily on word order and closed-class function words like prepositions, determiners, and conjunctions. Though function words bear little semantic content, they are nevertheless crucial to the meaning. Consider prepositions: they serve, for example, to convey place and time (*We met* ***at/in/outside*** *the restaurant* ***for/after*** *an hour*), to express configurational relationships like quantity, possession, part/whole, and membership (*the coats* ***of*** *dozens* ***of*** *children* ***in*** *the class*), and to indicate semantic roles in argument structure (*Grandma cooked dinner* ***for*** *the children*

(1) I was booked **for**/DURATION 2 nights **at**/LOCUS this hotel **in**/TIME Oct 2007 .

(2) I went **to**/GOAL ohm **after**/EXPLANATION⇝TIME reading some **of**/QUANTITY⇝WHOLE the reviews .

(3) It was very upsetting to see this kind **of**/SPECIES behavior especially **in_front_of**/LOCUS **my**/SOCIALREL⇝GESTALT four year_old .

Figure 1: Annotated sentences from our corpus.

vs. *Grandma cooked the children* ***for*** *dinner*). Frequent prepositions like **for** are maddeningly polysemous, their interpretation depending especially on the object of the preposition—*I rode the bus* ***for*** *5 dollars/minutes*—and the governor of the prepositional phrase (PP): *I Ubered/asked* ***for*** *$5*. Possessives are similarly ambiguous: *Whistler's mother/painting/hat/death*. Semantic interpretation requires some form of sense disambiguation, but arriving at a linguistic representation that is flexible enough to generalize across usages and types, yet simple enough to support reliable annotation, has been a daunting challenge (§2).

This work represents a new attempt to strike that balance. Building on prior work, we argue for an approach to describing English preposition and possessive semantics with broad coverage. Given the semantic overlap between prepositions and possessives (*the hood* ***of*** *the car* vs. *the car's hood* or ***its*** *hood*), we analyze them using the same inventory of semantic labels.[1] Our contributions include:

- a new hierarchical **inventory** ("SNACS") of 50 supersense classes, extensively documented in guidelines for English (§3);
- a gold-standard **corpus** with *comprehensive* annotations: all types and tokens of prepositions and possessives are disambiguated (§4; example sentences appear in figure 1);
- an **interannotator agreement** study that

[*] nathan.schneider@georgetown.edu

[1] Some uses of certain other closed-class markers—intransitive particles, subordinators, infinitive **to**—are also included (§3.1).

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 185–196
Melbourne, Australia, July 15 - 20, 2018.

shows the scheme is reliable and generalizes across genres—and for the first time demonstrating empirically that the lexical semantics of a preposition can sometimes be detached from the PP's semantic role (§5);
- **disambiguation** experiments with two supervised classification architectures to establish the difficulty of the task (§6).

2 Background: Disambiguation of Prepositions and Possessives

Studies of preposition semantics in linguistics and cognitive science have generally focused on the domains of space and time (e.g., Herskovits, 1986; Bowerman and Choi, 2001; Regier, 1996; Khetarpal et al., 2009; Xu and Kemp, 2010; Zwarts and Winter, 2000) or on motivated polysemy structures that cover additional meanings beyond core spatial senses (Brugman, 1981; Lakoff, 1987; Tyler and Evans, 2003; Lindstromberg, 2010). Possessive constructions can likewise denote a number of semantic relations, and various factors—including semantics—influence whether attributive possession in English will be expressed with **of**, or with **'s** and possessive pronouns (the 'genitive alternation'; Taylor, 1996; Nikiforidou, 1991; Rosenbach, 2002; Heine, 2006; Wolk et al., 2013; Shih et al., 2015).

Corpus-based computational work on semantic disambiguation specifically of prepositions and possessives[2] falls into two categories: the **lexicographic/word sense disambiguation** approach (Litkowski and Hargraves, 2005, 2007; Litkowski, 2014; Ye and Baldwin, 2007; Saint-Dizier, 2006; Dahlmeier et al., 2009; Tratz and Hovy, 2009; Hovy et al., 2010, 2011; Tratz and Hovy, 2013), and the **semantic class** approach (Moldovan et al., 2004; Badulescu and Moldovan, 2009; O'Hara and Wiebe, 2009; Srikumar and Roth, 2011, 2013; Schneider et al., 2015, 2016; Hwang et al., 2017, see also Müller et al., 2012 for German). The lexicographic approach can capture finer-grained meaning distinctions, at a risk of relying upon idiosyncratic and potentially incomplete dictionary definitions. The semantic class approach, which we follow here, focuses on commonalities in meaning across multiple lexical items, and aims to generalize more easily to new types and usages.

The most recent class-based approach to prepositions was our initial framework of 75 **preposition supersenses** arranged in a multiple inheritance taxonomy (Schneider et al., 2015, 2016). It was based largely on relation/role inventories of Srikumar and Roth (2013) and VerbNet (Bonial et al., 2011; Palmer et al., 2017). The framework was realized in version 3.0 of our comprehensively annotated corpus, STREUSLE[3] (Schneider et al., 2016). However, several limitations of our approach became clear to us over time.

First, as pointed out by Hwang et al. (2017), the one-label-per-token assumption in STREUSLE is flawed because it in some cases puts into conflict the semantic role of the PP with respect to a predicate, and the lexical semantics of the preposition itself. Hwang et al. (2017) suggested a solution, discussed in §3.3, but did not conduct an annotation study or release a corpus to establish its feasibility empirically. We address that gap here.

Second, 75 categories is an unwieldy number for both annotators and disambiguation systems. Some are quite specialized and extremely rare in STREUSLE 3.0, which causes data sparseness issues for supervised learning. In fact, the only published disambiguation system for preposition supersenses collapsed the distinctions to just 12 labels (Gonen and Goldberg, 2016). Hwang et al. (2017) remarked that solving the aforementioned problem could remove the need for many of the specialized categories and make the taxonomy more tractable for annotators and systems. We substantiate this here, defining a new hierarchy with just 50 categories (SNACS, §3) and providing disambiguation results for the full set of distinctions.

Finally, given the semantic overlap of possessive case and the preposition **of**, we saw an opportunity to broaden the application of the scheme to include possessives. Our reannotated corpus, STREUSLE 4.0, thus has supersense annotations for over 1000 possessive tokens that were not semantically annotated in version 3.0. We include these in our annotation and disambiguation experiments alongside reannotated preposition tokens.

3 Annotation Scheme

3.1 Lexical Categories of Interest

Apart from canonical prepositions and possessives, there are many lexically and semantically overlap-

[2] Of course, meanings marked by prepositions/possessives are to some extent captured in predicate-argument or graph-based meaning representations (e.g., Palmer et al., 2005; Fillmore and Baker, 2009; Oepen et al., 2016; Banarescu et al., 2013) and domain-centric representations like TimeML and ISO-Space (Pustejovsky et al., 2003, 2012).

[3] https://github.com/nert-gu/streusle/

ping closed-class items which are sometimes classified as other parts of speech, such as adverbs, particles, and subordinating conjunctions. *The Cambridge Grammar of the English Language* (Huddleston and Pullum, 2002) argues for an expansive definition of 'preposition' that would encompass these other categories. As a practical measure, we decided to encourage annotators to focus on the semantics of these functional items rather than their syntax, so we take an inclusive stance.

Another consideration is developing annotation guidelines that can be adapted for other languages. This includes languages which have *post*positions, *circum*positions, or *in*positions rather than prepositions; the general term for such items is *ad*positions.[4] English possessive marking (via **'s** or possessive pronouns like **my**) is more generally an example of *case* marking. Note that prepositions (4a–4c) differ in word order from possessives (4d), though semantically the object of the preposition and the possessive nominal pattern together:

(4) a. eat **in** a restaurant
b. the man **in** a blue shirt
c. the wife **of** the ambassador
d. the ambassador**'s** wife

Cross-linguistically, adpositions and case marking are closely related, and in general both grammatical strategies can express similar kinds of semantic relations. This motivates a common semantic inventory for adpositions and case.

We also cover multiword prepositions (e.g., **out_of**, **in_front_of**), intransitive particles (*He flew **away***), purpose infinitive clauses (*Open the door **to** let in some air*[5]), prepositions with clausal complements (*It rained **before** the party started*), and idiomatic prepositional phrases (***at**_large*). Our annotation guidelines give further details.

3.2 The SNACS Hierarchy

The hierarchy of preposition and possessive supersenses, which we call Semantic Network of Adposition and Case Supersenses (SNACS), is shown in figure 2. It is simpler than its predecessor—Schneider et al.'s (2016) preposition supersense hierarchy—in both size and structural complexity.

[4] In English, **ago** is arguably a postposition because it follows rather than precedes its complement: *five minutes **ago***, not **ago** *five minutes*.

[5] **To** can be rephrased as **in_order_to** and have prepositional counterparts like in *Open the door **for** some air*.

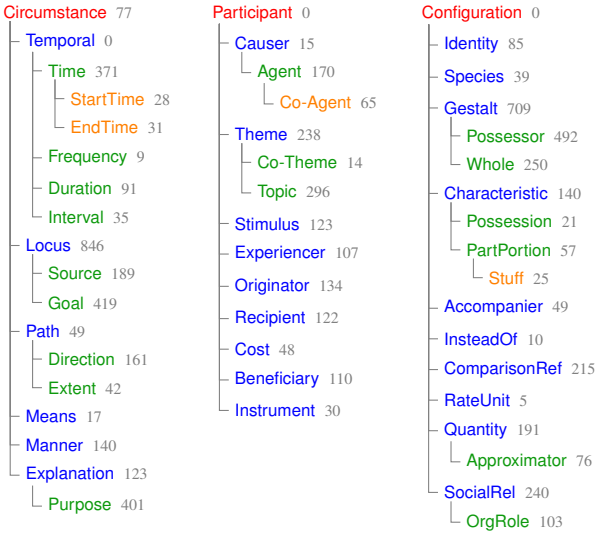

Figure 2: SNACS hierarchy of 50 supersenses and their token counts in the annotated corpus described in §4. Counts are of direct uses of labels, excluding uses of subcategories. Role and function positions are not distinguished (so if a token has different role and function labels, it will count toward two supersense frequencies).

SNACS has 50 supersenses at 4 levels of depth; the previous hierarchy had 75 supersenses at 7 levels. The top-level categories are the same:

- CIRCUMSTANCE: Circumstantial information, usually non-core properties of events (e.g., location, time, means, purpose)
- PARTICIPANT: Entity playing a role in an event
- CONFIGURATION: Thing, usually an entity or property, involved in a static relationship to some other entity

The 3 subtrees loosely parallel adverbial adjuncts, event arguments, and adnominal complements, respectively. The PARTICIPANT and CIRCUMSTANCE subtrees primarily reflect semantic relationships prototypical to verbal arguments/adjuncts and were inspired by VerbNet's thematic role hierarchy (Palmer et al., 2017; Bonial et al., 2011). Many CIRCUMSTANCE subtypes, like LOCUS (the concrete or abstract location of something), can be governed by eventive and non-eventive nominals as well as verbs: *eat **in** the restaurant*, *a party **in** the restaurant*, *a table **in** the restaurant*. CONFIGURATION mainly encompasses non-spatiotemporal relations holding between entities, such as quantity, possession, and part/whole. Unlike the previous hierarchy, SNACS does not use multiple inheritance, so there is no overlap between the 3 regions.

The supersenses can be understood as roles in fundamental types of scenes (or schemas) such as: LOCATION—THEME is located at LO-

CUS; MOTION—THEME moves from SOURCE along PATH to GOAL; TRANSITIVE ACTION—AGENT acts on THEME, perhaps using an INSTRUMENT; POSSESSION—POSSESSION belongs to POSSESSOR; TRANSFER—THEME changes possession from ORIGINATOR to RECIPIENT, perhaps with COST; PERCEPTION—EXPERIENCER is mentally affected by STIMULUS; COGNITION—EXPERIENCER contemplates TOPIC; COMMUNICATION—information (TOPIC) flows from ORIGINATOR to RECIPIENT, perhaps via an INSTRUMENT. For AGENT, CO-AGENT, EXPERIENCER, ORIGINATOR, RECIPIENT, BENEFICIARY, POSSESSOR, and SOCIALREL, the object of the preposition is prototypically animate.

Because prepositions and possessives cover a vast swath of semantic space, limiting ourselves to 50 categories means we need to address a great many nonprototypical, borderline, and special cases. We have done so in a 75-page annotation manual with over 400 example sentences (Schneider et al., 2018).

Finally, we note that the Universal Semantic Tagset (Abzianidze and Bos, 2017) defines a cross-linguistic inventory of semantic classes for content and function words. SNACS takes a similar approach to prepositions and possessives, which in Abzianidze and Bos's (2017) specification are simply tagged `REL`, which does not disambiguate the nature of the relational meaning. Our categories can thus be understood as refinements to `REL`.

3.3 Adopting the Construal Analysis

Hwang et al. (2017) have pointed out the perils of teasing apart and generalizing preposition semantics so that each use has a clear supersense label. One key challenge they identified is that the preposition itself and the situation as established by the verb may suggest different labels. For instance:

(5) a. Vernon works **at** Grunnings.
b. Vernon works **for** Grunnings.

The semantics of the *scene* in (5a, 5b) is the same: it is an employment relationship, and the PP contains the employer. SNACS has the label ORGROLE for this purpose.[6] At the same time, **at** in (5a) strongly suggests a locational relationship, which would correspond to the label LOCUS; consistent with this hypothesis, *Where does Vernon work?* is a perfectly good way to ask a question that could be answered by the PP. In this example, then, there is overlap between locational meaning and organizational-belonging meaning. (5b) is similar except the **for** suggests a notion of BENEFICIARY: the employee is working *on behalf of* the employer. Annotators would face a conundrum if forced to pick a single label when multiple ones appear to be relevant. Schneider et al. (2016) handled overlap via multiple inheritance, but entertaining a new label for every possible case of overlap is impractical, as this would result in a proliferation of supersenses.

Instead, Hwang et al. (2017) suggest a **construal analysis** in which the lexical semantic contribution, or henceforth the **function**, of the preposition itself may be distinct from the semantic role or relation mediated by the preposition in a given sentence, called the **scene role**. The notion of scene role is a widely accepted idea that underpins the use of semantic or thematic roles: semantics licensed by the governor[7] of the prepositional phrase dictates its relationship to the prepositional phrase. The innovative claim is that, in addition to a preposition's relationship with its head, the prepositional *choice* introduces another layer of meaning or **construal** that brings additional nuance, creating the difficulty we see in the annotation of (5a, 5b). Construal is notated by ROLE⇝FUNCTION. Thus, (5a) would be annotated ORGROLE⇝LOCUS and (5b) as ORGROLE⇝BENEFICIARY to expose their common truth-semantic meaning but slightly different portrayals owing to the different prepositions.

Another useful application of the construal analysis is with the verb *put*, which can combine with any locative PP to express a destination:

(6) Put it **on/by/behind/on_top_of/**... the door.
GOAL⇝LOCUS

I.e., the preposition signals a LOCUS, but the door serves as the GOAL with respect to the scene. This approach also allows for resolution of various se-

[6] ORGROLE is defined as "Either a party in a relation between an organization/institution and an individual who has a stable affiliation with that organization, such as membership or a business relationship."

[7] By "governor" of the preposition or prepositional phrase, we mean the head of the phrase to which the PP attaches in a constituency representation. In a dependency representation, this would be the head of the preposition itself or of the object of the preposition depending on which convention is used for PP headedness: e.g., the preposition heads the PP in CoNLL and Stanford Dependencies whereas the object is the head in Universal Dependencies. The governor is most often a verb or noun. Where the PP is a predicate complement (e.g. *Vernon is **with** Grunnings*), there is no governor to specify the nature of the scene, so annotators must rely on world knowledge and context to determine the scene.

	Train	Dev	Test	Total
Documents	347	192	184	723
Sentences	2,723	554	535	3,812
Tokens	44,804	5,394	5,381	55,579
Annotated targets	4,522	453	480	5,455
Role = function	3,101	291	310	3,702
P or PP	3,397	341	366	4,104
Multiword unit	256	25	24	305
Infinitive **to**	201	26	20	247
Genitive clitic (**'s**)	52	6	1	59
Possessive pronoun	872	80	93	1,045
Attested SNACS labels	47	46	44	47
Unique scene roles	46	43	41	47
Unique functions	41	38	37	41
Unique pairs	167	79	87	177
Role = function	41	33	34	41

Table 1: Counts for the data splits used in our experiments.

mantic phenomena including perceptual scenes (e.g., *I care* ***about*** *education*, where **about** is both the topic of cogitation and perceptual stimulus of caring: STIMULUS⇝TOPIC), and fictive motion (Talmy, 1996), where static location is described using motion verbiage (as in *The road runs* ***through*** *the forest*: LOCUS⇝PATH).

Both role and function slots are filled by supersenses from the SNACS hierarchy. Annotators have the option of using distinct supersenses for the role and function; in general it is not a requirement (though we stipulate that certain SNACS supersenses can only be used as the role). When the same label captures both role and function, we do not repeat it: *Vernon lives* ***in***/LOCUS *England*. Figure 1 shows some real examples from our corpus.

We apply the construal analysis in SNACS annotation of our corpus to test its feasibility. It has proved useful not only for prepositions, but also possessives, where the general sense of possession may overlap with other scene relations, like creator/initial-possessor (ORIGINATOR): *Da Vinci's*/ORIGINATOR⇝POSSESSOR *sculptures*.

4 Annotated Reviews Corpus

We applied the SNACS annotation scheme (§3) to prepositions and possessives in the STREUSLE corpus (§2), a collection of online consumer reviews taken from the English Web Treebank (Bies et al., 2012). The sentences from the English Web Treebank also comprise the primary reference treebank for English Universal Dependencies (UD; Nivre et al., 2016), and we bundle the UD version 2 syntax alongside our annotations. Table 1 shows the total number of tokens present and those that we annotated. Altogether, 5,455 tokens were annotated for scene role and function.

Rank	Role		Function	
1	LOCUS	636	LOCUS	780
2	POSSESSOR	381	GESTALT	699
⋮	⋮		⋮	
last	DIRECTION	1	POSSESSION	2

Table 2: Most and least frequent role and function labels.

The new hierarchy and annotation guidelines were developed by consensus. The original preposition supersense annotations were placed in a spreadsheet and discussed. While most tokens were unambiguously annotated, some cases required a new analysis throughout the corpus. For example, the functions of **for** were so broad that they needed to be (manually) clustered before mapping clusters onto hierarchy labels. Unusual or rare contexts also presented difficulties. Where the correct supersense remained unclear, specific instructions and examples were included in the guidelines. Possessives were not covered by the original preposition supersense annotations, and thus were annotated from scratch.[8] **Special labels** were applied to tokens deemed not to be prepositions or possessives evoking semantic relations, including uses of the infinitive marker that do not fall within the scope of SNACS (487 tokens: a majority of infinitives) and preposition-initial discourse expressions (e.g. ***after***_*all*) and coordinating conjunctions (***as***_*well_as*).[9] Other tokens requiring special labels are the opaque possessive slot in a multiword idiom (12 tokens), and tokens where unintelligble, incomplete, marginal, or nonnative usage made it impossible to assign a supersense (48 tokens).

Table 2 shows the most and least common labels occurring as scene role and function. Three labels never appear in the annotated corpus: TEMPORAL from the CIRCUMSTANCE hierarchy, and PARTICIPANT and CONFIGURATION which are both the highest supersense in their respective hierarchies. While all remaining supersenses are attested as scene roles, there are some that never occur as functions, such as ORIGINATOR, which is most often realized as POSSESSOR or SOURCE, and EXPERIENCER. It is interesting to note that every subtype of CIRCUMSTANCE (except TEMPORAL) appears as both scene role and function, whereas many of the subtypes of the other two hierarchies are lim-

[8] Blodgett and Schneider (2018) detail the extension of the scheme to possessives.

[9] In the corpus, lexical expression tokens appear alongside a **lexical category** indicating which inventory of supersenses, if any, applies. SNACS-annotated units are those with ADP (adposition), PP, PRON.POSS (possessive pronoun), etc., whereas DISC (discourse) and CCONJ expressions do not receive any supersense. Refer to the STREUSLE README for details.

ited to either role or function. This reflects our view that prepositions primarily capture circumstantial notions such as space and time, but have been extended to cover other semantic relations.[10]

5 Interannotator Agreement Study

Because the online reviews corpus was so central to the development of our guidelines, we sought to estimate the reliability of the annotation scheme on a new corpus in a new genre. We chose Saint-Exupéry's novella *The Little Prince*, which is readily available in many languages and has been annotated with semantic representations such as AMR (Banarescu et al., 2013). The genre is markedly different from online reviews—it is quite literary, and employs archaic or poetic figures of speech. It is also a translation from French, contributing to the markedness of the language. This text is therefore a challenge for an annotation scheme based on colloquial contemporary English. We addressed this issue by running 3 practice rounds of annotation on small passages from *The Little Prince*, both to assess whether the scheme was applicable without major guidelines changes and to prepare the annotators for this genre. For the final annotation study, we chose chapters 4 and 5, in which 242 markables of 52 types were identified heuristically (§6.2). The types *of*, *to*, *in*, *as*, *from*, and *for*, as well as possessives, occurred at least 10 times. Annotators had the option to mark units as false positives using special labels (see §4) in addition to expressing uncertainty about the unit.

For the annotation process, we adapted the open source web-based annotation tool UCCAApp (Abend et al., 2017) to our workflow, by extending it with a type-sensitive ranking module for the list of categories presented to the annotators.

Annotators. Five annotators (A, B, C, D, E), all authors of this paper, took part in this study. All are computational linguistics researchers with advanced training in linguistics. Their involvement in the development of the scheme falls on a spectrum, with annotator A being the most active figure in guidelines development, and annotator E not being involved in developing the guidelines and learning the scheme solely from reading the manual. Annotators A, B, and C are native speakers of English, while Annotators D and E are nonnative but highly fluent speakers.

	Labels	Role	Function
Exact	47	74.4%	81.3%
Depth-3	43	75.0%	81.8%
Depth-2	26	79.9%	87.4%
Depth-1	3	92.6%	93.9%

Table 3: Interannotator agreement rates (pairwise averages) on *Little Prince* sample (216 tokens) with different levels of hierarchy coarsening according to figure 2 ("Exact" means no coarsening). "Labels" refers to the number of distinct labels that annotators could have provided at that level of coarsening. Excludes tokens where at least one annotator assigned a non-semantic label.

Results. In the *Little Prince* sample, 40 out of 47 possible supersenses were applied at least once by some annotator; 36 were applied at least once by a majority of annotators; and 33 were applied at least once by all annotators. APPROXIMATOR, COTHEME, COST, INSTEADOF, INTERVAL, RATEUNIT, and SPECIES were not used by any annotator.

To evaluate interannotator agreement, we excluded 26 tokens for which at least one annotator has assigned a non-semantic label, considering only the 216 tokens that were identified correctly as SNACS targets and were clear to all annotators. Despite varying exposure to the scheme, there is no obvious relationship between annotators' backgrounds and their agreement rates.[11]

Table 3 shows the interannotator agreement rates, averaged across all pairs of annotators. Average agreement is 74.4% on the scene role and 81.3% on the function (row 1).[12] All annotators agree on the role for 119, and on the function for 139 tokens. Agreement is higher on the function slot than on the scene role slot, which implies that the former is an easier task than the latter. This is expected considering the definition of construal: the *function* of an adposition is more lexical and less context-dependent, whereas the *role* depends on the context (the scene) and can be highly idiomatic (§3.3).

The supersense hierarchy allows us to analyze agreement at different levels of granularity (rows

[10] All told, 41 supersenses are attested as both role and function for the same token, and there are 136 unique construal combinations where the role differs from the function. Only four supersenses are never found in such a divergent construal: EXPLANATION, SPECIES, STARTTIME, RATEUNIT. Except for RATEUNIT which occurs only 5 times, their narrow use does not arise because they are rare. EXPLANATION, for example, occurs over 100 times, more than many labels which often appear in construal.

[11] See table 7 in appendix A for a more detailed description of the annotators' backgrounds and pairwise IAA results.

[12] Average of pairwise Cohen's κ is 0.733 and 0.799 on, respectively, role and function, suggesting strong agreement. However, it is worth noting that annotators selected labels from a ranked list, with the ranking determined by preposition type. The model of chance agreement underlying κ does not take the identity of the preposition into account, and thus likely underestimates the probability of chance agreement.

2–4 in table 3; see also confusion matrix in supplement). Coarser-grained analyses naturally give better agreement, with depth-1 coarsening into only 3 categories. Results show that most confusions are local with respect to the hierarchy.

6 Disambiguation Systems

We now describe systems that identify and disambiguate SNACS-annotated prepositions and possessives in two steps. Target identification heuristics (§6.2) first determine which tokens (single-word or multiword) should receive a SNACS supersense. A supervised classifier then predicts a supersense analysis for each identified target. The research objectives are (a) to study the ability of statistical models to learn roles and functions of prepositions and possessives, and (b) to compare two different modeling strategies (feature-rich and neural), and the impact of syntactic parsing.

6.1 Experimental Setup

Our experiments use the reviews corpus described in §4. We adopt the official training/development/test splits of the Universal Dependencies (UD) project; their sizes are presented in table 1. All systems are trained on the training set only and evaluated on the test set; the development set was used for tuning hyperparameters. Gold tokenization was used throughout. Only targets with a semantic supersense analysis involving labels from figure 2 were included in training and evaluation—i.e., tokens with special labels (see §4) were excluded.

To test the impact of automatic syntactic parsing, models in the **auto syntax** condition were trained and evaluated on automatic lemmas, POS tags, and Basic Universal Dependencies (according to the v1 standard) produced by Stanford CoreNLP version 3.8.0 (Manning et al., 2014).[13] Named entity tags from the default 12-class CoreNLP model were used in all conditions.

6.2 Target Identification

§3.1 explains that the categories in our scheme apply not only to (transitive) adpositions in a very narrow definition of the term, but also to lexical items that traditionally belong to variety of syntactic classes (such as adverbs and particles), as well as possessive case markers and multiword expressions. 61.2% of the units annotated in our corpus are adpositions according to gold POS annotation, 20.2% are possessives, and 18.6% belong to other POS classes. Furthermore, 14.1% of tokens labeled as adpositions or possessives are not annotated because they are part of a multiword expression (MWE). It is therefore neither obvious nor trivial to decide which tokens and groups of tokens should be selected as targets for SNACS annotation.

To facilitate both manual annotation and automatic classification, we developed heuristics for identifying annotation targets. The algorithm first scans the sentence for known multiword expressions, using a blacklist of non-prepositional MWEs that contain preposition tokens (e.g., *take_care_of*) and a whitelist of prepositional MWEs (multiword prepositions like **out_of** and PP idioms like **in_town**). Both lists were constructed from the training data. From segments unaffected by the MWE heuristics, single-word candidates are identified by matching a high-recall set of parts of speech, then filtered through 5 different heuristics for adpositions, possessives, subordinating conjunctions, adverbs, and infinitivals. Most of these filters are based on lexical lists learned from the training portion of the STREUSLE corpus, but there are some specific rules for infinitivals that handle *for*-subjects (*I opened the door for Steve* ***to*** *take out the trash*—**to**, but not *for*, should receive a supersense) and comparative constructions with *too* and *enough* (*too short* ***to*** *ride*).

6.3 Classification

The next step of disambiguation is predicting the role and function labels. We explore two different modeling strategies.

Feature-rich Model. Our first model is based on the features for preposition relation classification developed by Srikumar and Roth (2013), which were themselves extended from the preposition sense disambiguation features of Hovy et al. (2010). We briefly describe the feature set here, and refer the reader to the original work for further details. At a high level, it consists of features extracted from selected neighboring words in the dependency tree (i.e., heuristically identified governor and object) and in the sentence (previous verb, noun and adjective, and next noun). In addition, all these features are also conjoined with the lemma of the rightmost word in the preposition token to capture

[13] The CoreNLP parser was trained on all 5 genres of the English Web Treebank—i.e., a superset of our training set. Gold syntax follows the UDv2 standard, whereas the classifiers in the auto syntax conditions are trained and tested with UDv1 parses produced by CoreNLP.

target-specific interactions with the labels. The features extracted from each neighboring word are listed in the supplementary material.

Using these features extracted from targets, we trained two multi-class SVM classifiers to predict the role and function labels using the LIBLINEAR library (Fan et al., 2008).

Neural Model. Our second classifier is a multilayer perceptron (MLP) stacked on top of a BiLSTM. For every sentence, tokens are first embedded using a concatenation of fixed pre-trained word2vec (Mikolov et al., 2013) embeddings of the word and the lemma, and an internal embedding vector, which is updated during training.[14] Token embeddings are then fed into a 2-layer BiLSTM encoder, yielding a list of token representations.

For each identified target unit u, we extract its first token, and its governor and object headword. For each of these tokens, we construct a feature vector by concatenating its token representation with embeddings of its (1) language-specific POS tag, (2) UD dependency label, and (3) NER label. We additionally concatenate embeddings of u's lexical category, a syntactic label indicating whether u is predicative/stranded/subordinating/none of these, and an indicator of whether either of the two tokens following the unit is capitalized. All these embeddings, as well as internal token embedding vectors, are considered part of the model parameters and are initialized randomly using the Xavier initialization (Glorot and Bengio, 2010). A NONE label is used when the corresponding feature is not given, both in training and at test time. The concatenated feature vector for u is fed into two separate 2-layered MLPs, followed by a separate softmax layer that yields the predicted probabilities for the role and function labels.

We tuned hyperparameters on the development set to maximize F-score (see supplementary material). We used the cross-entropy loss function, optimizing with simple gradient ascent for 80 epochs with minibatches of size 20. Inverted dropout was used during training. The model is implemented with the DyNet library (Neubig et al., 2017).

The model architecture is largely comparable to that of Gonen and Goldberg (2016), who experimented with a coarsened version of STREUSLE 3.0. The main difference is their use of unlabeled multilingual datasets to improve prediction by exploiting the differences in preposition ambiguities across languages.

[14] Word2vec is pre-trained on the Google News corpus. Zero vectors are used where vectors are not available.

Syntax	P	R	F
gold	88.8	89.6	89.2
auto	86.0	85.8	85.9

Table 4: Target identification results for disambiguation.

6.4 Results & Analysis

Following the two-stage disambiguation pipeline (i.e. target identification and classification), we separate the evaluation across the phases. Table 4 reports the precision, recall, and F-score (P/R/F) of the target identification heuristics. Table 5 reports the disambiguation performance of both classifiers with gold (left) and automatic target identification (right). We evaluate each classifier along three dimensions—role and function independently, and full (i.e. both role and function together). When we have the gold targets, we only report accuracy because precision and recall are equal. With automatically identified targets, we report P/R/F for each dimension. Both tables show the impact of syntactic parsing on quality. The rest of this section presents analyses of the results along various axes.

Target identification. The identification heuristics described in §6.2 achieve an F_1 score of 89.2% on the test set using gold syntax.[15] Most false positives (47/54=87%) can be ascribed to tokens that are part of a (non-adpositional or larger adpositional) multiword expression. 9 of the 50 false negatives (18%) are rare multiword expressions not occurring in the training data and there are 7 partially identified ones, which are counted as both false positives and false negatives.

Automatically generated parse trees slightly decrease quality (table 4). Target identification, being the first step in the pipeline, imposes an upper bound on disambiguation scores. We observe this degradation when we compare the Gold ID and the Auto ID blocks of table 5, where automatically identified targets decrease F-score by about 10 points in all settings.[16]

Classification. Along with the statistical classifier results in table 5, we also report performance

[15] Our evaluation script counts tokens that received special labels in the gold standard (see §4) as negative examples of SNACS targets, with the exception of the tokens labeled as unintelligible/nonnative/etc., which are not counted toward or against target ID performance.

[16] A variant of the target ID module, optimized for recall, is used as preprocessing for the agreement study discussed in §5. With this setting, the heuristic achieves an F_1 score of 90.2% (P=85.3%, R=95.6%) on the test set.

		Gold ID			Auto ID								
		Role	*Func.*	*Full*	*Role*			*Func.*			*Full*		
	Syntax	Acc.	Acc.	Acc.	P	R	F	P	R	F	P	R	F
Most frequent	N/A	40.6	53.3	37.9	37.0	37.3	37.1	49.8	50.2	50.0	34.3	34.6	34.4
Neural	gold	71.7	82.5	67.5	62.0	62.5	62.2	73.1	73.8	73.4	58.7	59.2	58.9
Feature-rich	gold	73.5	81.0	70.0	62.0	62.5	62.2	70.7	71.2	71.0	59.3	59.8	59.5
Neural	auto	67.7	78.5	64.4	56.4	56.2	56.3	66.8	66.7	66.7	53.7	53.5	53.6
Feature-rich	auto	67.9	79.4	65.2	58.2	58.1	58.2	66.8	66.7	66.7	55.7	55.6	55.7

Table 5: Overall performance of SNACS disambiguation systems on the test set. Results are reported for the role supersense (*Role*), the function supersense (*Func.*), and their conjunction (*Full*). All figures are percentages. *Left:* Accuracies with gold standard target identification (480 targets). *Right:* Precision, recall, and F_1 with automatic target identification (§6.2 and table 4).

for the most frequent baseline, which selects the most frequent role–function label pair given the (gold) lemma according to the training data. Note that all learned classifiers, across all settings, outperform the most frequent baseline for both role and function prediction. The feature-rich and the neural models perform roughly equivalently despite the significantly different modeling strategies.

Function and scene role performance. Function prediction is consistently more accurate than role prediction, with roughly a 10-point gap across all systems. This mirrors a similar effect in the interannotator agreement scores (see §5), and may be due to the reduced ambiguity of functions compared to roles (as attested by the baseline's higher accuracy for functions than roles), and by the more literal nature of function labels, as opposed to role labels that often require more context to determine.

Impact of automatic syntax. Automatic syntactic analysis decreases scores by 4 to 7 points, most likely due to parsing errors which affect the identification of the preposition's object and governor. In the auto ID/auto syntax condition, the worse target ID performance with automatic parses (noted above) contributes to lower classification scores.

6.5 Errors & Confusions

We can use the structure of the SNACS hierarchy to probe classifier performance. As with the interannotator study, we evaluate the accuracy of predicted labels when they are coarsened post hoc by moving up the hierarchy to a specific depth. Table 6 shows this for the feature-rich classifier for different depths, with depth-1 representing the coarsening of the labels into the 3 root labels. Depth-4 (Exact) represents the full results in table 5. These results show that the classifiers often mistake a label for another that is nearby in the hierarchy. Examining the most frequent confusions of both models, we observe that LOCUS is overpredicted (which makes sense as it is most frequent overall), and SOCIALROLE–ORGROLE and GESTALT–POSSESSOR are often confused (they are close in the hierarchy: one inherits from the other).

	Labels	Role	Function
Exact	47	67.9%	79.4%
Depth-3	43	67.9%	79.6%
Depth-2	26	76.2%	86.2%
Depth-1	3	86.0%	93.8%

Table 6: Accuracy of the feature-rich model (gold identification and syntax) on the test set (480 tokens) with different levels of hierarchy coarsening of its output. "Labels" refers to the number of labels in the training set after coarsening.

7 Conclusion

This paper introduced a new approach to comprehensive analysis of the semantics of prepositions and possessives in English, backed by a thoroughly documented hierarchy and annotated corpus. We found good interannotator agreement and provided initial supervised disambiguation results. We expect that future work will develop methods to scale the annotation process beyond requiring highly trained experts; bring this scheme to bear on other languages; and investigate the relationship of our scheme to more structured semantic representations, which could lead to more robust models. Our guidelines, corpus, and software are available at `https://github.com/nert-gu/streusle/blob/master/ACL2018.md`.

Acknowledgments

We thank Oliver Richardson, whose codebase we adapted for this project; Na-Rae Han, Archna Bhatia, Tim O'Gorman, Ken Litkowski, Bill Croft, and Martha Palmer for helpful discussions and support; and anonymous reviewers for useful feedback. This research was supported in part by DTRA HDTRA1-16-1-0002/Project #1553695, by DARPA 15-18-CwC-FP-032, and by grant 2016375 from the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel.

References

Omri Abend, Shai Yerushalmi, and Ari Rappoport. 2017. UCCAApp: Web-application for syntactic and semantic phrase-based annotation. In *Proc. of ACL 2017, System Demonstrations*, pages 109–114, Vancouver, Canada.

Lasha Abzianidze and Johan Bos. 2017. Towards universal semantic tagging. In *Proc. of IWCS*, Montpellier, France.

Adriana Badulescu and Dan Moldovan. 2009. A Semantic Scattering model for the automatic interpretation of English genitives. *Natural Language Engineering*, 15(2):215–239.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proc. of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria.

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

Austin Blodgett and Nathan Schneider. 2018. Semantic supersenses for English possessives. In *Proc. of LREC*, pages 1529–1534, Miyazaki, Japan.

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.

Melissa Bowerman and Soonja Choi. 2001. 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*, pages 475–511. Cambridge University Press, Cambridge, UK.

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.

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.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: a library for large linear classification. *Journal of Machine Learning Research*, 9(Aug):1871–1874.

Charles J. Fillmore and Collin Baker. 2009. A frames approach to semantic analysis. In Bernd Heine and Heiko Narrog, editors, *The Oxford Handbook of Linguistic Analysis*, pages 791–816. Oxford University Press, Oxford, UK.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proc. of AISTATS*, pages 249–256, Chia Laguna, Sardinia, Italy.

Hila Gonen and Yoav Goldberg. 2016. Semi supervised preposition-sense disambiguation using multilingual data. In *Proc. of COLING*, pages 2718–2729, Osaka, Japan.

Bernd Heine. 2006. *Possession: Cognitive Sources, Forces, and Grammaticalization*. Cambridge University Press, Cambridge, UK.

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

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.

Dirk Hovy, Ashish Vaswani, Stephen Tratz, David Chiang, and Eduard Hovy. 2011. Models and training for unsupervised preposition sense disambiguation. In *Proc. of ACL-HLT*, pages 323–328, Portland, Oregon, USA.

Rodney Huddleston and Geoffrey K. Pullum, editors. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

Jena D. Hwang, Archna Bhatia, Na-Rae Han, Tim O'Gorman, Vivek Srikumar, and Nathan Schneider. 2017. Double trouble: the problem of construal in semantic annotation of adpositions. In *Proc. of *SEM*, pages 178–188, Vancouver, Canada.

Naveen Khetarpal, Asifa Majid, and Terry Regier. 2009. Spatial terms reflect near-optimal spatial categories. In *Proc. of the 31st Annual Conference of the Cognitive Science Society*, pages 2396–2401, Amsterdam.

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*, revised edition. John Benjamins, Amsterdam.

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

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.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proc. of ACL: System Demonstrations*, pages 55–60, Baltimore, Maryland, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Dan Moldovan, Adriana Badulescu, Marta Tatu, Daniel Antohe, and Roxana Girju. 2004. Models for the semantic classification of noun phrases. In *HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, pages 60–67, Boston, Massachusetts, USA.

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*, pages 63–82. Walter de Gruyter, Berlin.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. DyNet: The dynamic neural network toolkit. *arXiv:1701.03980*.

Kiki Nikiforidou. 1991. The meanings of the genitive: a case study in semantic structure and semantic change. *Cognitive Linguistics*, 2(2):149–205.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: a multilingual treebank collection. In *Proc. of LREC*, pages 1659–1666, Portorož, Slovenia.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinkova, Dan Flickinger, Jan Hajic, Angelina Ivanova, and Zdenka Uresova. 2016. Towards comparability of linguistic graph banks for semantic parsing. In *Proc. of LREC*, pages 3991–3995, Paris, France.

Tom O'Hara and Janyce Wiebe. 2009. Exploiting semantic role resources for preposition disambiguation. *Computational Linguistics*, 35(2):151–184.

Martha Palmer, Claire Bonial, and Jena D. Hwang. 2017. VerbNet: Capturing English verb behavior, meaning and usage. In Susan E. F. Chipman, editor, *The Oxford Handbook of Cognitive Science*, pages 315–336. Oxford University Press.

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

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. In *IWCS-5, Fifth International Workshop on Computational Semantics*, Tilburg, Netherlands.

James Pustejovsky, Jessica Moszkowicz, and Marc Verhagen. 2012. A linguistically grounded annotation language for spatial information. *TAL*, 53(2):87–113.

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

Anette Rosenbach. 2002. *Genitive variation in English: conceptual factors in synchronic and diachronic studies*. Mouton de Gruyter, Berlin.

Patrick Saint-Dizier. 2006. PrepNet: a multilingual lexical description of prepositions. In *Proc. of LREC*, volume 6, pages 1021–1026, Genoa, Italy.

Nathan Schneider, Jena D. Hwang, Archna Bhatia, Na-Rae Han, Vivek Srikumar, Tim O'Gorman, Sarah R. Moeller, Omri Abend, Austin Blodgett, and Jakob Prange. 2018. Adposition and Case Supersenses v2: Guidelines for English. *arXiv:1704.02134*.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Meredith Green, Abhijit Suresh, Kathryn Conger, Tim O'Gorman, and Martha Palmer. 2016. A corpus of preposition supersenses. In *Proc. of LAW X – the 10th Linguistic Annotation Workshop*, pages 99–109, Berlin, Germany.

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.

Stephanie Shih, Jason Grafmiller, Richard Futrell, and Joan Bresnan. 2015. Rhythm's role in genitive construction choice in spoken English. In Ralf Vogel and Ruben van de Vijver, editors, *Rhythm in cognition and grammar: a Germanic perspective*, pages 207–234. De Gruyter Mouton, Berlin.

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. 2013. Modeling semantic relations expressed by prepositions. *Transactions of the Association for Computational Linguistics*, 1:231–242.

Leonard Talmy. 1996. Fictive motion in language and "ception". In Paul Bloom, Mary A. Peterson, Nadel Lynn, and Merrill F. Garrett, editors, *Language and Space*, pages 211–276. MIT Press, Cambridge, MA.

John R. Taylor. 1996. *Possessives in English: An Exploration in Cognitive Grammar*. Clarendon Press, Oxford, UK.

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. 2013. Automatic interpretation of the English possessive. In *Proc. of ACL*, pages 372–381, Sofia, Bulgaria.

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

Christoph Wolk, Joan Bresnan, Anette Rosenbach, and Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica*, 30(3):382–419.

Yang Xu and Charles Kemp. 2010. Constructing spatial concepts from universal primitives. In *Proc. of CogSci*, pages 346–351, Portland, Oregon.

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.

Joost Zwarts and Yoad Winter. 2000. Vector space semantics: a model-theoretic analysis of locative prepositions. *Journal of Logic, Language and Information*, 9:169–211.