Computational linking theory

Aaron Steven White   Drew Reisinger   Rachel Rudinger
Kyle Rawlins   Benjamin Van Durme
Johns Hopkins University

Abstract

A *linking theory* explains how verbs' *semantic arguments* are mapped to their *syntactic arguments*—the inverse of the *semantic role labeling* task from the *shallow semantic parsing* literature. In this paper, we develop the *computational linking theory* framework as a method for implementing and testing linking theories proposed in the theoretical literature. We deploy this framework to assess two crosscutting types of linking theory: *local* v. *global models* and *categorical* v. *featural models*. To further investigate the behavior of these models, we develop a measurement model in the spirit of previous work in *semantic role induction*: the *semantic proto-role linking model*. We use this model, which implements a generalization of Dowty's seminal proto-role theory, to induce *semantic proto-roles*, which we compare to those Dowty proposes.

1 Introduction

A *linking theory* explains how verbs' *semantic arguments* are mapped to their *syntactic arguments* (Fillmore, 1970; Zwicky, 1971; Jackendoff, 1972; Carter, 1976; Pinker, 1984, 1989; Grimshaw, 1990; Levin, 1993). For example, the verb *hit* has three semantic arguments—one for the HITTER, one for the HITTEE, and one for the hitting INSTRUMENT—and for each token of *hit*, a subset of those semantic arguments are mapped to its syntactic arguments—e.g. subject, direct object, or object of a preposition.

(1) a. [John]$_{\text{HITTER}}$ hit [the fence]$_{\text{HITTEE}}$.
  b. [The stick]$_{\text{INST}}$ hit [the fence]$_{\text{HITTEE}}$.

(2) a. #[The fence]$_{\text{HITTEE}}$ hit [John]$_{\text{HITTER}}$.
  b. #[The fence]$_{\text{HITTEE}}$ hit [the stick]$_{\text{INST}}$.

The main desideratum for selecting a linking theory is how well it explains *linking regularities*: which mappings do and do not occur. One example of a linking regularity is that HITTEE arguments cannot be mapped to subject, suggested by the fact that (1) and (2) cannot mean the same thing.

The task of constructing a linking theory that covers the entire lexicon is no small feat. One classic (though not the only) example of this difficulty concerns *psych verbs*, like *fear* and *frighten* (Lakoff, 1970; Postal, 1974; Perlmutter and Postal, 1984; Baker, 1988; Dowty, 1991; Pesetsky, 1995).

(3) a. [Mary]$_{\text{FEARER}}$ feared [John]$_{\text{FEAREE}}$.
  b. #[John]$_{\text{FEAREE}}$ feared [Mary]$_{\text{FEARER}}$.

(4) a. #[Mary]$_{\text{FEARER}}$ frightened [John]$_{\text{FEAREE}}$.
  b. [John]$_{\text{FEAREE}}$ frightened [Mary]$_{\text{FEARER}}$.

Psych verbs raise issues for theories that disallow mapping FEARER to subject, since *fear* does that, as well as those that disallow mapping FEAREE to subject, since *frighten* does that.[1]

Linking theory is intimately related to *semantic role labeling* (SRL) (Gildea and Jurafsky, 2002; Litkowski, 2004; Carreras and Marquez, 2004; Marquez et al., 2008), which is a form of *shallow semantic parsing*. Where a linking theory maps from semantic arguments to syntactic arguments, an SRL system maps from syntactic arguments to semantic

[1] See Hartshorne et al. 2015 for recent work on how children learn psych verbs' linking regularities.

arguments. Thus, SRL systems can be thought of as interpreting language, and linking theory implementations can be thought of as generating language.[2] But while much work has focused on building wide-coverage SRL systems, linking theory has not commanded similar attention.

In this paper, we introduce a framework for implementing linking theories—*computational linking theory* (CLT)—which substantially generalizes an idea first introduced by Grenager and Manning (2006). In CLT, the traditional linking theoretic notion of a mapping from the space of semantic arguments Sem to the space of syntactic arguments Syn is implemented as a classifier.

This classifier can take several forms based on the structure of Syn. For instance, following Lang and Lapata (2010), who build on Grenager and Manning 2006, Syn might be a set of syntactic positions, such as {subject, object, ...}, in which case Sem might be a set of thematic roles, such as {AGENT, PATIENT, ...}. Another possibility is that Syn is a set of syntactic position sequences such as {(subject, object), (subject, oblique), ...}, in which case Sem might similarly be a set of thematic role sequences {(AGENT, PATIENT), (AGENT, INST), ...}, and the classifier would involve *structured prediction*.

In the first part of this paper, we deploy CLT in conjunction with PropBank (Palmer et al., 2005), VerbNet (Kipper-Schuler, 2005), SemLink (Loper et al., 2007), and Reisinger et al.'s (2015) recently released Semantic Proto-Roles version 1 (SPR1) dataset to evaluate the efficacy of various linking theories proposed in the theoretical literature. In the second part, we show that CLT can be useful not only for evaluation, but also for exploratory analysis, by developing a measurement model—the *Semantic Proto-Role Linking Model* (SPROLIM)—for analyzing computational linking theories. And though our main aim is to compare and explore theoretical proposals using computational tools, we believe that those interested in *semantic role induction* (SRI) will find this measurement model useful for incorporating independent semantic annotations into SRI.

We focus on two cross-cutting types of linking theories that have been proposed in the theoretical literature: *local* v. *global models* and *categorical* v. *featural models*. The distinction between local and global models—which, as we discuss in §2, is analogous to the distinction between local and global SRL systems (cf. Toutanova et al., 2005, 2008a)—contacts a long-standing theoretical debate regarding whether semantic arguments are mapped to syntactic positions independently of other arguments (Baker, 1988) or whether there are dependencies among semantic arguments (Dowty, 1991). There is general consensus among theoreticians that this debate has been won in favor of localist theories—a consensus that we hope to break here.

The distinction between categorical and featural models contacts an independent debate as to whether semantic arguments, such as HITTER and HITTEE, fall into discrete semantic role categories, such as AGENT or PATIENT, or whether they are associated to a greater or lesser extent with fuzzy semantic role prototypes, such as PROTOAGENT and PROTOPATIENT (Dowty, 1991). Because the featural models have been far less developed in the theoretical literature—largely due to lack of good methodologies for understanding their behavior—our goal here will be to further develop measurement models for exploring featural theories.

Our main findings are:

1. Global models outperform local models (§4)
2. Categorical models outperform featural models, particularly for oblique arguments (§4)
3. Dowty's PROTOAGENT prototype is robustly discovered by our measurement model, but his PROTOPATIENT prototype appears to be a collection of multiple other prototype roles (§5)

We begin with a discussion of related work in the statistical machine translation and shallow semantic parsing literatures, and we give a brief introduction to linking theory (§2). We then describe the three datasets (PropBank, VerbNet, and SPR1) we build on to implement linking models (§3). Based on these data, we implement and test four linking models built from crossing the categorical-featural distinction with the local-global distinction, and we establish the unequivocal superiority of the global models (§4). These experiments reveal challenges faced by the featural model, which we investigate using our *Semantic Proto-Role Linking Model* (§5).

[2] See Flanigan et al. 2016 for recent work on semantics-based language generation with a looser coupling to the syntax.

2 Related work

A *linking theory* explains how verbs' *semantic arguments* are mapped to their *syntactic arguments*. Various types of theories have been proposed, differing mostly on how they define semantic roles. All of them share the feature that they predict syntactic position based on some aspect of the verb's semantics.

2.1 Predicting syntactic position

The task of predicting an argument's syntactic position based on some set of linguistic features is not a new one in computational linguistics and natural language processing (cf. Hajic et al., 2004). This problem has been particularly important in the area of *statistical machine translation* (SMT), where one needs to translate from morphologically poor languages like English to morphologically richer languages like Japanese and German (Koehn, 2005).

SMT researchers have focused for the most part on using morphological and syntactic predictors. Suzuki and Toutanova (2006, 2007) construct models for predicting Japanese morphological case (which marks syntactic position in languages that have such cases) using intralanguage positional and alignment-based features, and Jeong et al. (2010) extend this line of work to Bulgarian, Czech, and Korean. Koehn and Hoang (2007), Avramidis and Koehn (2008), and Toutanova et al. (2008b) use richer phrase-based features to do the same task.

Other approaches have incorporated semantic roles into SMT reranking components (Wu and Fung, 2009), similar to the reranking conducted in many SRL systems (cf. Gildea and Jurafsky, 2002; Pradhan et al., 2004, 2005b,a; Toutanova et al., 2005, 2008a, among others), but directly predicting syntactic position has not been explored in SMT (though see Minkov et al. 2007, who suggest using semantic role information in future work).

2.2 Semantic role labeling

A semantic role labeling (SRL) system implements the inverse of a linking theory. Where a linking theory aims to map a verb's semantic arguments to it syntactic arguments, an SRL system aims to map a verb's syntactic arguments to its semantic arguments (Gildea and Jurafsky, 2002; Litkowski, 2004; Carreras and Marquez, 2004; Marquez et al., 2008). Continuing with examples (1) and (2) from Section 1, a linking theory would need to explain why (and when) HITTERs and INSTRUMENTs, but not HITTEEs, are mapped to subject position; in contrast, an SRL system would need to label the subject position with HITTER or INSTRUMENT (or some abstraction of those roles like A0 or A2) and the object with HITTEE (or some abstraction like A1).

Local v. global models Toutanova et al. (2005, 2008a) introduce a distinction between local and global (joint) SRL models. In a local SRL model, a labeling decision is made based on only the features of the argument being labeled, while in a global system, features of the other arguments can be taken into account. The analogous distinction for a linking theory is between local linking models, which predict an argument's syntactic position based only on that argument's semantic role, and global linking models, which predict an argument's syntactic position based on its semantic role along with others'. In §4, we implement both local and global linking models for each representation of Sem we consider.

Semantic role induction Semantic role annotation is expensive, time-consuming, and hard to scale. This has led to the development of unsupervised SRL systems for *semantic role induction* (SRI). Work in SRI has tended to focus on using syntactic features to cluster arguments into semantic roles. Swier and Stevenson (2004) introduce the first such system, which uses a bootstrapping procedure to first associate verb tokens with frames containing typed slots (drawn from VerbNet), then iteratively compute probabilities based on cooccurrence counts and fill unfilled slots based on these probabilities.

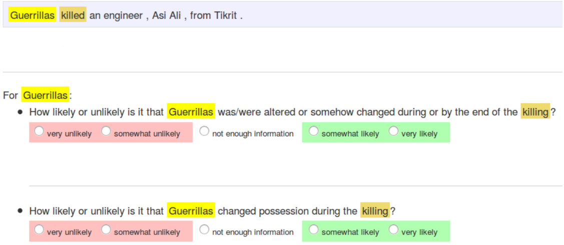
Grenager and Manning (2006) introduce the idea of predicting syntactic position based on a latent semantic role representation learned from syntactic and selectional features. Lang and Lapata (2010) expand on Grenager and Manning 2006 by introducing the notion of a *canonicalized linking*. We discuss these ideas further in §5, incorporating both into our Semantic Proto-Role Linking Model (SPROLIM).

Syntax-based clustering approaches which do not explicitly attempt to predict syntactic position have also been popular. Lang and Lapata (2011a, 2014) use graph clustering methods and Lang and Lapata (2011b) use a split-merge algorithm to cluster arguments based on syntactic context. Titov and Klementiev (2011) use a non-parametric clustering method based on the Pitman-Yor Process, and Titov and Klementiev (2012) propose two nonparametric clustering models based on the Chinese Restaurant Process (CRP) and distance dependent CRP.

2.3 Abstracting semantic roles

To predict syntactic position, linking theories aim to take advantage of linking regularities. One way theories take advantage of linking regularities is to abstract over semantic arguments in such a way that the abstractions correlate with syntactic position. Two main types of abstraction have been proposed. On the one hand are categorical theories, which group semantic arguments into a finite set of *semantic roles*—e.g., HITTERs are AGENTs, HITTEEs are PATIENTs, etc.—and then (deterministically) map these categories onto syntactic positions—e.g., subject, direct object, etc. (Fillmore, 1970; Zwicky, 1971; Jackendoff, 1972; Carter, 1976, *inter alia*; see Levin and Rappaport Hovav 2005; Williams 2015 for a review). In a categorical theory, Sem is thus some set of discrete indices, such as the core NP argument roles assumed in PropBank—i.e., $\{\text{A0}, \text{A1}, \ldots\}$ (Palmer et al., 2005)—or VerbNet—i.e., $\{\text{AGENT}, \text{PATIENT}, \ldots\}$ (Kipper-Schuler, 2005).

On the other hand are featural theories, which assign each semantic argument a set of feature values based on predicate entailments imposed on that argument. For instance, HITTERs are instigators and are thus assigned [+INSTIGATES]; they need not be volitional and are thus assigned [-VOLITIONAL]; and they are not affected by the event and are thus assigned [-AFFECTED]; in contrast, HITTEEs are [-INSTIGATES], [-VOLITIONAL], and [+AFFECTED]. A featural theory maps from (vectors of) those feature values onto syntactic positions. Thus, in a featural theory, Sem is (or is related to) some set of vectors representing some priveleged set of P entailments—e.g., $\{0, 1\}^P$, $\mathbb{R}^P$, etc.. A dataset that provides such a representation for verb-argument pairs in the Penn Treebank, which is also annotated for PropBank and (partially) for VerbNet roles, was recently made available by Reisinger et al. (2015).[3]



Guerrillas killed an engineer , Asi Ali , from Tikrit .

For Guerrillas:

- How likely or unlikely is it that Guerrillas was/were altered or somehow changed during or by the end of the killing?
  ○ very unlikely ○ somewhat unlikely ○ not enough information ○ somewhat likely ○ very likely

- How likely or unlikely is it that Guerrillas changed possession during the killing?
  ○ very unlikely ○ somewhat unlikely ○ not enough information ○ somewhat likely ○ very likely

Figure 1: Example of the SPR1 protocol

2.4 Role fragmentation

Featural theories were initially proposed to remedy certain failings of the categorical theories. In particular, reasonably wide-coverage categorical theories require an ever-growing number of roles to capture linking regularities—a phenomenon that Dowty (1991) refers to as *role fragmentation*.

In the face of role fragmentation, categorical theories have two choices: (i) use a large number of roles or (ii) force the fragmented roles into classes that predict the syntax well but map onto many distinct (possibly non-cohesive) semantic notions. These same choices must be made when developing a resource. For instance, FrameNet (Baker et al., 1998) makes the first choice; PropBank (Palmer et al., 2005) makes the second; and VerbNet (Kipper-Schuler, 2005) splits the difference.

Featural theories remedy the role fragmentation found in categorical theories by positing a small number of properties that a verb might entail about each of its arguments. The properties that hold of a particular argument determine which position it gets mapped to. Different configurations of these properties correspond to a large space of roles. For example, P binary properties generate 2^P potential roles. Dowty (1991) proposes the first (and, perhaps, most comprehensive) list of such properties as a part of specifying his *proto-role linking theory* (PRLT).

In PRLT, properties are grouped into two clusters: PROTOAGENT properties and PROTOPATIENT properties. These groupings are named this way for the fact that AGENTs in categorical theories, which tend to get mapped to subject position, tend to have PROTOAGENT properties and PATIENTs, which tend to get mapped to a non-subject position, tend to have PROTOPATIENT properties. With this idea in mind, Dowty proposed that semantic arguments

[3] `http://decomp.net`

were linked to syntactic arguments by (i) counting the number of PROTOAGENT and PROTOPATIENT properties each argument has and (iia) mapping the argument with the most PROTOAGENT properties to subject or, (iib) if there is a tie for PROTOAGENT properties, mapping the argument with the least PROTOPATIENT properties to subject.

An important thing to note before moving on is that, unlike other linking theories, which are concerned with determining more general linking phenomena Dowty concerns himself only with determining which semantic argument is mapped to subject, though he does briefly suggest that his system might extend to selecting which argument is mapped to, e.g., object v. oblique position. We assess such an extension in the experiments described below.

3 Data

For all experiments, datasets are based on the thematic role annotations of the Penn Treebank (Marcus et al., 1993) found in PropBank (Palmer et al., 2005) and VerbNet (Kipper-Schuler, 2005), mapped via SemLink (Loper et al., 2007), as well as the Semantic Proto-Roles version 1.0 (SPR1) dataset of crowd-sourced proto-role property annotations (Reisinger et al., 2015).

The SPR1 annotations consist of answers to simple questions about how likely, on a five-point scale, it is that particular relational properties hold of arguments of PropBank-annotated verbs (cp. Kako, 2006; Greene and Resnik, 2009; Hartshorne et al., 2013). Reisinger et al. constructed the questions to correspond to each of Dowty's proto-role properties. Figure 1 shows an example of the protocol.

We extracted syntactic position information about each SPR1 annotated argument by mapping Penn Treebank annotations to Universal Dependencies (De Marneffe et al., 2014; Nivre et al., 2015) using the Stanford Dependencies to Universal Dependencies converter available in `PyStanfordDependencies`.[4]

3.1 Argument and clause filtering

SPR1 contains only a subset of PTB sentences produced by applying various automated filters (see Reisinger et al. 2015 for details). All of our models, including the PropBank- and VerbNet-based models are trained on this subset—or in the case of the VerbNet-based models, a subset thereof (see §3.2). The most important of these filters for current purposes is one that retains only NP core arguments. This is useful here, since linking theories tend to only treat semantic arguments that surface as NPs. In spite of these filters, some non-NPs occur in SPR1. To remedy this, we filter all dependents that do not have a UD dependency label matching `nsubj`, `dobj`, `iobj`, or `nmod`.

Argument type	SPR1.0	VN subset
subject	5041	2357
object (direct)	2902	1307
oblique	1253	382
object (indirect)	28	18

Table 1: Count of each argument type after preprocessing

Clause type	SPR1.0	VN subset
NP V NP	2342	1144
NP V	1508	823
NP V PP+	625	221
NP V NP PP+	544	157
NP V NP NP	22	12

Table 2: Count of each clause type after preprocessing

3.2 VerbNet subset

VerbNet contains role annotations for only a subset of the SPR1 data. This has to do with the fact that only a subset of the verbs in the PTB that are annotated with PropBank rolesets are also annotated with VerbNet verb classes. A further subset of these verbs also have their arguments (whose spans are defined by PropBank) annotated with VerbNet semantic roles. Thus, there are three kinds of verbs in the corpus: those with no VerbNet annotation, those annotated only with verb class, and those with both verb class and semantic role annotations. We apply our models both to the full set of SPR1 annotations as well as the VerbNet-annotated subset.

3.3 Final datasets

Table 1 gives the counts for each syntactic position after the preprocessing steps listed above, and Ta-

[4] `https://pypi.python.org/pypi/PyStanfordDependencies`

ble 2 gives the analogous counts for each clause type. The argument type counts show only subjects (`nsubj`), direct objects (`dobj`), indirect objects (`iobj`), and obliques (`nmod`) because all other arguments are filtered out as non-core arguments.

There are few indirect objects—i.e., first direct objects in a double object construction—in this dataset. This is problematic for the cross-validation we employ in §4. To remedy this, we collapse the indirect object label to the direct object label. This is justified linguistically, since indirect objects in the sense employed in Universal Dependencies (and most other dependency parse standards) is really just a subtype of direct object.

4 Evaluating linking models

In this section, we implement categorical and featural linking models by constructing classifiers that predict the syntactic position (*subject*, *direct object*, *oblique*) of an argument based either on that argument's thematic role (e.g. AGENT, PATIENT, etc.) or on the entailments that the verb requires of that argument (e.g. INSTIGATION, VOLITION, etc.).

For each type of predictor, two linear classifiers are constructed to instantiate (i) a *local linking model* (Experiment 1), which predicts syntactic position irrespective of other arguments, and (ii) a *global linking model* (Experiment 2), which predicts syntactic position relative to other arguments. In both experiments, the PropBank- and SPR1-based models are fit to the full SPR1 dataset, and all three models are fit to the VerbNet-annotated subset.

The featural models—in particular, the global featural model—can be seen as a generalization of Dowty's proto-role model. Like Dowty's model, it groups properties based on how predictive they are of particular syntactic positions—e.g., PROTOAGENT properties are predictive of subject position. They are a generalization in two senses: (i) for Dowty, each role is weighted equally—one simply counts how many of each kind of property hold—while here, these properties can receive distinct weights (and are ordinal- rather than binary-valued); and (ii) instead of predicting only subject v. nonsubject; we attempt to also differentiate among nonsubjects—i.e., direct object v. obliques. another way of thinking about the featural models is that they, like Dowty's model, admit of role prototypes with piecewise linear boundaries in the property space.

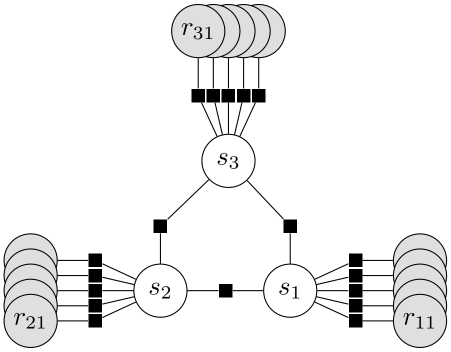


Figure 2: Factor graph for global model applied to a three argument sentence. Variable s_i represent the syntactic position of argument i. Variable $r_{ip} \equiv (a_{ip}, l_{ip})$ represents the likelihood l_{ip} that property p applies to argument i and the applicability a_{ip} of property p to arg i.

The main findings in this section are that (i) the global models substantially improve upon the local models for both categorical and featural predictors and (ii) the featural models perform worse overall than the categorical models. The first finding argues against local linking theories like that proposed by Baker (1988). We hypothesize that the second finding has two sources: (i) the set of properties in SPR1, which are essentially just Dowty's properties, is insufficient for capturing distinctions among nonsubject positions like *direct object* and *oblique*—likely because Dowty engineered his properties only to distinguish subjects from non-subjects; and (ii) because the models we use don't capture multimodality in the kinds of property configurations that exist. We explore this second possibility in §5.

4.1 Classifiers

L2-regularized maximum entropy models were used for classification in both Experiments 1 and 2. Experiment 1 uses simple logistic regression, and Experiment 2 uses a conditional random field (CRF) analogous to the logistic regression used for Experiment 1, but containing factors for each pair of arguments (see Figure 2 for a three argument example).

Experiment 1 Syntactic position $s_i \in \{$*subject*, *object*, *oblique*$\}$ was used as the dependent variable and either thematic role or property configuration as predictors. Entailment judgments from SPR1 were represented as a vector $\mathbf{l}_i$ of likelihood ratings

			PropBank			SPR1			VerbNet		
			F1	precision	recall	F1	precision	recall	F1	precision	recall
local	full	subject	0.87	1.00	0.76	0.82	0.83	0.82			
		object	0.75	0.65	0.89	0.68	0.58	0.81			
		oblique	0.73	0.71	0.76	0.09	0.53	0.05			
	subset	subject	0.89	1.00	0.81	0.85	0.87	0.83	0.88	0.90	0.87
		object	0.77	0.68	0.88	0.72	0.63	0.84	0.72	0.69	0.75
		oblique	0.65	0.61	0.70	0.07	0.40	0.04	0.56	0.60	0.54
global	full	subject	0.92	0.92	0.91	0.91	0.91	0.90			
		object	0.83	0.78	0.89	0.77	0.71	0.85			
		oblique	0.73	0.88	0.63	0.48	0.66	0.38			
	subset	subject	0.89	0.89	0.89	0.85	0.85	0.85	0.87	0.87	0.87
		object	0.81	0.76	0.88	0.77	0.73	0.82	0.78	0.73	0.83
		oblique	0.54	0.84	0.39	0.12	0.17	0.09	0.50	0.75	0.38

Table 3: Mean F1, precision, and recall on outer cross-validation test folds for local (Exp. 1) and global (Exp. 2) models on both full SPR1 dataset and the VerbNet subset. The VerbNet models were only run on the subset.

$l_{ij} \in \{1, 2, 3, 4, 5\}$ for each potential entailment j.

Ratings l_{ij} in SPR1 are furthermore associated with values $a_{ij} \in \{0, 1\}$ corresponding to the applicability of a particular entailment question. If a question i was annotated as not applicable ($a_{ij} = 0$), the combined rating r_{ij} was set to $a_{ij}l_{ij} = 0$. Because l_{ij} is strictly positive, by setting these ratings to 0, the classifier is effectively forced to project a probability from only the feature subspace corresponding to the applicable questions.

Experiment 2 The sequence of syntactic positions in each clause was used as the dependent variable. For instance, (5) would be labeled {*subj*, *obj*, *obl*}.

(5) [The bill]$_{subj}$ also **imposes** [the California auto-emissions standards]$_{object}$ on [all cars nationwide]$_{oblique}$.

These sequences were predicted using the CRF corresponding to the factor graph in Figure 2. Because the maximum number of core NP arguments and syntactic position types for any verb token is relatively small, exact inference for s is possible in our case by enumerating all configurations in the relevant cartesian product of syntactic positions $\mathcal{S}$—$\mathcal{S}^2$ for a two-argument verb, $\mathcal{S}^3$ for a three-argument verb, etc.—and computing their probabilities explicitly.[5]

4.2 Cross-validation

Nested stratified cross-validation with 10-folds at both levels of nesting was used to validate and test the models in both experiments. For each of the 10 folds in the outer CV, the L2 regularization parameter α was set using grid search over $\alpha \in \{0.01, 0.1, 1, 2, 5, 10\}$ on the 10 folds of the inner CV with highest average F1 as the selection criterion. For each of the outer folds, the model with this optimal α was refit to the full training set on that fold and tested on that fold's held-out data. In Experiment 2, a further constraint was imposed that folds not separate arguments in the same sentence.

All reported F1, precision, and recall values are computed from testing on the outer held-out sets, and all error analyses are conducted on errors when an item was found in an outer held-out set.

4.3 Results

Table 3 gives the mean F1, precision, and recall on the outer cross-validation test folds on both full SPR1 dataset and the VerbNet subset. There are two relevant patterns. First, across all predictor sets, the three measures for the global models improve substantially compared to the local models. This suggests that, regardless of one's representation of argument semantics, global models are to be preferred

[5] We suspect that this strategy will generally be possible since the vast majority of verbs have only one or two core arguments (see Table 2). Furthermore, even if the number of arguments were high on average, most configurations are not possible—e.g., one never finds two subjects. These sorts of syntax-aware constraints can significantly cut down the space of configurations that need to be considered.

over local models. It furthermore suggests that the current trend in theoretical linguistics to prefer local models likely needs to be reassessed.

Second, we find that the featural models do worse than the categorical models, particularly comparing the respective local models, but that this gap is closed to some extent when considering the respective global models. This change is particularly apparent for obliques, for which the local featural model's performance is abysmal and the global featural model's performance is middling, at least when validated on the full dataset.[6] As such, in the remainder of this section, we focus in on understanding the behavior of the global models.

Because these models are fairly simple, it is straightforward to analyze how they make their predictions by looking at their parameters. Figure 3 shows a heatmap of the mean coefficients for each global model across the 10 outer CV folds.

Turning first to the coefficients relating the syntactic positions to each other, we see that, across all models, each syntactic position disprefers occurring with a syntactic position of the same kind, and this dispreference is particularly strong for subjects. This makes sense in that we never find a sentence with two subjects, and duplicates of the other syntactic positions are relatively rare, only occurring in double object constructions or sentences with multiple prepositions. On the other hand, unlike syntactic positions attract, with strongest preference for sentences containing a non-subject to have a subject.[7]

Turning next to the categorical roles, we see that the majority of roles either prefer subject position or are agnostic, but none disprefer it. In contrast, all disprefer object to some extent. This likely arises because each of these roles can occur in intransitives, which always place their single argument in subject position.[8] This pattern also gives rise to an ordering on the roles with respect to which role will be mapped to subject position when others are present—e.g., A0 and AGENT will be mapped to subject before any other role. This is reminiscent of popular proposals from the theoretical literature regarding *role hierarchies* (see Levin and Rappaport Hovav 2005 for a review), and errors arise when that ordering is violated. For instance, because A2 does not disprefer objects as much as A1, (A1, A2) is mapped to (*object*, *subject*) by the model, when they should be mapped to (*subject*, *oblique*).

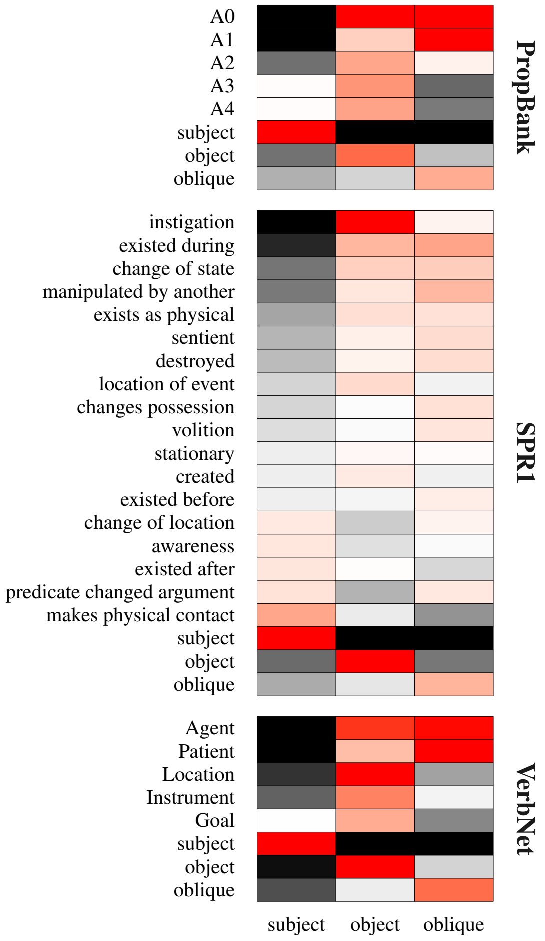


Figure 3: Heatmap of mean coefficients for global models. Black is +; red is –.

(6) [Approximately 85% of the total]$_{\mathrm{A1}}$ consisted [of nonperforming commercial real estate assets]$_{\mathrm{A2}}$.

The featural model coefficients are slightly harder to interpret. We see that being likely to INSTIGATE,

[6] There is a substantial decrement in F1 for the obliques when comparing the global featural model on the full dataset and VerbNet subset. This seems to arise from higher confusion with subject. We address this in the error analysis.

[7] This is likely due to the fact that English requires subjects; in a language that has no such requirement, we might expect a different pattern.

[8] There are languages that show, e.g., distinct case-marking behavior for different intransitives (Perlmutter, 1978; Burzio, 1986; Levin and Hovav, 1995; Hale and Keyser, 2002, see also Stevenson and Merlo 1999a,b), and so we might again expect a different pattern in those languages.

one of Dowty's PROTOAGENT properties, matters a lot for being mapped to subject position and not being mapped to object, but the rest of the relationships are quite weak and their relationship with subject position doesn't match well with the predictions of a featural theory such as Dowty's. For instance, CHANGE OF STATE is supposed to be a PROTOPATIENT property under his theory, but here it has its highest weight on subject. This appears to produce problems for many of the same sentences that the categorical models fail on—the featural model fails on 75% of the sentences the PropBank model fails on and 64% of the sentences the VerbNet model fails on—but it also produces problems for psych verbs and their kin—e.g., (7)—whose subjects are traditionally referred to as EXPERIENCERS.

(7) The real-estate market suffered even more severe setbacks.

The likely reason for this is that, as for the categorical models, the featural model must capture the fact that subjects of intransitives can be arguments that might occur in subject position in the presence of a 'better' subject. This might be further worsened by the existence of distinct clusters of properties that the model has no way of capturing. But because this representation doesn't provide a sense for which of the 6^{18} different combinations are extant, it is hard to tell what kinds of role combinations there are and, thus, which fall into each category. We could obtain a rough estimate of this by taking some statistic over the arguments that are classified as *subject*, *object*, and *oblique*, but this would fail to capture categories with multimodality in the property space. If fully spelled out categorical theories are even a good approximation, one would expect such multimodality.

4.4 Discussion

In this section, we established (i) that global models substantially improve upon the local models for both categorical and featural predictors and (ii) that featural models perform worse overall than the categorical models. In the next section, we demonstrate CLT's use as a framework for exploring linking theories—particularly, featural linking theories—by developing a measurement model that addresses the multimodality issue raised in this section.

5 Exploring linking models

In this section, we present the Semantic Proto-Role Linking Model (SPROLIM), which is a multi-view mixture model for inducing *semantic proto-roles* for each argument of a predicate from the property judgments employed in the last section. This model can be seen as a further generalization of Dowty's proto-role theory that incorporates the idea that semantic roles have a prototype structure.

In Experiment 3, we apply SPROLIM to the SPR1 data with the aim of discovering semantic protoroles. We investigate how the structure of the semantic protoroles changes as we allow for more distinct types of protoroles, finding that the one constant prototype is a PROTOAGENT role with exactly the structure proposed by Dowty. In contrast, Dowty's PROTOPATIENT role appears to rather be a collection multiple other protoroles.

5.1 Semantic Proto-Role Linking Model

There are four main components of SPROLIM. The first component is a representation of the relationship between a predicate's *l(exical)-thematic roles* (Dowty, 1989)—e.g., for the verb *hit*, the HITTER, HITTEE, and HITTING INSTRUMENT roles—and generalized *semantic proto-roles*—e.g., PROTOAGENT and PROTOPATIENT. To make clear that *l-thematic roles* are abstract, we refer to them by the more transparent name *argument types* and denote the set of argument types for a verb v with $\mathcal{A}_v$.

The second component of SPROLIM is a representation of the relationship between the semantic proto-role that an argument has and (i) the likelihood that a property is applicable to that argument and, (ii) if applicable, how likely it is that the property holds of the argument. We call this second component the *property model*. The property model represents this relationship probabilistically—i.e., each semantic proto-role is associated with a distribution over property applicability and likelihood. Thus, the property model is comparable to other mixture models implementing a *prototype theory*, and it is why we call the generalized roles semantic *proto*-roles.

The third component is a representation of the relationship between (a) the semantic proto-role that an argument (type) has and the syntactic positions an instantiation of that argument occupies in a par-

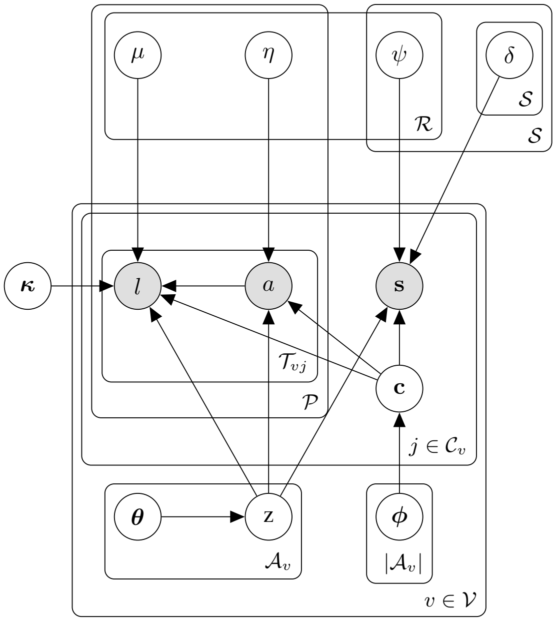


Figure 4: Plate diagram for SPROLIM

Algorithm 1 Semantic Proto-Role Linking Model

```
for verb type $v \in \mathcal{V}$ do
    for argument type $i \in \mathcal{A}_v$ do
        draw semantic protorole $z_{vi} \sim \text{Cat}(\boldsymbol{\theta}_{vi})$
    for verb token $j \in \mathcal{C}_v$ do
        draw canonicalization $k \sim \text{Cat}(\boldsymbol{\phi}_{v|\mathcal{T}_{vj}|})$
        $\mathbf{c}_{vj} \leftarrow$ element of symmetric group $S_{|\mathcal{T}_{vj}|,k}$
        let $\mathbf{r}$ : $|\mathcal{T}_{vj}|$-length tuple
        for argument token $t \in \mathcal{T}_{vj}$ do
            $r_t \leftarrow$ semantic protorole $z_{vc_{vjt}}$
            for property $p \in \mathcal{P}$ do
                draw $a_{vjt} \sim \text{Bern}(\eta_{r_{vjt}p})$
                if $a_{vjt} = 1$ then
                    draw $l_{vjt} \sim \text{Cat}(\text{Ord}_{\boldsymbol{\kappa}}(\mu_{r_tp}))$
        let $\boldsymbol{\rho}$ : $|\mathcal{S}^{|\mathcal{T}_{vj}|}|$-length vector
        for linking $\mathbf{s}' \in \mathcal{S}^{|\mathcal{T}_{vj}|}$ do
            $\rho_{\mathbf{s}'} \leftarrow \prod_t \text{softmax}\left(\psi_{r_t} + \sum_{o \neq t} \delta_{s'_t s'_o}\right)$
        draw linking $k \sim \text{Cat}(\boldsymbol{\rho})$
        $\mathbf{s}_{vj} \leftarrow \mathcal{S}_k^{|\mathcal{T}_{vj}|}$
```

ticular sentence—in our case, *subject*, *object*, and *oblique*—as well as (b) the other syntactic positions in that sentence (cf. the global model from §4). We call these instantiations *argument tokens*. Because this component determines how argument tokens are linked to syntactic positions, we refer to it as the *linking model*.[9]

The final component is a representation of the relationship between a predicate's argument tokens (in a given sentence) and that predicate's argument types. As noted by Lang and Lapata (2010), such a component is necessary to handle argument alternations like passivization (8), the double object alternation (9), and the causative-inchoative alternation (10). If we relied purely on, e.g., relative position to associate argument tokens with argument types, we would systematically make mistakes on such cases.

(8) a. Eight Oakland players hit homers
b. Homers were hit by eight Oakland players.

(9) a. Some 46% give foreign cars higher quality ratings.
b. Some 46% give higher quality ratings to foreign cars.

(10) a. The earthquake shattered windows at SFO's air-traffic control tower.
b. Windows at SFO's air-traffic control tower shattered.

Following Lang and Lapata, we refer to this component as the *canonicalizer*, and we refer to these final two components together as the *mapping model*, since they define how one maps from argument tokens to argument types (labeled with semantic proto-roles), and from semantic proto-roles to syntactic positions.

In the remainder of this section, we define the property and mapping models more formally then fit SPROLIM to our data. To guide the description, Algorithm 1 gives SPROLIM's generative story, and Figure 4 shows the corresponding plate diagram.

Property model The property model relates each semantic role to (i) the likelihood that a property is applicable to an argument that has that role and, (ii) if applicable, how likely it is that the property holds of that argument.

We implement this model using a cumulative link logit hurdle model (see Agresti, 2014).[10] In this model, each semantic role $r \in \mathcal{R}$ is associated with two $|\mathcal{P}|$-length real-valued vectors: a real-valued vector $\boldsymbol{\mu}_r$, which corresponds to the likelihood of

[9] Unless specifically noted, this is what we mean by *linking model*. When referring to the entire model, we use SPROLIM.

[10] Cumulative link logit models are standard in the analysis of ordinal judgments like those contained in SPR1.

each property $p \in \mathcal{P}$ when an argument has role r, and $\boldsymbol{\eta}_k$, which gives the probability that each property p is applicable to an argument that has role r. We first describe the cumulative link logit part of this model, which determines the probability of each likelihood rating, and then the hurdle part, which determines the probability that a particular property is applicable to a particular argument.

In the cumulative link logit portion of the model, a categorical probability mass function with support on the property likelihood ratings $l \in \{1, \ldots, 5\}$ is determined by the latent value μ and a nondecreasing real-valued cutpoint vector $\boldsymbol{\kappa}$.

$$\mathbb{P}(l = j \mid \mu, \boldsymbol{\kappa}) = \begin{cases} 1 - q_{j-1} & \text{if } j = 5 \\ q_j - q_{j-1} & \text{otherwise} \end{cases}$$

where $q_j \equiv \text{logit}^{-1}(\kappa_{j+1} - \mu)$ and $q_0 \equiv 0$. This model is known as a cumulative link logit model, since $\mathbf{q}$ is a valid cumulative distribution function for a categorical random variable with support on $\{1, \ldots, 5\}$. In Algorithm 1, we denote the parameters of this distribution with $\text{Ord}_{\boldsymbol{\kappa}}(\mu)$.

In the hurdle portion of the model, a Bernoulli probability mass function for applicability $a \in \{0, 1\}$ is given by $\mathbb{P}(a \mid \eta) = \eta^a(1 - \eta)^{1-a}$. What makes this model a hurdle model is that the rating probability only kicks in if the rating crosses the applicability "hurdle." The procedural way of thinking about this is that, first, a rater decides whether a property is applicable; if it is not, they stop; if it is, they generate a rating. The joint probability of l and a is then defined as

$$\begin{aligned} \mathbb{P}(l, a \mid \mu, \eta, \boldsymbol{\kappa}) &= \mathbb{P}(a \mid \eta)\mathbb{P}(l \mid a, \mu, \boldsymbol{\kappa}) \\ &\propto \mathbb{P}(a \mid \eta)\mathbb{P}(l \mid \mu, \boldsymbol{\kappa})^a \end{aligned}$$

This has the effect that the value of μ is estimated from only (l, a)-pairs where a=1—i.e., where the property was applicable to the argument.

Mapping model The mapping model defines how to map from argument tokens to argument types (labeled with semantic proto-roles), and from semantic proto-roles to syntactic positions. There are two components of this model: (i) the canonicalizer, which maps from argument tokens to argument types, and (ii) the linking model, which maps from argument types (labeled with semantic proto-roles) to syntactic positions.

We implement the canonicalizer by assuming that, for each predicate (verb) v, there is some canonical ordering of its argument types and that for each sentence (clause) $j \in \mathcal{C}_v$ that v occurs in, there is some permutation of v's argument tokens in that sentence that aligns them with their argument type in the canonical order. Denoting the set of argument tokens in sentence j with $\mathcal{T}_{vj}$, the set of possible mappings is the symmetric group $S_{|\mathcal{T}_{vj}|}$. We place a categorical distribution with parameter $\boldsymbol{\phi}_v$ on the elements of this group.

We implement the linking model using the same CRF described in §4, but replacing the factors for each property with a factor the the arguments semantic proto-role. Thus, in Figure 2, which gives the factor graph for this CRF, the factors linking the responses $r = (a, l)$ directly to the syntactic position nodes s are replaced with factors linking the semantic roles z to those syntactic positions.

5.2 Experiment 3

In this experiment, we fit SPROLIM to the SPR1 data and investigate the semantic protoroles it learns.

Model fitting We use projected gradient descent with AdaGrad (Duchi et al., 2011) to find an approximation to the maximum likelihood estimates (MLE) for $\boldsymbol{\Theta}$, $\boldsymbol{\Phi}$, $\mathbf{M}$, $\mathbf{E}$, $\boldsymbol{\Psi}$, $\boldsymbol{\Delta}$, and $\boldsymbol{\kappa}$, with the variables $\mathbf{Z}$ and $\mathbf{C}$ integrated out of the likelihood.

Determining a number of protoroles The one free parameter that we must set prior to fitting SPROLIM is the number of semantic protoroles $|\mathcal{R}|$. We are interested in the model's behavior as $|\mathcal{R}|$ increases but we cannot investigate the results for all possible values here. To cut down on this set, we use a stopping criterion based on the Akaike Information Criterion (AIC). We fit SPROLIM with increasing values of $|\mathcal{R}|$, stopping when AIC is minimized. We find that the $|\mathcal{R}|$ that maximizes AIC is 6.

Results Figure 5 shows the estimates of the property likelihood centroids $\mathbf{L}$ (top) and the role-syntax coefficients $\boldsymbol{\Psi}$ (bottom) for each value of $|\mathcal{R}|$ fit.[11] Columns give the values for a single protorole.

[11] The syntax-syntax coefficients show the same pattern seen among the corresponding coefficients represented in Figure 3.

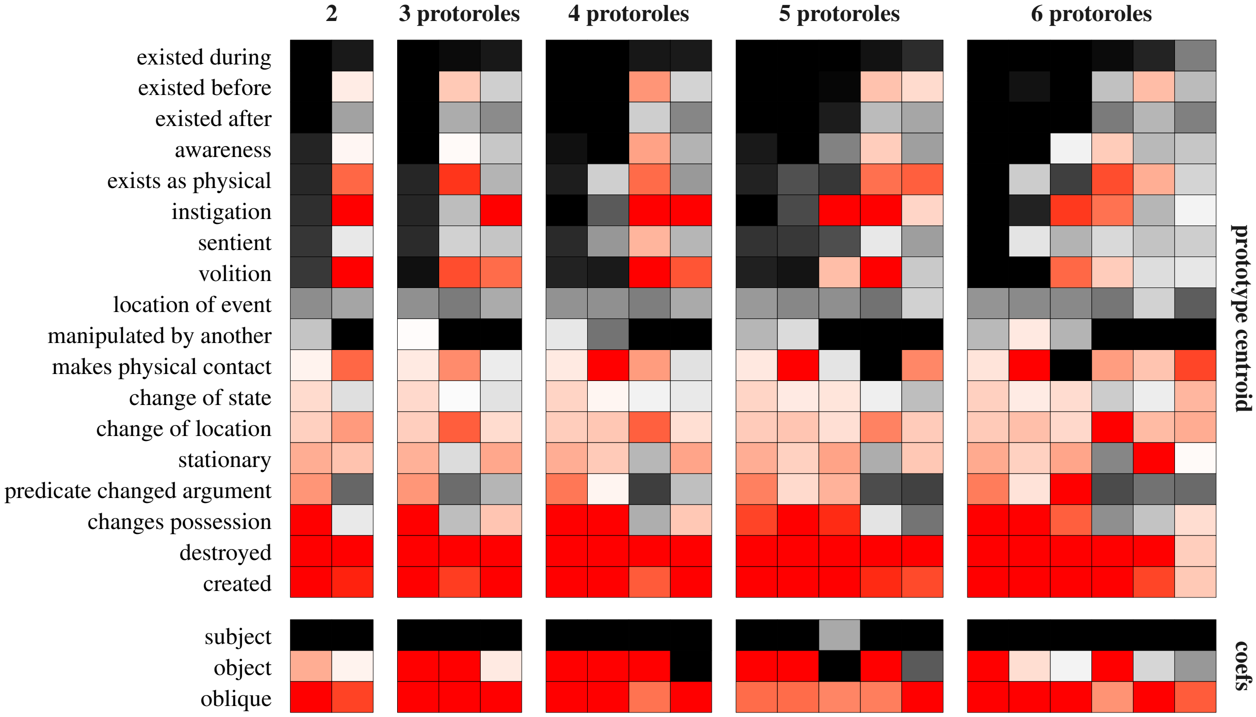


Figure 5: Heatmap of prototype centroids for likelihood ratings and coefficients for each role. Black is +; red is –.

Perhaps the most striking aspect of this figure is that, at each value of $|\mathcal{R}|$, we see a nearly identical protorole with strong positive values on exactly Dowty's PROTOAGENT properties and negatively (or near zero) on his PROTOPATIENT properties. As such, we refer to this role, which is always the first column, as the PROTOAGENT role.

The rest of the roles are more varied. For $|\mathcal{R}| \in \{2, 3\}$, the non-PROTOAGENT role loads negatively (or near zero) on all PROTOAGENT properties, and really, all other properties besides MANIPULATED BY ANOTHER. Comparing $|\mathcal{R}| = 2$ and $|\mathcal{R}| = 3$, it appears that, the non-PROTOAGENT role in $|\mathcal{R}| = 2$ is split in two based on INSTIGATION and EXISTS AS PHYSICAL, where the protorole that disprefers INSTIGATION is more likely to be an object.

Moving to $|\mathcal{R}| = 4$, we see the addition of what looks to be a second PROTOAGENT role with fewer PROTOAGENT properties. Upon investigation of the protorole mixtures $\mathbf{\Theta}$ for each argument, this appears to capture cases of nonsentient or abstract—but still relatively agentive—subjects, as in (11).

(11) The antibody then kills the cell.

This same protorole appears in $|\mathcal{R}| = 5$ with nearly all the PROTOAGENT properties, but dispreferring MAKES PHYSICAL CONTACT. It also appears in $|\mathcal{R}| = 6$ without SENTIENT or EXISTS AS PHYSICAL, bolstering its status as capturing abstract entities—e.g., corporations.

6 Conclusion

In this paper, we introduced a framework for *computational linking theory* (CLT) and deployed this framework for two distinct purposes: evaluation (§4) and exploration (§5). In §4, we evaluated four linking models based in theoretical proposals: local v. global linking models and categorical v. featural linking models. We found that global models outperform local models and categorical models outperform featural models. In §5, we developed the Semantic Proto-Role Linking Model in order to better understand how the property space employed in the featural models relates to the syntax. In investigating this model's behavior, we noted that it finds a protorole strikingly similar to the one proposed by Dowty (1991), but that others of Dowty's protoroles fall into multiple distinct prototypes.

References

Alan Agresti. *Categorical data analysis*. John Wiley & Sons, 2014. ISBN 1118710851.

Eleftherios Avramidis and Philipp Koehn. Enriching Morphologically Poor Languages for Statistical Machine Translation. In *Proceedings of ACL-08: HLT*, pages 763–770, 2008.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley Framenet Project. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pages 86–90. Association for Computational Linguistics, 1998.

Mark C. Baker. *Incorporation: A theory of grammatical function changing*. University of Chicago Press Chicago, 1988. ISBN 0226035425.

Luigi Burzio. *Italian syntax: A government-binding approach*, volume 1. Springer Science & Business Media, 1986. ISBN 90-277-2014-2.

Xavier Carreras and Llus Marquez. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 152–164. Association for Computational Linguistics, 2004.

Richard Carter. Some linking regularities. In *On Linking: Papers by Richard Carter*, Lexicon Project Working Papers (Vol. 25). MIT Center for Cognitive Science, Cambridge, MA, 1976.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of LREC*, volume 14, pages 4585–4592, 2014.

David Dowty. Thematic proto-roles and argument selection. *Language*, 67(3):547–619, 1991.

David R. Dowty. On the semantic content of the notion of thematic role. In *Properties, types and meaning*, pages 69–129. Springer, 1989.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

Charles John Fillmore. The grammar of hitting and breaking. In R.A. Jacobs and P.S. Rosenbaum, editors, *Readings in English Transformational Grammar*, pages 120–133. Ginn, Waltham, MA, 1970.

Jeffrey Flanigan, Chris Dyer, Noah A. Smith, and Jaime Carbonell. Generation from Abstract Meaning Representation using Tree Transducers. In *Proceedings of NAACL 2016*, 2016.

Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

Stephan Greene and Philip Resnik. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511. Association for Computational Linguistics, 2009. ISBN 1-932432-41-8.

Trond Grenager and Christopher D. Manning. Unsupervised discovery of a statistical verb lexicon. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 1–8. Association for Computational Linguistics, 2006. ISBN 1-932432-73-6.

Jane Grimshaw. *Argument structure*. MIT Press, Cambridge, MA, 1990. ISBN 0262071258.

Jan Hajic, Yuan Ding, Dan Gildea, Gerald Penn, and Dragomir Radevlo. Natural Language Generation in the Context of Machine Translation. 2004.

Ken Hale and Samuel Jay Keyser. *Prolegomena to a Theory of Argument Structure*. MIT Press, Cambridge, MA, 2002.

Joshua K. Hartshorne, Claire Bonial, and Martha Palmer. The VerbCorner Project: Toward an Empirically-Based Semantic Decomposition of Verbs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1438–1442, 2013.

Joshua K. Hartshorne, Amanda Pogue, and Jesse Snedeker. Love is hard to understand: the relationship between transitivity and caused events in the acquisition of emotion verbs. *Journal of Child Language*, 42(03):467–504, 2015.

Ray Jackendoff. *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, MA, 1972. ISBN 0-262-10013-4.

Minwoo Jeong, Kristina Toutanova, Hisami Suzuki, and Chris Quirk. A discriminative lexicon model for complex morphology. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*, 2010.

Edward Kako. Thematic role properties of subjects and objects. *Cognition*, 101(1):1–42, 2006.

Karin Kipper-Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, 2005.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, volume 5, pages 79–86, 2005.

Philipp Koehn and Hieu Hoang. Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 868–876, 2007.

George Lakoff. *Irregularity in syntax*. Holt, Rinehart, and Winston, 1970.

Joel Lang and Mirella Lapata. Unsupervised induction of semantic roles. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 939–947. Association for Computational Linguistics, 2010. ISBN 1-932432-65-5.

Joel Lang and Mirella Lapata. Unsupervised semantic role induction with graph partitioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1320–1331. Association for Computational Linguistics, 2011a. ISBN 1-937284-11-5.

Joel Lang and Mirella Lapata. Unsupervised semantic role induction via split-merge clustering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1117–1126. Association for Computational Linguistics, 2011b. ISBN 1-932432-87-6.

Joel Lang and Mirella Lapata. Similarity-driven semantic role induction via graph partitioning. *Computational Linguistics*, 40(3):633–669, 2014.

Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press, 1993. ISBN 0226475336.

Beth Levin and Malka Rappaport Hovav. *Unaccusativity: At the syntax-lexical semantics interface*, volume 26. MIT press, 1995. ISBN 0-262-62094-4.

Beth Levin and Malka Rappaport Hovav. *Argument realization*. Cambridge University Press, 2005. ISBN 0-521-66376-8.

Ken Litkowski. Senseval-3 task: Automatic labeling of semantic roles. *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 1:141–146, 2004.

Edward Loper, Szu-Ting Yi, and Martha Palmer. Combining lexical resources: mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Linguistics*, Tilburg, the Netherlands, 2007.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2): 313–330, June 1993. ISSN 0891-2017. URL `http://dl.acm.org/citation.cfm?id=972470.972475`.

Lluis Marquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. Semantic role labeling: an introduction to the special issue. *Computational linguistics*, 34(2):145–159, 2008.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. Generating complex morphology for machine translation. In *ACL*, volume 7, pages 128–135, 2007.

Joakim Nivre, Zeljko Agic, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Cristina Bosco, Sam Bowman, Giuseppe G. A. Celano, Miriam Connor, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Toma Erjavec, Richrd Farkas, Jennifer Foster, Daniel Galbraith, Filip Ginter, Iakes Goenaga, Koldo

Gojenola, Yoav Goldberg, Berta Gonzales, Bruno Guillaume, Jan Haji, Dag Haug, Radu Ion, Elena Irimia, Anders Johannsen, Hiroshi Kanayama, Jenna Kanerva, Simon Krek, Veronika Laippala, Alessandro Lenci, Nikola Ljubei, Teresa Lynn, Christopher Manning, Ctlina Mrnduc, David Mareek, Hctor Martnez Alonso, Jan Maek, Yuji Matsumoto, Ryan McDonald, Anna Missil, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Shunsuke Mori, Hanna Nurmi, Petya Osenova, Lilja vrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Prokopis Prokopidis, Sampo Pyysalo, Loganathan Ramasamy, Rudolf Rosa, Shadi Saleh, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simk, Kiril Simov, Aaron Smith, Jan tpnek, Alane Suhr, Zsolt Sznt, Takaaki Tanaka, Reut Tsarfaty, Sumire Uematsu, Larraitz Uria, Viktor Varga, Veronika Vincze, Zdenk abokrtsk, Daniel Zeman, and Hanzhi Zhu. Universal Dependencies 1.2. *http://universaldependencies.github.io/docs/*, November 2015. URL `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1548`.

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

David Perlmutter and Paul Postal. The 1-advancement exclusiveness law. *Studies in relational grammar*, 2(81):125, 1984.

David M. Perlmutter. Impersonal passives and the unaccusative hypothesis. In *annual meeting of the Berkeley Linguistics Society*, volume 4, pages 157–190, 1978. ISBN 2377-1666.

David Pesetsky. *Zero Syntax: Experiencers and Cascades*. MIT Press, 1995. ISBN 0-262-66100-4.

Steven Pinker. *Language Learnability and Language Development*. Harvard University Press, 1984. ISBN 0-674-04217-4.

Steven Pinker. *Learnability and Cognition: The Acquisition of Argument Structure*. MIT Press, Cambridge, MA, 1989. ISBN 0-262-51840-6.

Paul Martin Postal. *On raising: one rule of English grammar and its theoretical implications*. Current Studies in Linguistics. MIT Press, Cambridge, MA, 1974. ISBN 0262160579.

Sameer Pradhan, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James H. Martin, and Daniel Jurafsky. Support vector learning for semantic argument classification. *Machine Learning*, 60(1-3): 11–39, 2005a.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Daniel Jurafsky. Semantic role labeling using different syntactic views. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 581–588. Association for Computational Linguistics, 2005b.

Sameer S. Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Daniel Jurafsky. Shallow Semantic Parsing using Support Vector Machines. In *HLT-NAACL*, pages 233–240, 2004.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. Semantic Proto-Roles. *Transactions of the Association for Computational Linguistics*, 3:475–488, 2015.

Suzanne Stevenson and Paola Merlo. Automatic verb classification using distributions of grammatical features. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 45–52. Association for Computational Linguistics, 1999a.

Suzanne Stevenson and Paola Merlo. Automatic verb classification using distributions of grammatical features. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 45–52. Association for Computational Linguistics, 1999b.

Hisami Suzuki and Kristina Toutanova. Learning to predict case markers in japanese. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1049–1056. Association for Computational Linguistics, 2006.

Kristina Toutanova Hisami Suzuki and K. Toutanova. Generating case markers in machine translation. In *Proceedings of NAACL HLT*, pages 49–56, 2007.

Robert S. Swier and Suzanne Stevenson. Unsupervised semantic role labelling. In *Proceedings of EMNLP*, volume 95, page 102, 2004.

Ivan Titov and Alexandre Klementiev. A Bayesian model for unsupervised semantic parsing. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1445–1455. Association for Computational Linguistics, 2011. ISBN 1-932432-87-6.

Ivan Titov and Alexandre Klementiev. Crosslingual induction of semantic roles. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, volume 1, pages 647–656. Association for Computational Linguistics, 2012.

Kristina Toutanova, Aria Haghighi, and Christopher D. Manning. Joint learning improves semantic role labeling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 589–596. Association for Computational Linguistics, 2005.

Kristina Toutanova, Aria Haghighi, and Christopher D. Manning. A global joint model for semantic role labeling. *Computational Linguistics*, 34(2):161–191, 2008a.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. Applying Morphology Generation Models to Machine Translation. In *Proceedings of ACL-08: HLT*, pages 514–522, 2008b.

Alexander Williams. *Arguments in Syntax and Semantics*. Cambridge University Press, 2015. ISBN 1316239470.

Dekai Wu and Pascale Fung. Semantic roles for smt: a hybrid two-pass model. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 13–16. Association for Computational Linguistics, 2009.

Arnold M. Zwicky. In a manner of speaking. *Linguistic Inquiry*, 2(2):223–233, 1971.