# Computational linking theory

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

Submitted to *Transactions of the ACL*

## 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 propose the *computational linking theory* (CLT) framework and use it to implement two classic *linking theories*. We then propose the *semantic proto-role linking model* (SPROLIM), which implements a generalization of Dowty's seminal proto-role theory. We use SPROLIM to induce *natural semantic roles* that we evaluate using the CLT framework. We show that SPROLIM provides a viable theoretically motivated alternative to previous work in *semantic role induction*.

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

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

* Corresponding author: aswhite@jhu.edu

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

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

Psych verbs raise issues for theories that disallow mapping the FEARER to subject position, since *fear* does that, as well as those that disallow mapping the FEAREE to subject position, since *frighten* does that. (See Hartshorne et al. 2015 for recent work on how children learn psych verbs' linking regularities.)

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 arguments. Thus, SRL systems can be thought of as interpreting/comprehending language, and linking theory implementations can be thought of as gener-

ating/producing language. 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 Lang and Lapata (2010). In CLT, the traditional linking theoretic notion of a mapping $L$ from the space of semantic arguments $Sem$ to the space of syntactic arguments $Syn$ is implemented as a classifier.

$$L : Sem \longrightarrow Syn$$

This classifier can take several forms based on the structure of $Syn$. For instance, following Lang and Lapata, $Syn$ 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, INSTRUMENT), ...}, and $L$ would involve *structured prediction*.

In this paper, we deploy CLT in conjunction with Reisinger et al.'s (2015) recently released Semantic Proto-Roles version 1.0 (SPR1.0) dataset to assess the efficacy of Dowty's (1991) thematic proto-role theory as applied to naturally occurring text. Beyond the theoretical interest, a major reason for carrying out this assessment is to test the efficacy of inducing *natural semantic roles* from purely semantic information. This is a substantial contribution to the literature on *semantic role induction*, which has so far relied on purely syntactic information.

We consider two implementations of Dowty's theory: a direct implementation of the traditional proto-role model, which assumes only two proto-roles, and a generalized version, which allows for more than two roles. We compare these implementations to models based on more traditional categorical roles, like those annotated in PropBank and VerbNet. PropBank in particular provides a useful comparison, since the PropBank roles that correspond most closely to AGENT (A0) and PATIENT (A1) align closely with the syntax, and thus the PropBank models' performance stands as a sort of upper bound on how well a proto-role model might do.

Our main findings are:

1. The traditional categorical models outperform the direct implementation (Section 4)
2. The generalized model performs comparably to the PropBank model and outperforms the VerbNet model (Section 5)
3. The natural semantic roles induced by the generalized model are not just PropBank roles in disguise (Section 5)

We begin in Section 2 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, focusing in particular on the distinction between *categorical theories* and *featural theories*. In Section 3, we describe the three datasets (PropBank, VerbNet, and SPR1.0) we build on to implement categorical and featural linking theories. In Section 4, we implement these theories using both local and global semantic representations, analogous to the local v. global (joint) distinction found in the SRL literature. In Section 5, we generalize the model used for the SPRL data in order to solve these problems, implementing a multiview mixture model for inducing *natural semantic roles* (NSRs). In Section 6, we conclude.

## 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 Sections 4 and 5, 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.

Lang and Lapata (2010) introduce the idea of predicting syntactic position based on a latent semantic role representation learned from other syntactic features. We employ a similar multi-view model for inducing natural semantic roles in Section 5.

Syntax-based clustering approaches which do not necessarily 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. Gormley et al. (2014) propose a model that is based on syntactic features that themselves are latent.

### 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 [-CAUSE], [-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).

### 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).

| 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

In PRLT, properties are grouped into two clusters: *proto-agent properties* and *proto-patient 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 proto-agent properties and PATIENTs, which tend to get mapped to a non-subject position, tend to have proto-patient properties—cf. examples (1) and (2).

There are two ways one might model proto-roles. On the one hand, one might associate each property with a weight, where the sign of that weight corresponds to that property's proto-role. We implement this model in Section 4. One problem with this approach is that it cannot be easily extended to situations where more than two proto-roles are necessary (Grimm, 2011). For instance, one may need a proto-role corresponding to INSTRUMENTs (cf. Rissman et al., 2015). We capture this intuition in Section 5 by implementing a generalization of Dowty's theory.

## 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.,

| | | | PropBank | | | SPR1.0 | | | VerbNet | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | F1 | prec. | rec. | F1 | prec. | rec. | F1 | prec. | rec. |
| local | full | subject | 0.87 | 1.00 | 0.77 | 0.84 | 0.86 | 0.82 | | | |
| | | non-subject | 0.87 | 0.77 | 0.99 | 0.79 | 0.77 | 0.81 | | | |
| | subset | subject | 0.87 | 0.96 | 0.80 | 0.84 | 0.87 | 0.81 | 0.85 | 0.82 | 0.87 |
| | | non-subject | 0.86 | 0.78 | 0.95 | 0.79 | 0.77 | 0.83 | 0.78 | 0.81 | 0.75 |
| global | full | reversed | 0.99 | 0.99 | 0.99 | 0.94 | 0.93 | 0.94 | | | |
| | | true | 0.99 | 0.99 | 0.99 | 0.93 | 0.94 | 0.93 | | | |
| | subset | reversed | 0.99 | 0.99 | 0.99 | 0.94 | 0.94 | 0.94 | 0.97 | 0.97 | 0.97 |
| | | true | 0.99 | 0.99 | 0.99 | 0.94 | 0.94 | 0.94 | 0.97 | 0.97 | 0.97 |

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 SPR1.0 dataset and the VerbNet subset. The VerbNet models were only run on the subset.

2005) and VerbNet (Kipper-Schuler, 2005), mapped via SemLink (Loper et al., 2007), as well as the Semantic Proto-Roles version 1.0 (SPR1.0) dataset of entailment judgments (Reisinger et al., 2015). Syntactic position was extracted from the Universal Dependencies on the Penn Treebank (De Marneffe et al., 2014) using the Stanford Dependencies to Universal Dependencies converter available in `PyStanfordDependencies`.

### 3.1 Argument and clause filtering

SPR1.0 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 Section 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 SPR1.0. To remedy this, we filter all dependents that do not have a UD dependency label matching `nsubj`, `dobj`, `iobj`, or `nmod`.

A further 100 clauses are removed because they contained no subject. These tend to be participial phrases with only a `dobj` or `nmod` dependent.

### 3.2 VerbNet subset

VerbNet contains role annotations for only a subset of the SPR1.0 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 only use arguments with both kinds of annotations in our models, which reduces the size of our dataset for the VerbNet-based models.

## 4 Traditional linking models

In this section, categorical and featural linking models are implemented by constructing classifiers that predict the syntactic position (*subject* v. *non-subject*) 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 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 SPR1.0-based models are fit to the full SPR1.0 dataset, and all three models are fit to the VerbNet annotated subset of SPR1.0.

The featural model can be seen as a generalization of Dowty's proto-role model. Like Dowty's model, it groups properties into properties that are predictive of subject position—cp. proto-agent

| | | PB | SPR1.0 | VN |
|---|---|---|---|---|
| local | subject | 0.82 | 0.80 | 0.81 |
| | object (direct) | 0.75 | 0.73 | 0.67 |
| | object (indirect) | 0.84 | 0.71 | 0.73 |
| | oblique | 0.85 | 0.68 | 0.71 |
| global | subject | 0.96 | 0.91 | 0.96 |
| | object (direct) | 0.96 | 0.90 | 0.95 |
| | object (indirect) | 0.99 | 0.84 | 1.00 |
| | oblique | 0.97 | 0.86 | 0.94 |

Table 4: Mean probability of true syntactic position (*subject* v. *non-subject*).

properties—and properties that are predictive of non-subject position—cp. proto-patient properties. They are a generalization in the sense that, 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).

### 4.1 Classifier

An L2-regularized maximum entropy model was used for classification in both Experiments 1 and 2.

**Experiment 1** Syntactic position (*subject* v. *non-subject*) was used as the dependent variable and either thematic role or property configuration as predictors. Entailment judgments from SPR1.0 were represented as a vector $\mathbf{x}$ of likelihood ratings $x_i \in \{1, 2, 3, 4, 5\}$ for each potential entailment $i$.

Entailment ratings in SPR1.0 are furthermore associated with values corresponding to the applicability of a particular entailment question. If a question was annotated as not applicable, the rating was set to 0. 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 (the *applicable subspace*).

**Experiment 2** Instead of classifying arguments for their syntactic position (*subject* v. *non-subject*), we aim to classify what we refer to as *mapping type*: whether a particular sequence of predictors represents an observed mapping from semantic argument to syntactic position, or one that was unobserved.

To do this, we concatenate the predictors for subject and non-subject for each predicate in the data with either subject predictors first or non-subject predictors first. We call the resulting subject-first vectors the *true* mapping type and the resulting non-subject-first predictors the *reversed* mapping type.

### 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 $\alpha$ 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 $\alpha$ 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 that folds should not split two datapoints created from concatenating the same two argument predictor vectors was imposed. (If this constraint were not enforced, information from the training set would likely leak into the test set.)

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 SPR1.0 dataset and the VerbNet subset. We see that, in Experiment 1, the PropBank model outperforms the SPR1.0 and VerbNet models by 4-5 points in macro F1 (mean across subject and non-subject F1). Further, across all predictor sets, all three measures improve compared to the local models. This is to be expected because the global models get a superset of the information that the local models do.

All models do at least as well or better (in terms of F1) on classifying subjects compared to non-subjects. The featural model in particular tends to do much worse on the latter case compared to the former. This may have to do with the fact that non-subjects are a more heterogenous group, mapping to different syntactic positions—e.g. direct object and object of a preposition (oblique).

Table 4 shows the mean of the probabilities assigned to the true syntactic position in each model,

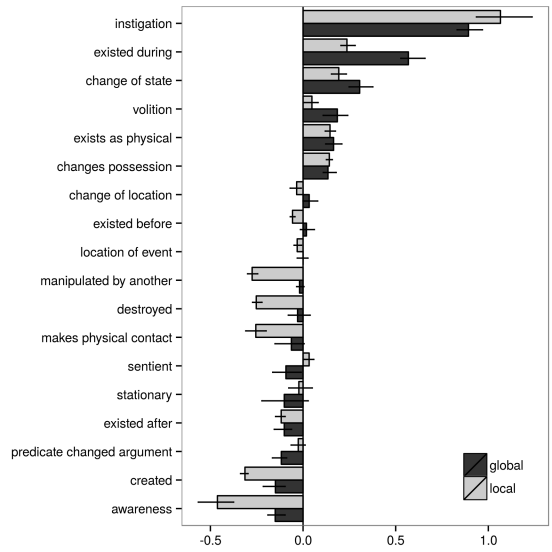


Figure 1: Coefficients for featural models (positive = predictive of subject; negative = predictive of non-subject)

broken out by syntactic position. While all models do relatively well at predicting that subjects are subjects, there is substantial variability with respect to predicting different kinds of non-subjects. For instance, the local VerbNet model tends to assign lower probability to direct objects being non-subjects, and the SPR1.0 model tends to assign lower probability to indirect objects and obliques being non-subjects.

The VerbNet global model improves substantially in this regard, suggesting that the direct objects it assigns low probability to in the local model have semantic roles that can appear in both subject and object position. In contrast, the SPR1.0 global does not improve as much.

This appears to be because, where the categorical datasets mark obliques with categories like A2-5, INSTRUMENT, RECIPIENT, etc., the featural dataset associates many of these arguments with proto-agent properties to some extent. This is particularly bad for obliques and indirect objects, since these often denote, e.g., RECIPIENTs, which tend to differ little from, e.g., SOURCEs in terms of properties. For instance, *the chefs* in (5) is misclassified by the SPR1.0 object as a subject.

(5) The execs gave *the chefs* a standing ovation.

Figure 1 shows the logistic regression coefficients for the featural model. Bars give the mean coefficient across outer CV folds and error bars give the range. A positive coefficient means 'predictive of subject position' and a negative coefficient means 'predictive of non-subject position.' We see here that the INSTIGATION and VOLITION properties, which are proto-agent properties, are predictive of being a subject in both the local and global models; and the CREATED property (along with the DESTROYED and MANIPULATED BY ANOTHER properties for the local model), which are proto-patient properties, are predictive of being a non-subject.

Beyond this, however, this model does not yield a split that looks much like Dowty's. For example, CHANGE OF STATE, CHANGE OF POSSESSION are proto-patient property for Dowty—and thus should be negative—while AWARENESS and SENTIENT are proto-agent properties—and thus should be positive.

One reason for the latter case may have to do with the existence of examples like (5), in which the non-subject is very likely both aware and sentient of the ovation. More generally, these results may be indicative that these properties should be partitioned into more than two roles, since it could be the case that only a subset of Dowty's proto-role properties map onto obliques, while another subset—perhaps, the full set maps—onto direct objects.

In the next section, we investigate this possibility by generalizing Dowty's theory. We do this by allowing as many proto-roles as are necessary to capture the data and find that, as these data suggest, two proto-roles is not enough.

## 5 Natural semantic roles

In this section, we present the Semantic Proto-Role Linking Model (SPROLIM), which is a multi-view mixture model for inducing *natural semantic roles* (NSRs) 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.

SPROLIM goes beyond Dowty's theory in two ways. First, like the SPR1.0-based featural model discussed in Section 4, SPROLIM allows for different levels of association between a property and a proto-role, and it takes into account that some prop-

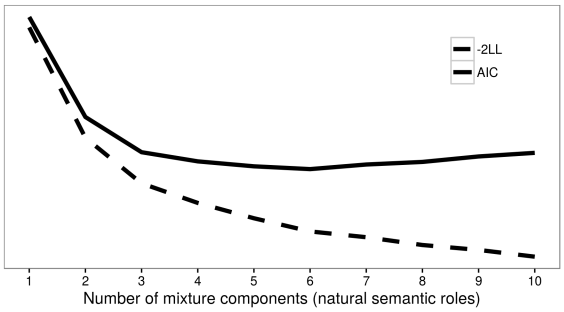


Figure 2: AIC and -2LL for various numbers of NSRs.

| | N0 | N1 | N2 | N3 | N4 | N5 |
|---|---|---|---|---|---|---|
| A0 | 0.55 | 0.29 | -0.39 | -0.56 | -0.42 | -0.22 |
| A1 | -0.30 | -0.12 | 0.36 | 0.50 | 0.16 | 0.19 |
| A2 | -0.24 | -0.15 | 0.06 | 0.07 | 0.22 | 0.04 |
| A3 | -0.12 | -0.06 | -0.02 | 0.03 | 0.14 | 0.01 |
| A4 | -0.14 | -0.10 | -0.03 | 0.01 | 0.17 | -0.02 |
| A5 | -0.06 | -0.05 | -0.06 | -0.04 | 0.07 | -0.04 |

Table 5: Spearman rank correlation between PropBank roles and natural semantic roles.

erties may be inapplicable for particular predicates. This contrasts with Dowty's theory, which assumes a simple counting mechanism that weights features evenly (see discussion in Section 4).

Second, instead of positing two property groupings (proto-roles)—e.g., proto-agent and proto-patient properties—as Dowty's theory does, SPROLIM can assume any number of proto-roles. In Experiment 3, we use a model selection procedure based on AIC to determine how many proto-roles, which SPROLIM implements as mixture components, are needed to describe the data. These mixture components constitute our NSR representation.

### 5.1 Semantic Proto-Role Linking Model

SPROLIM associates each argument of each predicate type with a discrete (mixture) distribution over latent components, which constitutes that argument's NSR. It furthermore associates each latent component with two models. The first, which we refer to as the *property model*, predicts property judgments and property applicability based on an argument's mixture distribution, and the second, which constitutes SPROLIM's linking model, predicts syntactic position based on that argument's mixture. Thus, the mixture distribution is used in the same way the categorical roles and raw SPR1.0 property judgments were in Experiments 1 and 2.

**Mixture distribution** Let the mixture distribution $\mathbf{m}_i$ for argument $i$ of some predicate lie on the $K$-simplex, where $K \in \mathbb{N}$ represents the number of mixing components (NSRs). We select $K$ using an AIC-based selection procedure (see Section 5.2).

**Property model** We implement the property model using a cumulative link logit hurdle model (see Agresti, 2014). In this model, each mixture component $k$ is associated with two real-valued vectors, $\mathbf{r}_k$ (for rating) and $\mathbf{s}_k$, with length equal to the number of properties. We first describe the cumulative link logit part of this model, corresponding to $\mathbf{r}_k$, and then the hurdle part, corresponding to $\mathbf{s}_k$.

In the cumulative link logit portion of the model, a categorical probability mass function with support on the property likelihood ratings $x \in \{1, \ldots, 5\}$ is associated with a value $r$ using a real-valued cut-point vector $\mathbf{c} = (0, c_2, \ldots)$ and defined as

$$\mathbb{P}(x = j \mid r, \mathbf{c}) = \begin{cases} 1 - q_{j-1} & \text{if } j = 5 \\ q_j - q_{j-1} & \text{otherwise} \end{cases}$$

where $q_j = \text{logit}^{-1}(c_{j+1} - r)$ and $c_j$ for all $j > 2$ are free parameters. 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 the hurdle portion of the model, a bernoulli probability mass function for applicability $s \in \{\text{applicable, not applicable (NA)}\}$ is given by $\mathbb{P}(a = \text{applicable} \mid s) = \text{logit}^{-1}(s)$

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 applicability; if it is not, they stop, but if it is, they generate a rating. The joint probability of $x$ and $a$ is then defined as

$$\mathbb{P}(x, a \mid r, s, \mathbf{c}) = \begin{cases} \mathbb{P}(a \mid s) & \text{if } a = \text{NA} \\ \mathbb{P}(x \mid r, \mathbf{c})\mathbb{P}(a \mid s) & \text{otherwise} \end{cases}$$

The expected log-likelihood of the data $\mathbf{X}$, $\mathbf{A}$ is then

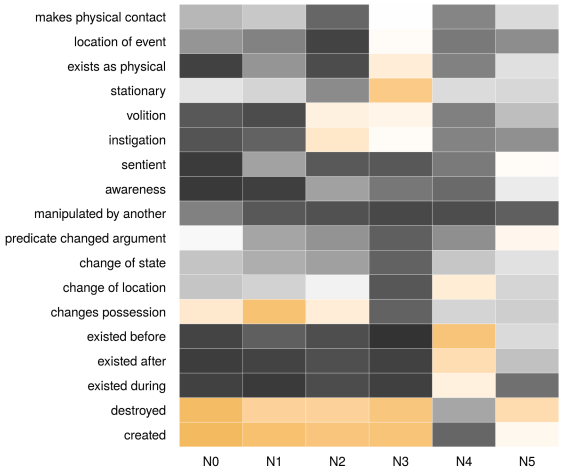


Figure 3: Rating weights for each SPR1.0 property on each mixture component in the optimal model (orange=more unlikely, white=uncertain, grey=more likely).

$$\mathbb{E} \log \mathcal{L}(\mathbf{M}, \mathbf{R}, \mathbf{S}) = \sum_{n,p} \log \left( \sum_k m_{g(n)k} p_{nkp} \right)$$

where $N$ is the total number of argument tokens and $P$ is the total number of properties, $g$ maps from datapoints (argument tokens) to argument types, and $p_{nkp} = \mathbb{P}(x_{np}, a_{np} \mid r_{kp}, s_{kp}, \mathbf{c})$.

**Linking model** Each mixture component $k$ is associated with a real number $l_k$. The probability of syntactic position $y \in \{\text{subject}, \text{non-subject}\}$ is given by $\mathbb{P}(y = \text{subject} \mid l) = \text{logit}^{-1}(l)$. The expected log-likelihood of the data $\mathbf{y}$ is then

$$\mathbb{E} \log \mathcal{L}(\mathbf{M}, \mathbf{L}) = \sum_n \log \left( \sum_k m_{g(n)k} \mathbb{P}(y_n \mid l_k) \right)$$

**Objective** Assuming some fixed $K$, we attempt to maximize the expected log-likelihood of the full model, which is just the sum of the above expectations. This yields the maximum likelihood estimate for the model with that $K$.

### 5.2 Experiment 3

In this experiment, we fit SPROLIM to the SPR1.0 data with different numbers $K$ of mixture components. As $K$ grows the likelihood is guaranteed to increase, but the explanatory power gained from adding a component is guaranteed to decrease.

**Information criterion** To balance out the increase in likelihood and the decrease in explanatory power, we use the Akaike Information Criterion (AIC; Akaike, 1974). AIC, which measures the information lost by representing the true model $\mathcal{M}_{\text{true}}$ in terms of the proposed model $\mathcal{M}$. Our aim is to find a number of mixture components $K$ such that $K = \arg\min_i \text{AIC}(\mathcal{M}_i)$, where $\mathcal{M}_i$ is SPROLIM with $i$ components.

**Model fitting** We use projected gradient descent with AdaGrad to find an approximation to $\boldsymbol{\theta}_{\text{MLE}}$ for each model $\mathcal{M}_i$. Since this optimization problem is non-convex, we randomly restart the optimizer for each $\mathcal{M}_i$ 10 times and select the model with the highest log-likelihood. (These restarts did not turn out to be necessary, since models tended to converge to the same solution across restarts.)

**Results** Figure 2 shows the log-likelihood (dashed line) and AIC (solid line) for each of setting of $K \in \{1, \ldots, 10\}$. We see that $K = 6$ minimizes the AIC. We assume this $K$ for the remainder of the paper.

It is striking that our model finds the same number of NSRs as there are core argument PropBank roles. This could suggest that SPROLIM is simply a method for labeling PropBank roles. This would be interesting, since Palmer et al. (2005) suggest that core argument PropBank roles like A0 and A1 correspond to proto-roles like proto-agent and proto-patient.

To check this, we compute Spearman rank correlations between each PropBank role and the probability associated with each mixture component. Table 5 shows these correlations. We see that for A0 and A1 have NSRs that correspond to them to some extent—N0 to A0 and N3 to A1—but these correlations are far from perfect. Further, A2-5 do not show high correlations with any NSRs.

Thus, it appears that, despite positing the same number of roles, SPROLIM finds a very different role space than PropBank. It furthermore suggests that core argument PropBank roles do not actually capture Dowty's proto-roles. But the fact that PropBank roles predict the syntax so well raises the question of how well the NSRs induced by SPROLIM do. We take this question up in Section 5.3.

| | | Discrete | | | Simplex | | | Ball | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | f1 | precision | recall | f1 | precision | recall | f1 | precision | recall |
| local | subject | 0.80 | 0.73 | 0.89 | 0.84 | 0.80 | 0.87 | 0.84 | 0.82 | 0.85 |
| | non-subject | 0.79 | 0.88 | 0.72 | 0.85 | 0.88 | 0.81 | 0.85 | 0.87 | 0.84 |
| | subject | 0.80 | 0.73 | 0.89 | 0.84 | 0.80 | 0.87 | 0.84 | 0.82 | 0.85 |
| global | subject | 0.91 | 0.91 | 0.92 | 0.98 | 0.98 | 0.98 | 0.97 | 0.97 | 0.97 |
| | non-subject | 0.91 | 0.92 | 0.91 | 0.98 | 0.98 | 0.98 | 0.97 | 0.97 | 0.97 |

Table 6: Mean F1, precision, and recall on outer CV test folds for local (Exp. 4) and global (Exp. 5) models.

We now turn to how SPROLIM groups proto-role properties. Figure 3 gives a representation of the rating weights $\mathbf{R}$ for the 6 components. Orange means the property is unlikely given that component, white means it is uncertain, and grey means it is likely. We see that N0 (the NSR most closely associated with A0) very closely approximates Dowty's proto-agent proto-role, while N3 (the NSR most closely associated with A1) is a mix of proto-patient properties like CHANGE OF STATE, CHANGE OF POSSESSION, MANIPULATED BY ANOTHER, but also some proto-agent properties SENTIENT and AWARENESS. In spite of these latter two properties N3 seems like the best candidate for Dowty's proto-patient roles.

### 5.3 Experiment 4 and 5

In these experiments, we construct local (Experiment 4) and global (Experiment 5) classifiers for predicting an argument's syntactic position based on the SPROLIM mixture representation as predictors. We use the same methodology as in Section 4.

We consider three potential NSR representations constructed from the mixture induced by SPROLIM: (i) the raw simplex representation induced by the model; (ii) a discrete form of this simplex representation converted to a one-hot by taking the max; and (iii) a representation constructed by mapping the simplex representation into the unit ball.

**Classifier** The classifier used in Experiment 1 is used in Experiment 4 and the classifier used in Experiment 2 is used in Experiment 5. The only difference in both cases is that, instead of categorical roles or raw properties ratings, the SPROLIM mixture representations described above are used.

**Cross-validation** The nested cross-validation procedure used in Experiment 1 is used in Experiment 4 and the nested cross-validation procedure used in Experiment 2 is used in Experiment 5.

**Results** Table 6 gives the mean F1, precision, and recall on the outer cross-validation test folds. We see that the simplex and ball models in Experiment 1 outperform the discrete model by 4-5 points in macro F1 (mean across subject and non-subject F1). Further, across all predictor sets, all three measures improve compared to the local models, as in Experiments 1 and 2.

Compared to Experiment 1, the simplex and ball models do as well at classifying subjects as the original SPR1.0 model, but they outperform that model by 6 points in F1 in classifying non-subjects. Much of this gain comes from better classification of indirect objects and obliques, which we noted as a problem for the Experiment 1 model.

Compared to Experiment 2, the simplex model outperforms the VerbNet model—with the ball model doing about as well. Further, the simplex model approaches the performance of the PropBank model, which outperforms it by one point in F1. We hypothesize that this gap can be closed by adding more property questions to the SPR1.0 data.

## 6 Conclusion

We introduced a framework for *computational linking theory* (CLT). We then deployed this framework to assess the efficacy of Dowty's (1991) thematic proto-role theory as applied to naturally occurring text. Beyond the theoretical interest, a major reason for this assessment is to test the possibility of inducing natural semantic roles (NSRs) from purely semantic information. We showed that the NSRs we induced using our generalization of Dowty's theory nearly match PropBank in CLT predictive power.

## References


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

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

Eleftherios Avramidis and Philipp Koehn. Enriching Morphologically Poor Languages for Statistical Machine Translation. In *ACL*, 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.

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. *On Linking: Papers by Richard Carter Cambridge MA: Center for Cognitive Science, MIT (Lexicon Project Working Papers No. 25)*, 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*, pages 4585–4592, 2014.

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

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.

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

Matthew R. Gormley, Margaret Mitchell, Benjamin Van Durme, and Mark Dredze. Low-Resource Semantic Role Labeling. In *ACL (1)*, pages 1177–1187, 2014.

Scott Grimm. Semantics of case. *Morphology*, 21 (3-4):515–544, 2011.

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.

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 (AMTA 2010)*, 2010.

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. Citeseer, 2005.

Philipp Koehn and Hieu Hoang. Factored Translation Models. In *EMNLP-CoNLL*, 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. *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.

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 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 0674042174.

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.

Lilia Rissman, Kyle Rawlins, and Barbara Landau. Using instruments to understand argument structure: Evidence for gradient representation. *Cognition*, 142:266–290, September 2015. ISSN 0010-0277. doi:

10.1016/j.cognition.2015.05.015. URL http://www.sciencedirect.com/science/article/pii/S0010027715001122.

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 *ACL*, 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.