A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences

Katrin Erk*
University of Texas at Austin

Sebastian Padó**
Heidelberg University

Ulrike Padó†
Vico Research and Consulting GmbH

We present a vector space–based model for selectional preferences that predicts plausibility scores for argument headwords. It does not require any lexical resources (such as WordNet). It can be trained either on one corpus with syntactic annotation, or on a combination of a small semantically annotated primary corpus and a large, syntactically analyzed generalization corpus. Our model is able to predict inverse selectional preferences, that is, plausibility scores for predicates given argument heads.

We evaluate our model on one NLP task (pseudo-disambiguation) and one cognitive task (prediction of human plausibility judgments), gauging the influence of different parameters and comparing our model against other model classes. We obtain consistent benefits from using the disambiguation and semantic role information provided by a semantically tagged primary corpus. As for parameters, we identify settings that yield good performance across a range of experimental conditions. However, frequency remains a major influence of prediction quality, and we also identify more robust parameter settings suitable for applications with many infrequent items.

1. Introduction

Selectional preferences or **selectional constraints** describe knowledge about possible and plausible fillers for a predicate's argument positions. They model the fact that there is often a semantically coherent set of concepts that can fill a given argument position. Selectional preferences can help for many text analysis tasks which involve comparing different attachment decisions. Examples include syntactic disambiguation (Hindle and Rooth 1993; Toutanova et al. 2005), word sense disambiguation (WSD,

* Department of Linguistics, Calhoun Hall 512, 1 University Station B 5100, Austin, TX 78712. E-mail: katrin.erk@mail.utexas.edu.

** E-mail: pado@cl.uni-heidelberg.de.

† E-mail: ulrike.pado@vico-research.com.

Submission received: 26 November 2009; revised submission received: 6 May 2010; accepted for publication: 29 June 2010. Part of the work reported in this article was done while S. P. was a postdoctoral scholar and U. P. a visiting scholar at Stanford University.

McCarthy and Carroll 2003), semantic role labeling (SRL, Gildea and Jurafsky 2002), and characterizing the conditions under which entailment holds between two predicates (Zanzotto, Pennacchiotti, and Pazienza 2006; Pantel et al. 2007). Furthermore, selectional preferences are also helpful for determining linguistic properties of predicates and predicate–argument combinations, for example in compositionality assessment (McCarthy, Venkatapathy, and Joshi 2007) or the detection of diathesis alternations (McCarthy 2000). In psycholinguistics, selectional preferences predict human plausibility judgments for predicate–argument combinations (Resnik 1996) and effects in human sentence reading times (Padó, Crocker, and Keller 2009).

All these applications rely on the availability of broad-coverage, reliable selectional preferences for predicates and their argument positions. Given the immense effort necessary for manual semantic lexicon building and its associated reliability problems (see, e.g., Briscoe and Boguraev 1989), all contemporary models of selectional preferences acquire selectional preferences automatically from large corpora.

The simplest strategy is to extract triples (v, r, a) of a predicate, role, and argument headword (or filler) from a corpus, and then to compute selectional preference as relative frequencies. However, due to the Zipfian nature of word frequencies, the first step on its own results in a very sparse list of headwords, in particular for less frequent predicates. As an example, the verb *anglicize* only appears with nine direct objects in the 100-million word British National Corpus (BNC, Burnard 1995). Only one of them, *name*, appears more than once. Many highly plausible fillers are missing from the list, such as *word* or *spelling*.

In order to make sensible predictions for triples that are unseen at training time, it is crucial to add a **generalization** step that infers a degree of preference for new, unseen headwords for a given predicate and role.[1] The result is, in the ideal case, an assignment to every possible headword of some degree of compatibility (or **plausibility**) with the predicate's preferences. In the case of *anglicize*, the desired result would be a high plausibility for words like the (previously seen) *wordlist* and *surname* as well as the (unseen) *word* and *spelling*, and a low plausibility for (likewise unseen) words like *cow* and *machine*.

The predominant approach to generalizing over headwords, first introduced by Resnik (1996), is based on semantic hierarchies such as WordNet (Miller et al. 1990). The idea is to map all observed headwords onto synsets, and then generalize to a characterization of the selectional preference in terms of the WordNet noun hierarchy. This can be achieved in many different ways (Abe and Li 1996; Resnik 1996; Ciaramita and Johnson 2000; Clark and Weir 2001). The performance of these models relies on the coverage of the lexical resources, which can be a problem even for English (Gildea and Jurafsky 2002). An alternative approach to generalization uses **co-occurrence information**, either in the form of distributional models or through a clustering approach. These models, which avoid dependence on lexical resources, use corpus data for generalization (Dagan, Lee, and Pereira 1999; Rooth et al. 1999; Bergsma, Lin, and Goebel 2008).

In this article, we present a lightweight model for the acquisition and representation of selectional preferences. Our model is fully distributional and does not require any knowledge sources beyond a large corpus where subjects and objects can be identified with reasonable accuracy. Its key point is to use vector space similarity (Lund and Burgess 1996; Laundauer and Dumais 1997) to generalize from seen to unseen

[1] Some approaches also fix a role and headword list and generalize from seen predicates to other, similar predicates.

headwords. The vector space representations which serve as a basis for computing similarity can in principle be computed from any arbitrary corpus, given that it is large enough. In particular, this need not be the same corpus as the one on which we observe predicate–headword co-occurrences. Our model thus distinguishes between a **primary corpus**, from which the predicate–role–headword triples are extracted, and a **generalization corpus** for computing the vector space representations. This distinction makes it possible to apply our model to primary corpora with rich information that are too small for efficient generalization, such as domain-specific corpora or corpora with deeper linguistic analysis, as long as a larger, even if potentially noisier, generalization corpus is available. We empirically demonstrate the benefit of this distinction. We use FrameNet (Fillmore, Johnson, and Petruck 2003) as primary corpus and the BNC as generalization corpus, modeling selectional preferences for semantic roles with near-perfect coverage and low error rate.[2]

We evaluate our model on two tasks. The first task is pseudo-disambiguation (Yarowsky 1993), where the model decides which of two randomly chosen words is a better filler for the given argument position. This task tests model properties that are needed for concrete semantic analysis tasks, most notably word sense disambiguation, but also for semantic role labeling. The second task is the prediction of human plausibility ratings, which is a standard task-independent benchmark for the quality of selectional preferences. We test our model across a range of parameter settings to identify best-practice values and show that it robustly outperforms both WordNet-based and other distributional models on both tasks.

Finally, we investigate **inverse preferences**, that is, preferences that arguments have for their predicates. Although there is ample cognitive evidence for the existence of such preferences (e.g., McRae et al. 2005), to our knowledge, they have not been investigated systematically in linguistics. However, statistics about inverse preferences have been used implicitly in computational linguistics (e.g., Hindle 1990; Rooth et al. 1999). We investigate the properties of inverse selectional preferences in comparison to regular selectional preferences, and show that it is possible to predict inverse preferences with our selectional preference model as well.

The model that we discuss in this article, EPP, was first introduced in Erk (2007) (using a pseudo-disambiguation task for evaluation) and further studied by Padó, Padó, and Erk (2007) (evaluating against human plausibility judgments). In the current text, we perform a more extensive evaluation and analysis, including the new evaluation on inverse preferences, and we introduce a new similarity measure, nGCM, which achieves excellent performance in many settings.

2. Computational Models of Selectional Preferences

In this section, we provide an overview of corpus-based models of selectional preferences. See Table 1 for a summary of the notation that we use.

[2] As descriptions of semantic classes of participants in events, selectional preferences are most naturally applied to semantic argument positions, that is, *semantic roles* (such as agent or patient). In contrast, syntactic argument positions (like subject and object) can comprise several semantic argument positions, due to the presence of diathesis alternations, and thus show less consistent selectional preferences. Nevertheless, work in computational linguistics also makes use of selectional preferences for *syntactic* argument positions, considering them noisy approximations of semantic argument positions.

Table 1
Notation used throughout the article.

$w \in \mathsf{Lemmas}$	**Word**. We assume lemmatization throughout.
$v \in \mathsf{Preds}$	**Predicate**. Preds may be a subset of Lemmas, or a set of semantic classes.
$r \in \mathsf{Roles}$	**Role/Argument slot**. Roles may be a set of grammatical functions, or of semantic roles.
$a \in \mathsf{Args} \subseteq \mathsf{Lemmas}$	**(Potential) argument** headword.
$c \in C$	**Semantic class** on which selectional preferences are conditioned, for example, WordNet sense, FrameNet frame, or latent semantic class.
$VS = (\mathsf{DTrans}, \mathsf{Basis}, \mathsf{sim}, \mathsf{STrans})$	**Vector space**. Basis is a set of basis elements, sim a similarity measure, DTrans a transformation of raw counts, and STrans a transformation of the space. We write $\vec{w} = \langle w_{b_1}, \ldots, w_{b_n} \rangle$ for the representation of $w \in \mathsf{Lemmas}$ in a vector space with $\mathsf{Basis} = \{b_1, \ldots, b_n\}$.
$wt_{r,v}(a)$	**Weight** of argument headword a for predicate v and role r.

2.1 Historical Models

In formal linguistics, selectional restrictions were employed as strict Boolean restrictions by Katz and Postal (Katz and Fodor 1963; Katz and Postal 1964) as input to a mutual disambiguation process between predicates and their modifiers. Sentences are semantically anomalous if there are no mutually consistent readings for the two words. Semantically anomalous sentences would receive no reading, whereas ambiguous sentences would receive several readings.

The strict dismissal as meaningless of sentences that violate selectional restrictions was later criticized. A case in point is metaphors, which often combine predicates and arguments from different domains (Lakoff and Johnson 1980). Wilks (1975:329) stated that "rejecting utterances is just what humans do not. They try to understand them." He proposes to reconceptualize selectional restrictions as preferences whose violation is dispreferred, but not fatal. His proposal for a semantic interpretation mechanism still uses semantic primitives, but always produces a single most plausible interpretation by choosing the senses of each word that maximize the compatibility between selectional preferences and semantic types. In this manner, he is able to compute semantic representations for sentences that violate selectional restrictions, including metaphors such as "my car drinks gasoline."

2.2 Semantic Hierarchy–Based Models

The first broad-coverage computational model of selectional preferences, and still one of the best-known ones, namely that of Resnik (1996), belongs to the class of semantic hierarchy–based models. These models generalize over observed headwords using a semantic hierarchy or ontology for nouns. The two main advantages of such models are that (a) they can make predictions for all words covered by the hierarchy, even for very infrequent ones for which distributional representations tend to be unreliable; and (b) the hierarchy robustly guides generalization even for few observed headwords.

Resnik's model instantiates the set of relations `Roles` with grammatical functions which can be observed in syntactically analyzed corpora. More specifically, it concentrates on selectional preferences for subjects and objects. For the generalization step, Resnik's model maps all headwords onto WordNet synsets (or classes) c. Resnik first computes the overall **selectional preference strength** for each verb–relation pair (v, r), that is, the degree to which the pair constrains possible fillers. To estimate this quantity, the distribution of WordNet synsets for this particular verb–relation pair is compared to the distribution of synsets over all verbs, given the relation r. Technically, this is achieved using Kullback–Leibler divergence:

$$\mathsf{SelStr}(v,r) = D(P(c|v,r)||P(c|r)) = \sum_{c \in C} P(c|v,r) log(\frac{P(c|v,r)}{P(c|r)}) \tag{1}$$

The parameters $P(c|v,r)$ and $P(c|r)$ are estimated from the corpus frequencies of tuples (v, r, a) and the membership of nouns a in WordNet classes c: The observed frequency of (v, r, a) is split equally among all WordNet classes for a. This avoids word sense disambiguration, but incurs a certain share of wrong attributions. The intuition of $\mathsf{SelStr}(v, r)$ is that a verb–relation pair that only allows a limited range of argument heads will have a posterior distribution over classes that strongly diverges from the prior.

Next, the **selectional association** of the triple, $\mathsf{SelAssoc}(v, r, c)$, is computed as the ratio of the selectional preference strength for this particular class c to the overall selectional preference strength of the verb–relation pair (v, r). This is shown in Equation (2).

$$\mathsf{SelAssoc}(v,r,c) = \frac{P(c|v,r) log \frac{P(c|v,r)}{P(c|r)}}{\mathsf{SelStr}(v,r)} \tag{2}$$

Finally, the selectional preference between a verb, a relation, and an argument head is defined as the maximal selectional association of the verb, the relation, and any WordNet class c that the argument can instantiate. We will refer to this model as RESNIK herein.

In subsequent years, a number of WordNet-based models were developed that differ from Resnik's model in the details of how the generalization in the WordNet hierarchy is performed. Abe and Li (1996) characterize selectional preferences by a tree cut through the WordNet noun hierarchy that minimizes tree cut length while maximizing accuracy of prediction. Clark and Weir (2001) perform generalization by ascending the WordNet noun hierarchy as long as the degree of selectional preference among siblings is not significantly different. Ciaramita and Johnson (2000) encode WordNet in a Bayesian Network to take advantage of the Bayes nets' ability to "explain away" ambiguity. Grishman and Sterling (1992) perform generalization on the basis of a manually constructed semantic hierarchy specifically developed on the same corpus.

2.3 Distributional Models

Distributional models do not make use of any lexicon resource for the generalization step. Instead, they use word co-occurrence—typically obtained from the same corpus as the observed headwords—for generalization. This independence from manually constructed resources gives distributional models a good cost–benefit ratio and makes them especially attractive for domain-specific applications. These models, like the semantic hierarchy–based models, usually use grammatical functions as the set `Roles` for which selectional preferences are predicted.

Pereira, Tishby, and Lee (1993) and Rooth et al. (1999) generalize by discovering latent classes of noun–verb pairs with soft clustering. They model the probability of a word a as the argument of a predicate v as the probability of generating v and a independently from the latent classes c:

$$P(v,a) = \sum_{c \in C} P(c,v,a) = \sum_{c \in C} P(c)P(v|c)P(a|c) \qquad (3)$$

Pereira, Tishby, and Lee (1993) develop a task-specific procedure to optimize $P(c)$, $P(v|c)$, and $P(a|c)$. Their procedure supports hierarchical clustering and can optimize the number of clusters. Rooth et al. (1999) present a simpler Expectation Maximization–based estimation procedure which takes the number of clusters as input parameter. We refer to this model as ROOTH ET AL. herein.

Dagan, Lee, and Pereira (1999) introduce a general model for computing co-occurrence probabilities with similarity-based smoothing. Although not intended as a model of selectional preferences, it can also be interpreted as such. Given a similarity measure *sim* defined on word pairs, they compute the smoothed occurrence probability of a word w_2 given w_1 as

$$P_{sim}(w_2|w_1) = \sum_{w \in \mathsf{Simset}(w_1)} \frac{sim(w_1,w)}{Z(w_1)} P(w_2|w) \qquad (4)$$

where $\mathsf{Simset}(w)$ is the set of words most similar to w according to *sim*, and $Z(w_1) = \sum_{w \in \mathsf{Simset}(w_1)} sim(w_1,w)$ is a normalizing factor. This model predicts w_2 given w_1 by backing off from w_1 to words w similar to w_1. The contribution of each w in predicting $P(w_2|w_1)$ is weighted by $sim(w_1,w)$. The similarity $sim(w_1,w)$ is computed on vector space representations.

Recently, Bergsma, Lin, and Goebel (2008) have adopted a discriminative approach to the prediction of selectional preferences. The features they use are mainly co-occurrence statistics, enriched with morphological context features to alleviate sparse data problems for low-frequency argument heads. They train one SVM per verb–argument position pair, using unobserved verb–argument combinations as negative examples, which makes their approach independent of manually annotated training data. Schulte im Walde et al. (2008) present a model that combines features of the semantic hierarchy–based and the distributional approaches by integrating WordNet into an EM-based clustering model; Schulte im Walde (2010) shows that integrating noun–modifier relations improves the prediction of human plausibility judgments.

2.4 Semantic Role–Based Models

The third class of models takes advantage of semantic resources beyond simple semantic hierarchies, notably of corpora with semantic role annotation. Such corpora allow the prediction of selectional preferences for semantic roles rather than grammatical functions. From a linguistic perspective, semantic roles represent a more appropriate level for defining selectional preferences. For that reason, the role annotation provides cleaner and more specific training data than even a manually syntactically annotated corpus would. These advantages, however, come at the cost of considerably greater sparsity issues.

Padó, Crocker, and Keller (2009) present a model based on FrameNet (Fillmore, Johnson, and Petruck 2003). This model estimates selectional preferences with a generative probability model that equates the plausibility of a (v, r, a) triple with the joint probability of observing the thematic role r, the verb v, and the argument a, plus the verb's FrameNet sense c and the grammatical function gf of the argument. This joint probability can be decomposed using the chain rule:

$$P(v, c, r, gf, a) = P(v)P(c|v)P(r|v, c)P(gf|r, v, c)P(a|gf, r, v, c) \quad (5)$$

The model does not make any independence assumptions. To counteract sparse data issues for the more complex terms, the model applies WordNet-based generalization (for nouns), distributional clustering (for verbs), and Good–Turing smoothing. We refer to this model as **PADO ET AL.** Another semantic role–based model was proposed by Vandekerckhove, Sandra, and Daelemans (2009). It acquires selectional preferences for PropBank roles from a PropBank-labeled corpus, generalizing to unseen headwords with memory-based learning.

3. A Distributional Exemplar-Based Model of Selectional Preferences: EPP

We now present the EPP model of selectional preferences. It falls into the category of distributional models. More specifically, it is an **exemplar model** that remembers all seen headwords for a given argument position and computes the degree of plausibility for a new headword candidate through its similarity to the stored exemplars. Exemplars are modeled as vectors in a semantic space.

Exemplar models are a well-known modeling framework that is used in psychology (Nosofsky 1986), in computational linguistics (under the name of memory-based learning [Daelemans and van der Bosch 2005]), and in linguistics, particularly phonetics (Hay, Nolan, and Drager 2006). The appeal of exemplar models is that they provide a cognitively plausible process of learning as storing exemplars, and categorization as similarity computation that is grounded in features of the exemplars (e.g., formants in phonetics, and contexts in lexical semantics).

The representation of selectional preferences through feature vectors also fits in well with work in psycholinguistics by McRae, Ferretti, and Amyote (1997), who studied the characterization of verb selectional preferences through features elicited from human subjects. They found high overlap between features used to characterize the selectional preferences on the one hand, and features listed for typical role fillers on the other hand. For example, features generated for the agent role of *frighten* include *mean*, *scary*, and *ugly*, features that were also highly relevant for the typical filler noun *monster*.

As briefly mentioned in Section 1, we consider selectional preferences to be characterizations of typical fillers for the *semantic roles* of a predicate. Still, we keep our model modular to different notions of argumenthood, such that it is also applicable to the computation of selectional preferences for syntactic dependents of a predicate, as this is an important case for computational applications. When we compute selectional preferences for syntactic dependents rather than semantic roles, we view syntactic argument positions as noisy approximations of semantic roles.

3.1 The Model

As stated previously, we assume that we have two corpora which assume different functions in the model: the primary corpus, which provides information about predicate–argument co-occurrences but may be too sparse for generalization; and the large, but potentially noisy, generalization corpus, from which we obtain reliable semantic similarity estimates.

Thus, the first step is the extraction of triples (v, r, a) of a predicate $v \in \mathsf{Preds}$, a relation $r \in \mathsf{Roles}$, and a headword $a \in \mathsf{Args}$ from the primary corpus. Let $\mathsf{Seenargs}(r, v)$ be the set of argument headwords seen with an argument position r of a predicate v in the primary corpus. Given these triples, we predict the plausibility for an arbitrary noun a_0 in position (v, r) through the **semantic similarity** of a_0 to all the members of $\mathsf{Seenargs}(r, v)$. We obtain these similarity ratings by first computing vector space representations for both and the members of $seen(r, v)$ from the generalization corpus, and then using a standard vector space similarity measure. We compute the plausibility for a_0 as

$$\mathsf{Selpref}_{\mathsf{EPP}\,r,v}(a_0) = \sum_{a \in \mathsf{Seenargs}(r,v)} \frac{wt_{r,v}(a)}{Z_{r,v}} \cdot \mathsf{sim}(a_0, a) \quad (6)$$

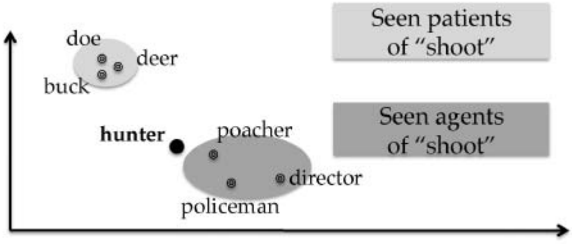
where $\mathsf{sim}(a_0, a)$ is the similarity between the vector space representations of a_0 and a, $wt_{r,v}(a)$ a weight for the seen headword a, and $Z_{r,v}$ a normalization constant, $Z_{r,v} = \sum_{a \in \mathsf{Seenargs}(r,v)} wt_{r,v}(a)$, so that the number of observed exemplars for each (v, r) pair does not matter. Because $\mathsf{Selpref}_{\mathsf{EPP}}$ is basically a weighted average over similarity values, the range of $\mathsf{Selpref}_{\mathsf{EPP}}$ is identical to the range of the employed similarity function sim. For example, the range is $[-1, 1]$ for cosine similarity, or $[0, 1]$ for the Jaccard coefficient (cf. Section 3.3). We discuss possible choices of both the similarity sim and the weight $wt_{r,v}$ in Section 3.3.

3.2 Vector Space Representations

We use vector space representations for generalization. In a vector space model, each **target word** is represented as a vector, typically constructed from co-occurrence counts with context words in a large corpus (the so-called **basis elements**). The underlying assumption, which goes back to Firth (1957) and Harris (1968), is that words with similar meanings occur in similar contexts and will be assigned similar vectors. Thus, the distance between the vectors of two target words, as given by some distance measure (e.g., Cosine or Jaccard), reflects their **semantic similarity**.

Vector space models are simple to construct, and the semantic similarity they provide has found a wide range of applications. Examples in NLP include information retrieval (Salton, Wong, and Yang 1975), automatic thesaurus extraction (Grefenstette 1994), and predominant sense identification (McCarthy et al. 2004). Lexical resources based on distributional similarity (e.g., Lin [1998]'s thesaurus) are used in a wide range of applications that profit from knowledge about word similarity. In cognitive science, they have been used, for example, to account for the influence of context on human lexical processing (McDonald and Brew 2004) and lexical priming (Lowe and McDonald 2000).

An idealized example for a semantic space representation of selectional preferences is shown in Figure 1(a). The two ellipses represent the exemplar clouds formed by the



(a) Ideal topology

Target Basis elements	deer	hunter
shoot	10	10
escape	12	12

Target Basis elements	deer	hunter
shoot-SUBJ	0	8
shoot-OBJ	10	2
escape-SUBJ	10	5
escape-PP$_{\text{from}}$	2	7

(b) Two co-occurrence matrices

Figure 1
An idealized vector space for the plausibilities of (*shoot*, *agent*, *hunter*) and (*shoot*, *patient*, *hunter*).

fillers of the agent and patient position of *shoot*, respectively. In order to judge whether a hunter is a plausible agent of *shoot*, the vector space representation of *hunter* is compared to the members of the exemplar cloud for the agent position—namely, *poacher*, *policeman*, and *director*. Due to the high average similarity of the *hunter* vector to these vectors, *hunter* will be judged a fairly good agent of *shoot*. Compare this with the result for the patient role: *hunter* is rather distant from *roe*, *deer*, and *buck*, and is therefore predicted to be a bad patient of *shoot*. However, note that *hunter* is still more plausible as a patient of *shoot* than, for example, *director*.

3.3 Formalization and Parameter Choice

Vector space models have been formalized by Lowe (2001) as tuples VS = (DTrans, Basis, sim, STrans), where Basis is a set of basis elements or dimensions, DTrans is a transformation of raw co-occurrence counts, sim is a similarity measure, and STrans is a transformation of the whole space, typically dimensionality reduction. An additional parameter that becomes relevant for our use of vector spaces (cf. Equation [6]) is the weighting function *wt* that determines the contribution of each exemplar to the overall similarity. We discuss the parameters in turn and discuss our reasons for either exploring them or fixing them.

Basis elements Basis. Traditionally, context words are used as basis elements, and co-occurrence is defined in terms of a surface window. Such bag-of-words spaces tend to group words by topics. They ignore the syntactic relation between context items and the target, which is a problem for selectional preference modeling. The top table in Figure 1(b) illustrates the problem: *deer* and *hunter* receive identical vectors, even though they show complementary plausibility ratings. The reason is that *deer* and *hunter* often co-occur in similar lexical bag-of-words contexts (namely, hunting-related activities). The bottom table in Figure 1(b) indicates a way out of this problem, namely the use of word-relation pairs as basis elements (Grefenstette 1994; Padó and Lapata 2007). This space splits the co-occurrences with context words such as *shoot* based on the grammatical relation between target and context word, and this split looks different for different words: whereas *deer* occurs exclusively as the object of *shoot*, *hunter* predominantly occurs as the subject. We find the reverse pattern for *escape*. In consequence,

Table 2
Similarity measures explored in this article. Notation: We assume $\mathsf{Basis} = \{b_1, \ldots, b_n\}$. We write I for mutual information, and $\mathsf{BE}(a)$ for the set of basis elements that co-occur at least once with a.

$$
\begin{array}{llllll}
\mathsf{sim}_{\mathrm{Lin}}(a,a') & = & \frac{\sum_{(r,v)\in \mathsf{BE}(a)\cap \mathsf{BE}(a')} I(a,r,v)+I(a',r,v)}{\sum_{(r,v)\in \mathsf{BE}(a)} I(a,r,v) \sum_{(r,v)\in \mathsf{BE}(a')} I(a',r,v)} & \mathsf{sim}_{\mathrm{cosine}}(a,a') & = & \frac{\sum_{i=1}^{n} a_{b_i}\cdot a'_{b_i}}{||\vec{a}||\cdot||\vec{a'}||} \\
\mathsf{sim}_{\mathrm{Dice}}(a,a') & = & \frac{2\cdot|\mathsf{BE}(a)\cap \mathsf{BE}(a')|}{|\mathsf{BE}(a)|+|\mathsf{BE}(a')|} & \mathsf{sim}_{\mathrm{Jaccard}}(a,a') & = & \frac{|\mathsf{BE}(a)\cap \mathsf{BE}(a')|}{|\mathsf{BE}(a)\cup \mathsf{BE}(a')|} \\
\mathsf{sim}_{\mathrm{nGCM}}(a,a') & = & exp\Big(-\sqrt{\sum_{i=1}^{n}(\frac{a_{b_i}}{||\vec{a}||}-\frac{a'_{b_i}}{||\vec{a'}||})^2}\Big) & \text{where } ||\vec{a}|| = \sqrt{\sum_{i=1}^{n} a_{b_i}^2} & & \\
\mathsf{sim}_{\mathrm{Hindle}}(a,a') & = & \sum_{i=1}^{n} \mathsf{sim}_{\mathrm{Hindle}}(a,a',i) & \text{where} & & \\
\mathsf{sim}_{\mathrm{Hindle}}(a,a',i) & = & \begin{cases} \min(I(a,b_i),I(a',b_i)) & \text{if } I(a,b_i)>0 \text{ and } I(a',b_i)>0 \\ \mathrm{abs}(\max(I(a,b_i),I(a',b_i))) & \text{if } I(a,b_i)<0 \text{ and } I(a',b_i)<0 \\ 0 & \text{else} \end{cases} & & &
\end{array}
$$

the resulting spaces gain the ability to distinguish between words like *hunter* and *deer*, based on differences in typical occurrences in argument positions.

On the downside, dependency-based spaces are more expensive to compute than word-based spaces because they require a corpus with syntactic analysis. Thus, we explore both options. The word-based space records co-occurrences within a surface window of 10 (lemmatized) words.[3] We refer to it as WORDSPACE. The dependency-based space, called DEPSPACE, has basis elements consisting of a grammatical function concatenated with a word, as in the bottom example in Figure 1(b) (Padó and Lapata 2007). Following earlier experiments on the representation of selectional preferences in word-dependency-relation spaces (Padó, Padó, and Erk 2007), we use a subject–object context specification that only considers co-occurrences between verbs and their subjects and direct objects.[4] In each case, we adopt the 2,000 most frequent context items as basis elements.

Similarity measure sim. In principle, any similarity measure for vectors can be plugged into our model. Previous studies that compared similarity measures came to various conclusions about the usefulness of different measures. Cosine similarity is very popular in Information Retrieval. Lee (1999) obtains good results for the Jaccard coefficient in pseudo-disambiguation. In the synonymy prediction task of Curran (2004), Dice emerged in first place. Padó and Lapata (2007) found good results with Lin's measure for predominant word sense identification.

Because it is unclear whether the findings about best similarity measures generalize to new tasks, we will investigate a range of similarity measures shown in Table 2: **Cosine**, the **Dice** and **Jaccard** coefficients, **Hindle**'s (1990) and **Lin**'s (1998) mutual information-based metrics, and an adaptation of Nosofsky's (1986) Generalized Context Model (GCM), a model for exemplar-based similarity from psychology. The original GCM includes normalization by summed similarity over all classes of exemplars, which introduces competition between categories. Our version, which we call **nGCM,** instead normalizes by vector length to alleviate the influence of overall target frequency, but

[3] We do not remove stop words for reasons of simplicity, as there is no unequivocal definition of this set, and we do not wish to remove potentially informative contexts.

[4] This context specification is available as `soonly` in the DependencyVectors software package (`http://www.nlpado.de/~sebastian/dv.html`) starting from Release 2.5.

preserves the central idea that similarity decreases exponentially with distance (Shepard 1987).

All similarity measures from Table 2 are applicable to semantic spaces with arbitrary basis elements, with the exception of the Lin measure, whose definition applies only to dependency-based spaces. The reason is that it decomposes the basis elements into relation–word pairs (r, v). For semantic spaces with words as basis elements, the Lin measure can be adapted by omitting the random variable r (cf. Padó and Lapata 2007).

Transformations DTrans *and* STrans. Next, we come to transformations on counts and vector spaces. Concerning the count transformations DTrans, all counts are log-likelihood transformed (Dunning 1993), a standard procedure for word-based semantic space models which alleviates the problematic effects of the Zipfian distribution of lexical items, as proposed by Lowe (2001). As for transformations on the complete space STrans, many studies do not perform dimensionality reduction at all. Others, like the LSA family of vector spaces (Landauer and Dumais 1997), regard it as a crucial ingredient. To gauge the impact of STrans, we compare unreduced spaces (2,000 dimensions) to 500-dimensional spaces created using Principal Component Analysis (PCA), a standard method for dimensionality reduction that identifies the directions of highest variance in a high-dimensional space.

Weight functions wt. Exemplar-based models are usually applied in conjunction with a function that can assign each exemplar an individual weight, which can be interpreted cognitively as degree of activation (Nosofsky 1986). We assess a small number of weight functions to investigate their importance within the EPP model. The first one, UNI, assumes a uniform distribution, $wt_{r,v}(a) = 1$. The second one, FREQ, uses the co-occurrence frequency as weight, $wt_{r,v}(a) = freq(a, r, v)$, with the intuition that more frequent exemplars should be both more activated and more reliable. Finally, we consider a weight function that is an analogue of inverse document frequency in Information Retrieval. It weights words higher that occur with a smaller number of verb–role pairs: $wt_{r,v}(a) = \log \frac{|\bigcup_{a'} \mathsf{Seen}_{\mathsf{rv}}(a')|}{|\mathsf{Seen}_{\mathsf{rv}}(a)|}$, where we write $\mathsf{Seen}_{\mathsf{rv}}(a)$ for the set of verb–role pairs (r, v) for which a occurs as a headword.[5] We abbreviate this weight function by DISCR for 'discrimination'.

3.4 Discussion

Our EPP model can be seen as a straightforward implementation of the intuition to model selectional preference by generalizing from seen headwords to other, similar, words. We use vector space representations to judge the similarity of words, obtaining a completely corpus-driven model that does not require any additional resources and is very flexible. A complementary view on this model is as a generalization of traditional vector space models that represent semantic similarities between pairs of words. The EPP model goes beyond this by computing similarity between a vector and a set of other vectors. By instantiating the set with the vectors for seen headwords of some relation r, the similarity turns into a plausibility prediction that is specific to this relation.

Like other distributional models, the EPP model is applicable whenever corpus data are available; no lexical resource is required. Additionally, it does not require the headword observation step and the generalization step (cf. Section 1) to use the same

[5] By keeping the constant $|\bigcup_{a'} \mathsf{Seen}_{\mathsf{rv}}(a')|$, we guarantee that the fraction remains larger than one, and $wt_{r,v}(a)$ remains positive. This is to ensure that the weighted average in Equation (6) yields correct results.

corpus.[6] This allows us to work with a relatively small and deeply linguistically analyzed corpus of seen headwords, the FrameNet corpus, while using a much bigger data set to generalize over seen headwords. It also allows us to make predictions for the potentially deeper relations annotated in the primary corpus, for example, semantic roles. We will investigate the potential of this setup in our Experiments 1 and 2.

As a distributional model, EPP avoids the two pitfalls of resource-based models. One is a coverage problem due to the limited size of the resource (see the task-based evaluation in Gildea and Jurafsky [2002]). For example, the semantic role–based PADO ET AL. model resorts to class-based smoothing methods to improve coverage, which EPP does not need. The other problem of resource-based models is that the shape of the WordNet hierarchy determines the generalizations that the models make. These are not always intuitive. For example, Resnik (1996) observes that (*answer, obj, tragedy*) receives a high preference because *tragedy* in WordNet is a type of written communication, which is a preferred argument class of *answer*.

The ROOTH ET AL. model (Rooth et al. 1999) shares the resource independence of EPP, but has complementary benefits and problems. Querying the probabilistic ROOTH ET AL. model takes only constant time, whereas querying the exemplar-based EPP model takes time linear in the number of seen arguments for the argument position. However, the ROOTH ET AL. model requires a dedicated training phase with a space complexity linear in the total number of verbs and nouns, which can lead to practical problems for large corpora (cf. Section 5.1). The separation of similarity computation and headword observation in EPP also gives the experimenter more fine-grained control over the types and sources of information in the model.

The EPP model looks superficially similar to the model of Dagan, Lee, and Pereira (1999). However, they differ in the role of the similarity measure: The Dagan, Lee, and Pereira model computes a co-occurrence probability, and it uses similarity as a weighting scheme. The EPP model computes similarity (of a word to the typical fillers of an argument position), and its weighting schemes are separate from the similarity measure. The two models also differ in the kinds of items they consider as a basis for generalization (or smoothing): In computing the probability of seeing a word w_2 after w_1, the sum in the Dagan, Lee, and Pereira model runs over all words that are similar to w_1, whereas the sum in the EPP model runs over all words that have been seen as headwords in the argument position in question. Given that occurrence in an argument position is a form of co-occurrence, and similarity (in both models) is computed on the basis of vectors derived from co-occurrence counts, one could say that the sum in the EPP model runs over words determined by first-order co-occurrence, whereas the sum in Dagan, Lee, and Pereira runs over words chosen through second-order co-occurrence (where w_1 and w_2 are second-order co-occurring if they both tend to occur with the same words w_3).

4. Design of the Experimental Evaluation

In this section, we give a high-level overview over the experiments and experimental settings we will use subsequently. Details will be provided in the following sections.

We evaluate the EPP model in three ways: We test the prediction of verbal selectional preference models with a pseudo-disambiguation task (Experiment 1). Then, we address the task of predicting human verb–argument plausibility ratings (Experiment 2). Finally, we investigate inverse selectional preferences—preferences of

[6] Dagan, Lee, and Pereira (1999) could in principle do the same, but do not explore this option.

nouns for the predicates that they co-occur with—again using pseudo-disambiguation (Experiment 3).

We compare the EPP model to models from the three model categories presented in Section 2: RESNIK as a hierarchical model; ROOTH ET AL. as a distributional model; and PADO ET AL. as a semantic role–based model. As both Brockmann and Lapata (2003) and Padó (2007) have argued, no WordNet-based model systematically outperforms the others, and the RESNIK model shows the most consistent behavior across different scenarios. Among the distributional models, we choose ROOTH ET AL. as a model that performs soft clustering and thus shows a marked difference to the EPP model. To our knowledge, this is the first comparison of all three generalization paradigms: semantic hierarchy–based, distributional, and semantic role–based.[7]

As mentioned earlier, we employ two tasks to evaluate the four models: pseudo-disambiguation and the prediction of human plausibility ratings. The pseudo-disambiguation task (Yarowsky 1993) has become a standard evaluation measure for selectional preference models (Dagan, Lee, and Pereira 1999; Rooth et al. 1999). Given a choice of two potential headwords, the task of a selectional preference model is to pick the more plausible one to fill a particular argument position of a given predicate. Pseudo-disambiguation can be viewed as a word sense disambiguation task in which the two potential headwords together form a "pseudo-word," for example *herb/struggle* from the original words *herb* and *struggle*. The task is to "disambiguate" the pseudo-word to the word that fits better in the given context. It can also be viewed as an in vitro version of semantic role labeling and dependency parsing (depending on whether the relations are semantic roles or grammatical functions) (Zapirain, Agirre, and Màrquez 2009). In this case, the scenario is that of a sentence containing a predicate and two words that could potentially fill an argument position of that predicate, for example, the predicate *recommend* with the potential headwords *herb* and *struggle* for the grammatical relation of direct object. The task is to decide which of the two potential headwords is better suited to fill the argument position.

Human plausibility ratings, on the other hand, make considerably more fine-grained distinctions than those occurring in pseudo-disambiguation tasks. Here, models predict the exact human ratings for verb–argument–role triples. Ratings are collected to further control carefully selected experimental items for psycholinguistic studies (Trueswell, Tanenhaus, and Garnsey 1994; McRae, Spivey-Knowlton, and Tanenhaus 1998), or are solicited for corpus-derived triples specifically to create evaluation data for plausibility models (Brockmann and Lapata 2003; Padó 2007).

We contrast two different levels of semantic analysis for the predicates and argument positions. In the **SEM PRIMARY** setting, the predicates are FrameNet frames, each of them potentially instantiated by multiple different verbs. The argument positions in these settings are frame-semantic roles. This setting most closely matches the notion of selectional preferences as characterizations of semantic arguments of an event. In addition, we study the **SYN PRIMARY** setting, where predicates are verbs, and argument positions are grammatical functions (subject and direct object). Viewing grammatical functions as shallow approximations of semantic roles, we can expect the selectional preference models for this setting to yield noisier estimates than in the SEM PRIMARY setting. The two settings will differ only in the choice of primary corpus, but will use the same generalization corpus.

[7] Erk (2007) has a comparison between hierarchy-based and distributional models, but does not include a semantic role–based model.

Table 3 illustrates the difference between the SEM PRIMARY setting and the SYN PRIMARY setting on an example from a pseudo-disambiguation task: The SEM PRIMARY setting has predicates like the FrameNet frame (predicate sense) ADORNING, with the semantic role THEME as argument position. In contrast, the SYN PRIMARY setting has predicates that are verb lemmas, such as *cause*, and argument positions that are grammatical functions (subj). In both settings, the two potential headwords (here called **headword** and **confounder**, to be explained in more detail in the next section) to be distinguished in the pseudo-disambiguation task are noun lemmas.

The verb–dependency–headword tuples of the SYN PRIMARY setting yield much more coarse-grained and noisy characterizations of selectional preferences; however, they can be extracted from corpora with only syntactic annotation. We are therefore able to use the 100-million word BNC (Burnard 1995) as the primary corpus for this setting by parsing it with the Minipar dependency parser (Lin 1993). Minipar could parse almost all of the corpus, resulting in 6,005,130 parsed sentences.

For the SEM PRIMARY setting, we require a primary corpus with role-semantic annotation. We use the much smaller FrameNet corpus (Fillmore, Johnson, and Petruck 2003). FrameNet is a semantic lexicon for English that groups words in semantic classes called frames and lists fine-grained semantic argument roles for each frame. Ambiguity is expressed by membership of a word in multiple frames. Each frame is exemplified with annotated example sentences extracted from the BNC. The FrameNet release 1.2 comprises 131,582 annotated sentences (roughly three million words). To determine headwords of the semantic roles, the corpus was parsed using the Collins (1997) parser.

As generalization corpus, we use the Minipar-parsed BNC in both settings. The experimentation with two different primary corpora allows us to directly study the influence of the disambiguation of predicates and the semantic characterization of argument positions on the performance of selectional preference models. Note, however, that the comparison is complicated by differences between the two corpora: The primary corpus for the SYN PRIMARY setting is parsed automatically, which can introduce noise in the determination of predicates, grammatical functions, and headwords. The primary corpus for the SEM PRIMARY setting is manually annotated for semantics but is parsed automatically to determine headwords. This can introduce noise in the headwords, but not in the determination of predicates and semantic roles. Also, the primary corpus for the SYN PRIMARY setting is much larger than the one used in the SEM PRIMARY setting.

5. Experiment 1: Pseudo-Disambiguation

The first experiment uses a pseudo-disambiguation task to evaluate the models' performance on modeling the plausibility of nouns as headwords of argument positions of verbal predicates.

Table 3
Pseudo-disambiguation items for the SYN PRIMARY setting and the SEM PRIMARY setting.

Setting	Predicate (v)	Arg. pos. (r)	Headword (a)	Confounder (a')
SYN	cause	subj	succession	island
	appear	subj	feasibility	desire
SEM	ADORNING	THEME	illustration	axe
	ROPE_MANIPULATION	ROPE	cord	literature

```
Require: Some corpus $T$: a list of triples $(v, r, a)$ of seen predicates, roles, and arguments.
Require: Some corpus $N$: a list of noun lemmas, along with a function $freq_N : N \rightarrow \mathbb{N}$
    that associates each noun $n \in N$ with its corpus frequency.
$N_{mid} = \{n \in N \mid freq_N(n) \geq 30 \text{ and } freq_N(n) \leq 3,000\}$
We define a probability distribution $p_N$ over the $n \in N_{mid}$ by $p_N(n) = \frac{freq_N(n)}{\sum_m freq_N(m)}$
$conf = \{\ \}$ # set of headword/confounder mappings, starts empty
$A_T = \{a \mid (v, r, a) \in T\}$ # set of seen headwords
for every $a$ in $A_T$ do
    choose a confounder $a' \in N_{mid}$ according to $p_N$
    $conf = conf \cup \{\ a \mapsto a'\ \}$
end for
Return: $conf$
```

Figure 2
Algorithm for choosing confounders.

5.1 Setup

Task and data. In a data set of tuples (v, r, a) of a predicate v, argument position r, and headword a, each tuple is paired with a confounder a'. The task is to pick the original headword by comparing the tuples (v, r, a) and (v, r, a'). Table 3 shows some examples.

We begin by collecting all triples (v, r, a) observed in the respective primary corpus. In the SYN PRIMARY setting, this corresponds to all headwords observed in subject or direct object position of a verbal predicate in the BNC, and in the SEM PRIMARY setting, to all nouns observed as headword of some semantic role in a frame introduced by a verb. From this set of triples (v, r, a) for a given primary corpus, we draw an **evaluation sample** that is balanced by the corpus frequency of predicates and argument position. As test set, we choose 100 (v, r) pairs at random, drawing 20 pairs each from five frequency bands: 50–100 occurrences; 100–200 occurrences; 200–500; 500–1,000; and more than 1,000 occurrences. For any chosen predicate–relation pair, we sample triples (v, r, a) equally from six frequency bands of arguments a: 1–50 occurrences; 50–100; 100–200; 200–500; 500-1,000; and more than 1,000 occurrences. These evaluation samples contain a total of 213,929 (SYN) and 65,902 (SEM) tuples.

Next, we pair each headword with a confounder sampled from the primary corpus as described in Figure 2.[8] In the literature, there have been two different approaches to choosing confounders for pseudo-disambiguation tasks: The first approach, used by Dagan, Lee, and Pereira (1999), chooses confounders to match the headword a in frequency. The second approach, used in Rooth et al. (1999), sets the probability that a word is drawn as a confounder to its relative frequency. The advantage and disadvantage of the first approach is that it largely eliminates the frequency bias that is a general problem of vector space-based approaches. This is an advantage in that it allows the generalization achieved by the model to be evaluated without any distortion from frequency bias. It is a disadvantage in that in any practical application making use of selectional preferences, the data will not be frequency-balanced. For example, selectional preferences could be used by a dependency parser to decide which word in the sentence to link to a given verb via a subject edge, or selectional preferences could

[8] The confounder is the same for all instances of the headword a in the evaluation sample, regardless of the values for r and v. As confounder candidates, we only use words with between 30 and 3,000 occurrences in the BNC, following Rooth et al. (1999).

be used by a semantic role labeler to decide which constituent is the overall best filler for the AGENT role for a given predicate. In such cases, it does not appear warranted to assume that the frequencies of different headword candidates are balanced. We choose the second option for our experiments, using relative corpus frequency to approximate the probability of encountering different headword candidates.

Training of models. As stated earlier, we evaluate all models in the SYN PRIMARY setting and the SEM PRIMARY setting. In all experiments herein, we perform two 2-fold cross-validations runs. In each run, we randomly split the respective (SYN or SEM) evaluation sample into a training and a test set at the token level. Figure 3 describes the experimental procedure in pseudo-code.

The EPP, RESNIK, and PADO ET AL. models are trained on the training split of the evaluation sample. The EPP model additionally uses the BNC as generalization corpus in both the SYN PRIMARY setting and the SEM PRIMARY setting. This generalization corpus is used to compute either a WORDSPACE or a DEPSPACE vector space, as discussed in Section 3.3. For the ROOTH ET AL. model, we had to employ a frequency

```
Require: A set $Formalisms$ of formalisms to test
Require: A primary corpus $T$: a list of triples $(v, r, a)$ of seen predicates, argument
    positions, and arguments, along with a function $freq_T : T \rightarrow \mathbb{N}$ that associates each
    triple $(v, r, a) \in T$ with its corpus frequency
Require: A mapping $conf$ : Lemmas $\rightarrow$ Lemmas of headwords to confounders such that
    $\{a \mid (v, r, a) \in T\} \subseteq Domain(conf)$
$eval\_results = \{\ \}$
for splitno in 1:2 do
    # prepare two independent splits
    $half1 = \{\ \}, half2 = \{\ \}$ # mappings from headwords to counts
    for each tuple $t$ in $T$ do
        # decide how many occurrences of t to put in half1, half2 by drawing from the binomial
          distribution
        Sample $k \sim \mathrm{B}(freq_T(t), 0.5)$
        $half1 = half1 \cup \{\ t \mapsto k\ \}, half2 = half2 \cup \{\ t \mapsto freq_T(t) - k\ \}$
    end for
    $splits = \{\ (half1, half2), (half2, half1)\ \}$
    for $(f_{train}, f_{test})$ in $splits$ do
        for each formalism $F$ in $Formalisms$ do
            train a model $m_F$ according to formalism $F$ using the training set defined by
              the frequency function $f_{train}$.
            for each tuple $(v, r, a)$ in $T$ do
                for i in 1:$f_{test}(v, r, a)$ do
                    Evaluate the performance of $m_F$ on the tuple $(v, r, a, conf(a))$ and add the
                      result to $eval\_results$
                end for
            end for
        end for
    end for
end for
Return: $eval\_results$
```

Figure 3
Algorithm for running a pseudo-disambiguation experiment

cutoff of five in the SYN PRIMARY setting to reduce the amount of training data due to memory limitations. The PADO ET AL. model is only used in the SEM PRIMARY setting: FrameNet is an integral part of this model, and it cannot be used in a syntax-only setting without major changes. For details on training, see Section 2.4. Note that no verb classes had to be induced from the data, because the predicates v are already instantiated by verb classes, namely, FrameNet frames (see Table 3).

Finally, we report three baselines. The first baseline, **headword frequency** (HW), is very simple. It decides between the headword a and the confounder a' by comparing the frequencies $f(a)$ and $f(a')$. The second, more informed, baseline is **triple frequency** (TRIPLE). It votes for a if $f(v, r, a) > f(v, r, a')$, and vice versa. The third baseline, a **bigram language model** (LM), was constructed by training a 2-gram language model from the large English ukWAC Web corpus (Baroni et al. 2009) using the SRILM toolkit (Stolcke 2002) with default Good–Turing smoothing. We retained only verbs, nouns, adjectives, and adverbs in order to maximize the proximity between verbs and their subjects and objects. We defined the preference score for verb–subject triples as the probability of the sequence av, that is, $Pref(v, subj, a) = P(v|a)$. Conversely, the preference score for verb–object triples was defined as the probability of the sequence va, that is, $Pref(v, obj, a) = P(a|v)$. Again, the model compares $Pref(v, r, a)$ and $Pref(v, r, a')$ to make its decision.

Evaluation. For all models, we report two evaluation figures. One is **coverage**: A tuple is covered if the model assigns some preference to *both* a and a', and the preferences are not equal. The second is **error rate**, which is the relative frequency, among all **covered** tuples, of instances where the confounder was at least equally preferred. Both coverage and error rate are averages over the 2 x 2 cross-validation runs in each setting.

We determine the statistical significance of differences between error rates using bootstrap resampling (Efron and Tibshirani 1994). This procedure samples corresponding model predictions with replacement from the set of predictions made by the models to be compared and computes the difference in error rates. On the basis of n such samples (n = 1,000), the empirical 95% confidence interval for the difference in strength on the basis of all observed differences is computed. If the interval includes 0, the difference is not statistically significant.

5.2 SYN PRIMARY Setting: Results

Table 4 shows the results for the SYN PRIMARY setting. The overall best error rate is achieved by a variant of the EPP model, with the RESNIK model coming in second (the performance difference is significant at the 0.05 level). The EPP variants also show near-perfect coverage, whereas the RESNIK model delivers results only for 63% of the data points. We found a very high error rate and a comparatively low coverage for ROOTH ET AL., which most likely stems from the data pruning necessary to reduce the training data (compare the subsequent results in the SEM PRIMARY setting). The PADO ET AL. model was not tested in the SEM PRIMARY setting, because it requires semantic role annotation. The HW baseline is somewhat below chance (50%), which is an effect of our by-token sampling procedure, according to which confounders often have higher corpus frequencies than the real arguments. The TRIPLE baseline has a better error rate than the LM baseline, but has very low coverage. Both the RESNIK and the EPP models outperform the baselines in terms of error rate. That they outperform the TRIPLE baseline in terms of error rate indicates that we sometimes have confounders that have actually been seen more often with the verb–argument pair than the headword, but that

Table 4
SYN PRIMARY setting: Pseudo-disambiguation results for different weighting schemes.

Model	Similarity	Error rate (%)			Coverage (%)
		UNI	FREQ	DISCR	
EPP:DEPSPACE	Cosine	32.8	30.3	31.2	98.5
	Dice	49.4	48.2	47.5	97.1
	nGCM	27.6	27.5	**25.7**	98.5
	Hindle	53.7	52.3	52.8	96.6
	Jaccard	49.5	48.2	47.6	97.1
	Lin	35.5	34.3	33.2	**98.8**
EPP:DEPSPACE, PCA	Cosine	30.2	28.7	28.8	98.1
	Dice	29.9	30.8	28.6	**98.2**
	nGCM	26.4	26.4	**25.6**	98.1
	Hindle	45.0	44.4	44.2	95.7
	Jaccard	29.7	30.7	28.5	**98.2**
	Lin	28.7	29.1	26.7	97.7
EPP:WORDSPACE	Cosine	35.3	35.8	34.0	97.4
	Dice	51.0	50.7	50.3	96.0
	nGCM	33.2	34.7	31.8	97.4
	Hindle	52.7	52.8	52.4	96.0
	Jaccard	51.8	52.0	51.3	96.0
	Lin	32.0	31.8	**31.4**	**98.2**
EPP:WORDSPACE, PCA	Cosine	30.3	31.3	29.4	97.1
	Dice	31.3	32.4	30.5	**97.8**
	nGCM	30.0	30.9	29.0	97.1
	Hindle	40.2	41.0	40.4	95.3
	Jaccard	31.0	32.1	30.2	**97.8**
	Lin	27.8	29.8	**26.9**	97.3
RESNIK			**28.1**		63.4
ROOTH ET AL.			58.1		61.5
PADO ET AL.			–		–
HW			60.0		**100.0**
TRIPLE			32.0		4.0
LM			37.0		86.0

are dissimilar from other seen headwords, which allows RESNIK and EPP to identify them as confounders in spite of their higher co-occurrence frequency.

We now turn to a comparison of the EPP variants. The coverage of all EPP models is very high (0.95 or higher), independent of space, similarity measure, and dimensionality reduction. We generally observe that error rates are lower when word meaning is represented in DEPSPACE, and when discrimination weighting is used. In DEPSPACE, nGCM works best, yielding the overall best result with an error rate of 25.6–25.7%. In WORDSPACE, the Lin measure shows the best error rates with an error rate of just below 27%. These results hold both for the unreduced and the reduced spaces and are highly significant ($p \leq 0.01$). Hindle is clearly the worst measure at around random performance.

The difference between UNI and DISCR is significant throughout; the difference between FREQ and DISCR is less uniform. In DEPSPACE, the difference between the best measure with and without PCA (nGCM in both cases) is not significant; in WORDSPACE, the difference between the best measure with and without PCA (Lin in both cases) is significant ($p \leq 0.01$).

For both WORDSPACEs and DEPSPACEs without PCA, the similarity measures divide into two distinct groups: Lin, nGCM, and Cosine on the one hand and Jaccard, Dice, and Hindle on the other, with a significant difference in performance between the groups ($p \leq 0.01$). The use of dimensionality reduction through PCA improves performance for all similarity measures, in WORDSPACE as well as DEPSPACE. The improvement is especially marked for the Dice and Jaccard measures, which perform at the level of a random baseline for unreduced spaces. We assume that these set intersection-based measures benefit from the independent dimensions that PCA produces. For the similarity measures with best performance, the improvement through PCA is less marked. Thus, PCA-reduced spaces show more similar error rates across similarity measures. After PCA, only nGCM and Lin still significantly ($p \leq 0.01$) outperform the others in DEPSPACE, and in WORDSPACE, Lin is the only measure that performs significantly differently from the rest ($p \leq 0.01$).

As arguments are sampled from six frequency bins, we can inspect the effect of argument frequency on error rate. Figure 4 examines the performance of the EPP model with different similarity measures and weighting schemes by argument frequency bins (cf. the subsection *Task and Data* in Section 5). We find that the overall best weighting scheme, DISCR, also works best for all except the highest argument frequency bin. In the DEPSPACE setting (upper row), all similarity measures show a frequency bias in that

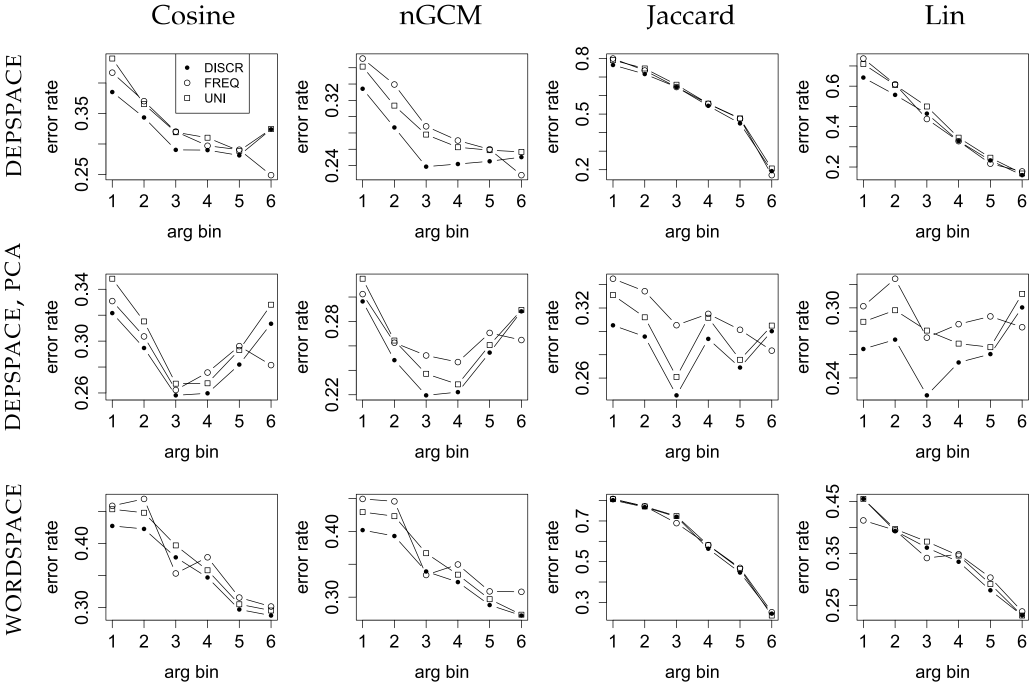


Figure 4
SYN PRIMARY setting: Error rate by *argument* frequency bin. Bins: 1 = 1–50; 2 = 50–100; 3 = 100–200; 4 = 200–500; 5 = 500–1,000; 6 > 1,000.

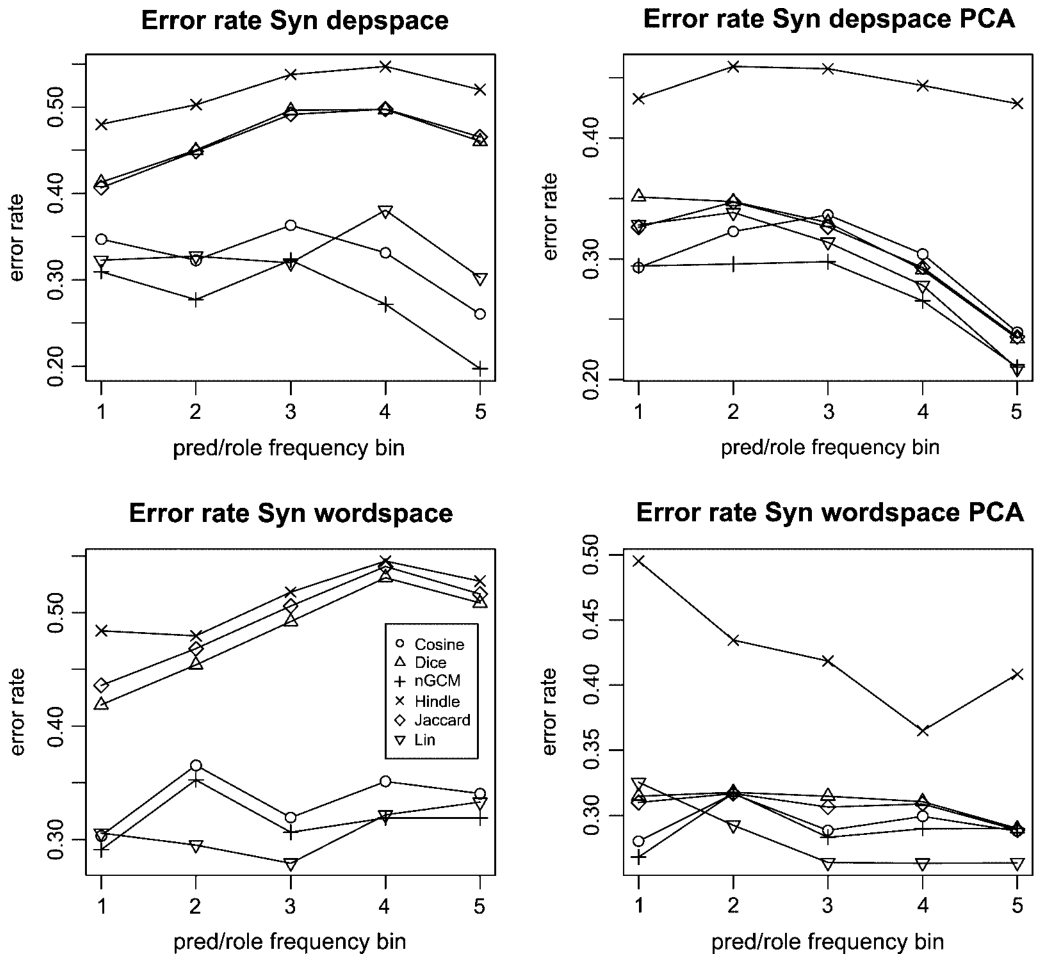


Figure 5
SYN PRIMARY setting: Error rate by *predicate* frequency bin: DISCR weighting. Bins: 1 = 50–100; 2 = 100–200; 3 = 200–500; 4 = 500–1,000; 5 > 1,000.

error rate is lower for more frequent arguments, but this bias is much less pronounced in Cosine and nGCM than in the other measures, with error rates varying between 45% and 25% rather than 80% and 20%. (Dice and Hindle, not shown here, exhibit similar behavior to Jaccard.) In PCA-transformed DEPSPACE (middle row), this frequency bias largely disappears for all similarity measures. In WORDSPACE (bottom row), although there is again a frequency bias in all similarity measures, Lin now joins Cosine and nGCM in being much less biased than Jaccard, Dice, and Hindle. For WORDSPACE with PCA-transformation, not shown here, the curves resemble those of DEPSPACE with PCA-transformation.

Figure 5 examines the effect of (predicate, argument position) pair frequency on error rate. Predicate–argument position pairs were sampled from five frequency bins. The figure shows DISCR weighting only. In the spaces without dimensionality reduction, there is a clear division between Cosine, nGCM, and Lin on the one hand, and Jaccard, Dice, and Hindle on the other. In PCA spaces, all measures except for Hindle are similar in their performance. In both DEPSPACE conditions, error rate decreases towards the higher frequency predicate bins, although this is not so in

WORDSPACE. It seems that in the sparser DEPSPACE, models can still profit from the additional seen headwords in the highest predicate frequency bins, whereas in the less sparse but noisier WORDSPACE, the added noise is stronger than the added signal in the highest predicate frequency bins. For the lowest predicate frequency bins, the best results in WORDSPACE are better than those in DEPSPACE.

5.3 SEM PRIMARY Setting: Results

Table 5 shows the results for the SEM PRIMARY setting, where we predict head words for pairs of a frame (predicate sense) and semantic role. In comparison to the SYN PRIMARY setting (Table 4), error rates are lower across the board. The difference for the EPP models is on average around 10%.

Table 5
SEM PRIMARY setting: Pseudo-disambiguation results for different weighting schemes.

Model	Similarity	Error rate (%)			Coverage (%)
		UNI	FREQ	DISCR	
EPP:DEPSPACE	Cosine	19.8	16.9	19.0	97.1
	Dice	42.3	32.4	39.5	96.3
	nGCM	20.2	**16.3**	20.8	97.1
	Hindle	48.3	46.8	47.8	93.4
	Jaccard	41.5	31.5	38.5	96.3
	Lin	31.1	20.2	29.0	**97.7**
EPP:DEPSPACE, PCA	Cosine	18.5	17.0	17.8	96.9
	Dice	19.3	19.2	18.0	97.6
	nGCM	16.9	**15.7**	16.4	96.9
	Hindle	44.9	45.6	44.7	89.3
	Jaccard	18.2	18.5	17.5	97.6
	Lin	18.8	19.5	18.3	**98.9**
EPP:WORDSPACE	Cosine	24.1	20.4	23.4	**93.1**
	Dice	23.7	24.5	22.5	89.6
	nGCM	21.1	**17.8**	19.4	**93.1**
	Hindle	31.8	33.1	31.8	83.1
	Jaccard	24.8	26.5	24.2	89.6
	Lin	22.3	18.4	21.9	92.8
EPP:WORDSPACE, PCA	Cosine	21.0	17.6	20.5	93.1
	Dice	18.5	16.4	17.8	96.8
	nGCM	19.7	16.4	19.3	93.1
	Hindle	41.0	39.8	40.7	90.6
	Jaccard	18.1	**16.2**	17.6	96.8
	Lin	21.3	17.1	20.7	**98.3**
RESNIK			16.5		62.8
ROOTH ET AL.			24.9		**100.0**
PADO ET AL.			**7.1**		59.0
HW			65.0		**100.0**
TRIPLE			44.0		2.0
LM			NA		NA

The error rate of the PADO ET AL. model, at 7%, is the best by a large margin. We attribute this to the extensive generalization mechanisms that the model uses, which draw on an array of lexical–semantic resources. However, with a coverage of 59%, the model is still unable to make predictions for many of the test items. Error rates for the RESNIK and the EPP models are comparable, at 16.5% for RESNIK and 15.7% for the best EPP variant. The two models differ sharply in coverage, however: 62.8% for RESNIK, consistent with the findings of Gildea and Jurafsky (2002), and between 90% and 98% for EPP variants. The RESNIK model also profits from the presence of semantic disambiguation in the SEM PRIMARY setting (in the SYN PRIMARY setting its error rate was 28%), which underlines the substantial impact that properties of the training data have on semantic hierarchy–based models of selectional preferences. ROOTH ET AL. now has perfect coverage, affirming our assumption that the very bad results of the ROOTH ET AL. model in the SYN PRIMARY setting were an artifact of the data sampling necessary for that data set. Although its error rate of 24.9% is a substantial improvement over all baselines, the EPP model achieves error rates that are up to 9 points lower at a comparable coverage. Among the baselines, HW shows that here, as in the SYN PRIMARY setting, arguments have some tendency of having lower frequency than the confounders. The TRIPLE baseline shows near-random performance, at very low coverage, a result of the very small size of the corpus. Because there is no large corpus with frame-semantic roles, nor is the annotation easily linearizable, we could not compute a LM baseline in the SEM PRIMARY setting.

Among EPP models, the DEPSPACEs and WORDSPACEs perform comparably, with a non-significant advantage for DEPSPACE among the best models. Overall error rates show the same clear divide between the three high-performing similarity measures (Cosine, nGCM, and Lin) and the three weaker ones (Dice, Jaccard, and Hindle). Dimensionality reduction again dramatically improves the weaker models, with Jaccard yielding the best result for the PCA-reduced WORDSPACE.[9] Whereas all best parametrizations in the SYN PRIMARY setting used DISCR weighting, it is now FREQ weighting that yields the best results.

Figure 6 again analyzes the influence of argument frequency on performance by showing the performance of different variants of the EPP model over six argument frequency bins. The upper row shows DEPSPACE without dimensionality reduction. Note that FREQ weighting now works especially well for the lowest argument frequency bin, much better than DISCR and PLAIN. This is the opposite of what we saw for the SYN PRIMARY setting in Figure 4. With DISCR and PLAIN weighting, Jaccard and Lin again have noticeable problems with the lowest argument frequency bins—as in the SYN PRIMARY setting—but not with FREQ weighting. With DEPSPACE and dimensionality reduction (middle row), we get error rates of $\leq$ 26% for all settings and all frequency bins. On the lowest frequency bin, we again see a large advantage of FREQ weighting over the two other weighting schemes. The bottom row shows WORDSPACE without dimensionality reduction. Note that there is much less variation in error rates across frequency bins here than in unreduced DEPSPACE.

Figure 7 charts error rate by predicate frequency bin, showing FREQ weighting only, as this showed the best results on this data set. The figure clearly illustrates the divide between the top and the bottom three similarity measures in DEPSPACE, as well as the disappearance of this divide for both PCA settings. In unreduced WORDSPACE,

[9] The differences to other similarity metrics in the FREQ setting are insignificant, with the exception of Hindle.

the divide is not as clearly visible. The figure also indicates a slight tendency for error rates to rise for the lowest-frequency as well as the highest-frequency predicates, across all spaces.

5.4 Discussion

The resource-based approaches that we tested, RESNIK and PADO ET AL., show superior performance when they have coverage (which coincides with findings in other lexical semantics tasks that supervised data, when available, always increases performance), but showed low coverage, at most 63% (RESNIK, SYN PRIMARY setting). The EPP model achieves near-perfect coverage at good error rates: In the SYN PRIMARY setting, the RESNIK model achieved an error rate of 28%, and the best EPP variant was at 26%. In the SEM PRIMARY setting, error rates were 7% for the PADO ET AL. model, 16.5% for the RESNIK model, and 16% for the best EPP variant. Comparing the EPP and ROOTH ET AL. models in the SEM PRIMARY setting, we find that the use of an additional generalization corpus in the EPP model seems to offset any advantages introduced by the joint clustering of predicates and arguments.

The difference in model performance on the two primary corpora (SYN and SEM) is striking. Even though the FrameNet corpus is smaller and a sparse data problem might be expected, models perform at considerably lower error rates in the SEM PRIMARY setting than when the primary corpus is the larger BNC. This underscores the point that selectional preferences belong to a predicate *sense* rather than a predicate lemma, and that they describe the semantics of fillers of semantic roles rather than of

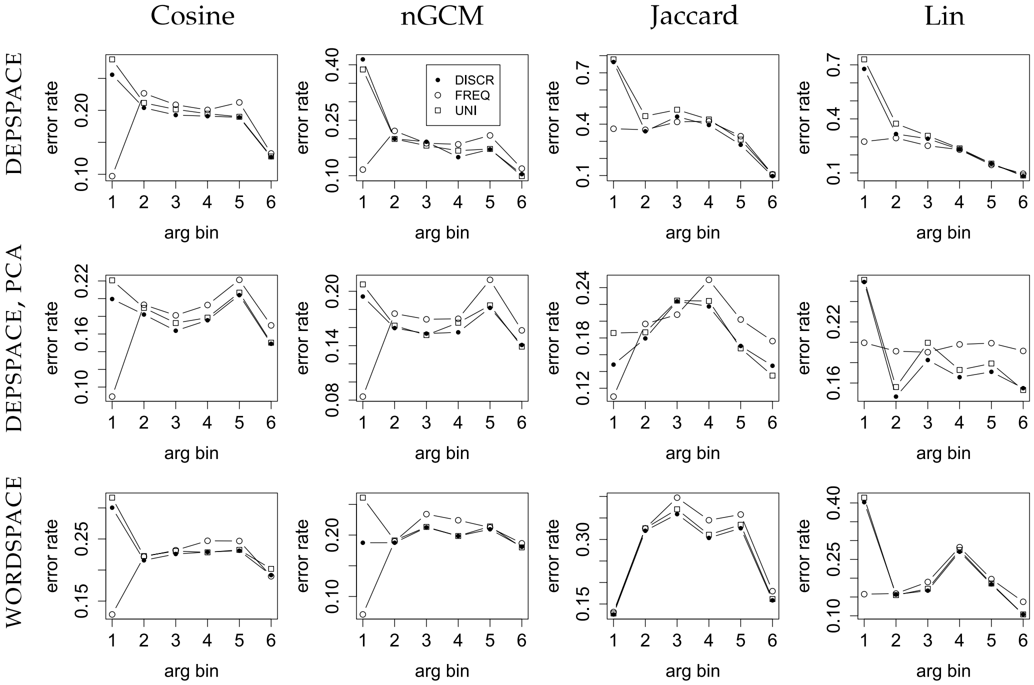


Figure 6
SEM PRIMARY setting: Error rate by *argument* frequency bin. Bins: 1 = 1–50; 2 = 50–100; 3 = 100–200; 4 = 200–500; 5 = 500–1,000; 6 > 1,000.

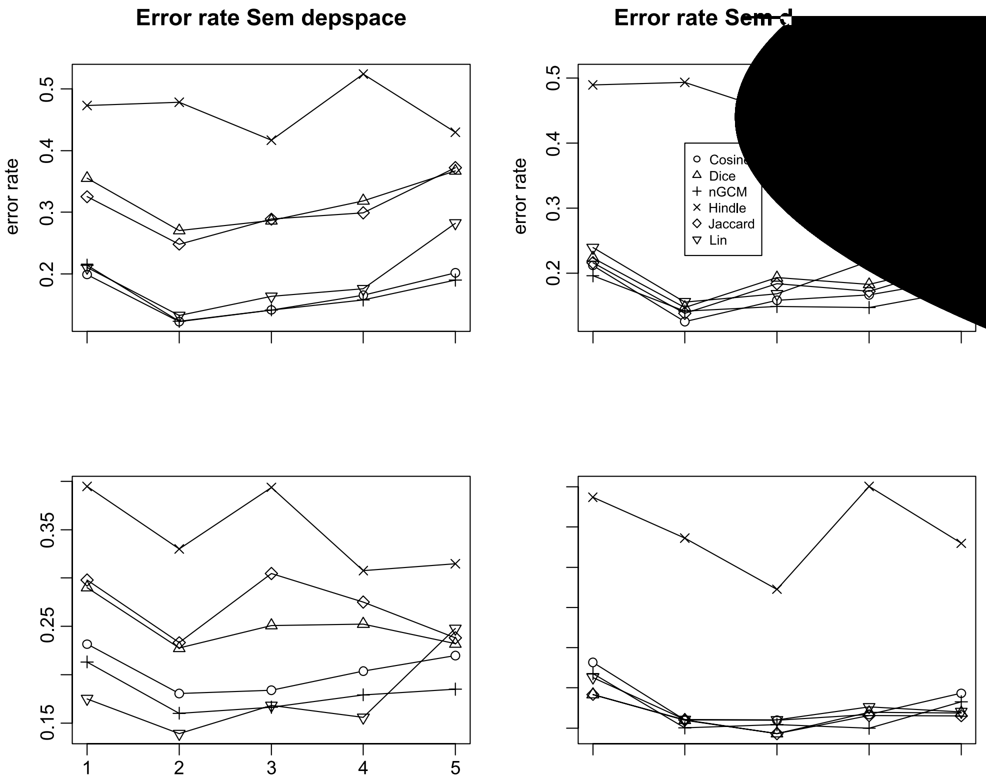

Error rate Sem depspace
Error rate Sem d
error rate
error rate
0.5
0.4
0.3
0.2
0.35
0.25
0.15
1
2
3
4
5
Cosin
Dice
nGCM
Hindle
Jaccard
Lin

Table 6
Verb–argument position–noun triples with plausibility judgments on a 7-point scale (McRae et al., 1998).

Verb	Argument position	Noun	Plausibility
shoot	agent	hunter	6.9
shoot	patient	hunter	2.8
shoot	agent	deer	1.0
shoot	patient	deer	6.4

the SYN PRIMARY setting, as well as in all SEM conditions except reduced WORDSPACE. The Lin measure seems to work well with noisier data: It is the best EPP model when using WORDSPACE in the SYN PRIMARY setting. Cosine, although never showing the top performance, is among the best models in any setting. Although dimensionality reduction only improves the overall error rates of the best models by a few points, it has two important consequences: First, dimensionality reduction reduces dependence of the results on the exact similarity measure chosen, as all measures except Hindle show nearly indistinguishable error rates on reduced spaces (Figures 5 and 7). Second, low-frequency arguments profit by a huge margin when PCA is used (Figures 4 and 6). Among weighting schemes, DISCR weighting seems to be most useful when the data is sparse but somewhat noisy (as is the case in the lower argument frequency bins in the SYN PRIMARY setting). Frequency weighting seems to work best when the data is either not sparse (as in the highest argument frequency bin in the SYN PRIMARY setting) or very clean but sparse (as in the lowest argument frequency bin in the SEM PRIMARY setting). A comparison of the two vector spaces, DEPSPACE and WORDSPACE, shows no clear winner. When the collections of seen headwords are noisier, as they are in the SYN PRIMARY setting, DEPSPACE, with its more aggressive filtering, yields the better results. Sets of headwords collected by predicate sense, as in the SEM PRIMARY setting, are sparser but cleaner, and WORDSPACE shows lower error rates.

6. Experiment 2: Human Plausibility Judgments

Experimental psycholinguistics affords a second perspective on selectional preferences: The plausibility of verb–argument pairs has been shown to have an important effect on human sentence processing (e.g., Trueswell, Tanenhaus, and Gransey 1994; Garnsey et al. 1997; McRae, Spivey-Knowlton, and Tanenhaus 1998). In these studies, plausibility was operationalized as the thematic fit or selectional preference between a verb and its argument in a specific argument position. Models of human sentence processing therefore need selectional preference models (Padó, Crocker, and Keller 2009). Conversely, psycholinguistic plausibility judgments can be used to evaluate computational models of selectional preferences.

6.1 Experimental Materials

We present evaluations on two plausibility judgment data sets used in recent studies.

The first data set consists of 100 data points[10] from McRae, Spivey-Knowlton, and Tanenhaus (1998). Our example in Table 6, which is taken from this data set, was elicited by asking study participants to rate the plausibility of, for example, a hunter shooting (AGENT) or being shot (PATIENT). The data point demonstrates the McRae set's balanced structure: 25 verbs are paired with two argument headwords in two argument positions each, such that each argument is highly plausible in one argument position but implausible in the other (*hunters shoot*, but are seldom shot, and vice versa for *deer*). The resulting distribution of ratings is thus highly bimodal. Models can only reliably predict the human ratings in this data set if they can capture the difference between verb argument positions as well as between individual fillers. However, because the verb–argument pairs were created by hand and with strict requirements, many of the arguments are infrequent in standard corpora (e.g., *wimp, bellboy*, or *knight*). When FrameNet is used to annotate senses for the verbs, no appropriate senses are available for 28 of the 100 verb–argument pairs, reducing the test set to 72 data points.

The second, larger data set addresses this sparseness issue. Its triples are constructed on the basis of corpus co-occurrences (Padó 2007). Eighteen verbs are combined with their three most frequent subjects and objects found in the Penn Treebank and FrameNet corpora, respectively, up to a total of 12 arguments. Each verb–argument pair was rated both as an agent and as a patient (i.e., both in the observed and an unobserved argument position), which leads to a total of 24 rated triples per verb. The data set contains ratings for 414 triples. The resulting judgments show a more even distribution of data. With FrameNet annotation for the verbs, appropriate senses are not attested for six verb–argument pairs, reducing the test set to 408 data points.

6.2 Setup

We evaluate the same four models as in Experiment 1: EPP, the WordNet-based RESNIK model, the distributional ROOTH ET AL. model, and the semantic role–based PADO ET AL. model. We again compare a SYN PRIMARY setting, where the models make predictions for pairs of a verb and a grammatical function, with a SEM PRIMARY setting, for which the two test data sets were annotated with verb sense and semantic roles in the FrameNet paradigm (Padó 2007) and where models make predictions for pairs of a frame and a semantic role. As before, the PADO ET AL. model is only tested in the SEM PRIMARY setting.[11] For the EPP model, we focus on parsed, dimensionally unreduced spaces and DISCR weighting, following earlier results (Padó, Padó, and Erk 2007). We provide results for the best WORDSPACE models from Experiment 1 for comparison. The primary corpora for training selectional preference models were prepared as in Experiment 1 (cf. Section 5.1). The generalization corpus for EPP was again the BNC. For the ROOTH ET AL. model in the SYN PRIMARY setting, we again used a frequency cutoff. We found the RESNIK model to perform better when using just a subset of the BNC (namely, all the triples for verbs present in the test set).

6.3 Evaluation Procedure

We evaluate our models by correlating the predicted plausibility values with the human judgments, which range between 1 and 7. Because we do not assume a priori that there

[10] The original data set has 60 data points more, which were used as the development set for the PADO ET AL. model.

[11] The PADO ET AL. model now uses automatically induced verb clusters instead of FrameNet frames.

Table 7
Comparison of EPP DEPSPACE models on McRae data. Unreduced spaces, DISCR weighting. ***p < 0.001.

	SEM			SYN		
Sim	Coverage	Spearman's ρ		Coverage	Spearman's ρ	
Dice	100%	0.038	ns	98%	0.148	ns
Jaccard	100%	0.045	ns	98%	0.153	ns
Cosine	100%	**0.162**	**ns**	98%	0.197	ns
Hindle	100%	0.060	ns	98%	0.108	ns
Lin	100%	0.085	ns	98%	0.094	ns
nGCM	100%	0.154	ns	98%	**0.325**	***

is a linear correlation between the two variables, we do not use Pearson's product-moment correlation, but instead Spearman's ρ, a non-parametric rank-order correlation coefficient.[12] Note that significance is harder to reach the smaller the number of data points is.

In line with Experiment 1, we include a simple frequency baseline FREQ, which predicts the plausibility of each item as its frequency in the BNC (SYN) and in FrameNet (SEM), respectively. With regard to an upper bound, we assume that automatic models of plausibility should not be expected to surpass the typical human agreement on the plausibility judgment. This is roughly $\rho \approx 0.7$ for the Pado data set.

6.4 McRae Data Set: Results and Discussion

Table 7 focuses on EPP variants with unreduced DEPSPACE for the McRae data set. We see that this data set is rather difficult to model. None of the models trained in the SEM PRIMARY setting achieves a significant correlation.[13] Apparently, the FrameNet corpus is too small to acquire selectional preferences that generalize well to the infrequent items that make up the McRae data set. In the SYN PRIMARY setting, the nGCM model's predictions reach significance.

Table 8 shows results on the McRae data set for all selectional preference models that we are considering. For EPP, we only show nGCM as the best-performing similarity measure from the pseudo-disambiguation task, and Cosine as a widely used vanilla measure. The results for the SEM PRIMARY setting (left-hand side) mirror the results for the SEM PRIMARY setting in Experiment 1: The deep PADO ET AL. model shows the best correlation (it is the only model to predict human judgments significantly). It overcomes the sparseness in the FrameNet corpus by using semantic verb classes that are particularly geared towards grouping the existing verb occurrences in the way that is most meaningful for this task. It covers about 80% of the test data. EPP has full coverage, and although it does not make statistically significant predictions, it shows substantially higher correlation coefficients than ROOTH ET AL. and RESNIK.

[12] A second concern is the computation of significance values: The methods most widely used for the Pearson coefficient (Student's t-distribution, Fisher transformation) assume that the variables are normally distributed, which is not the case in our data set. For Spearman's ρ, we use the algorithm by Best and Roberts (1975), which does not make this assumption.

[13] *Significance* here refers to significance of correlation with the human data, not significance of differences between models.

Table 8
Comparison across models on McRae data. $^{**}p < 0.01$, $^{***}p < 0.001$.

	SEM			SYN		
Model	Coverage	Spearman's ρ		Coverage	Spearman's ρ	
EPP (DEPSPACE nGCM)	100%	0.154	ns	98%	**0.325**	***
EPP (DEPSPACE Cosine)	100%	**0.162**	**ns**	98%	0.197	ns
RESNIK	100%	–0.041	ns	100%	0.123	ns
ROOTH ET AL.	67%	0.078	ns	48%	**0.465**	***
PADO ET AL.	78%	**0.415**	**	–	–	
EPP (WORDSPACE Lin)	100%	0.138	ns	98%	0.062	ns
EPP (WORDSPACE nGCM)	100%	0.167	ns	98%	0.110	ns
FREQ	18%	0.087	ns	36%	0.103	ns

The DEPSPACE and WORDSPACE variants of EPP perform similarly here, and the simple frequency baseline has very low coverage and correlation.

As the right-hand side of Table 8 shows, both ROOTH ET AL. and EPP achieve better results in the SYN PRIMARY setting than in the SEM PRIMARY setting. The ROOTH ET AL. model obtains a highly significant correlation. The combination of infrequent headwords in the McRae data set and the large primary corpus brings out the benefits that the ROOTH ET AL. model can derive from generalizing from verbs and nouns to the latent classes via soft clustering. Unfortunately, its coverage is still quite low (48%), and for this reason, the difference from the best EPP model is not significant.[14] In the SYN PRIMARY setting, the EPP DEPSPACE models clearly outperform the WORDSPACE because of the DEPSPACE models' more aggressive filtering. Interestingly, RESNIK still performs poorly in the SYN PRIMARY setting: WordNet does not make the right generalizations to capture the selectional preferences at play in the McRae data, no matter how much training data is available. This is underscored by an analysis of which WordNet classes were most frequently determined as the strongest association with the target verbs: The classes *entity*, *person*, and *physical object* are assigned in 60 out of 100 test cases for the McRae data (SYN PRIMARY setting), a data set where plausibility is determined by factors much more fine-grained than animacy. (In the SEM PRIMARY setting, the picture is similar with classes *person*, *organism*, and *entity* assigned in 48 out of 72 test cases.) The frequency baseline again performs badly.

6.5 Pado Data Set: Results and Discussion

We now turn to the Pado data set. Again, we first focus on the performance of different similarity measures in EPP using unreduced DEPSPACE (Table 9). Correlation with human judgments is much better than for the McRae data set, and highly significant for all SEM PRIMARY setting models and three of the SYN PRIMARY setting models. In both settings, Cosine and Lin are the best measures (difference not significant), followed by nGCM. Hindle comes out worst once more. The difference between the strong and

[14] As in Experiment 1, we apply bootstrap resampling to determine the significance of differences between models. This procedure also takes differences in coverage into account—specifically, a significant difference becomes harder to achieve as the number of data points shared between the models shrinks.

Table 9
Comparison of EPP DEPSPACE parametrizations on Padó data. Unreduced spaces, DISCR weighting. **p < 0.01; ***p < 0.001.

	SEM			SYN		
Sim	Coverage	Spearman's ρ		Coverage	Spearman's ρ	
Dice	100%	0.289	***	100%	0.026	ns
Jaccard	100%	0.285	***	100%	0.023	ns
Cosine	100%	**0.508**	***	100%	**0.403**	***
Hindle	100%	0.160	**	100%	–0.004	ns
Lin	100%	0.498	***	100%	0.229	***
nGCM	100%	0.384	***	100%	0.156	**

weak measures is more pronounced for the SYN PRIMARY setting, compared with the SEM PRIMARY setting. Coverage is at or close to 100% throughout.

Table 10 shows results on the Pado data set for all selectional preference models that we consider. In the SEM PRIMARY setting (where both the data and the primary corpus have FrameNet annotation), EPP and the deep PADO ET AL. model predict the human judgments similarly well (difference not significant). Because all verbs in this data set are covered by FrameNet, the PADO ET AL. model also shows a nearly perfect coverage. EPP and PADO ET AL. do much better than ROOTH ET AL. (differences significant at $p \leq 0.01$). ROOTH ET AL. has the lowest coverage at 88%, but this is still higher than its coverage of the McRae data. As with the McRae data, ROOTH ET AL. achieves better correlation in the SYN PRIMARY setting than the SEM PRIMARY setting, indicating that the frequency cutoff does not harm performance as much in Experiment 2 as it did in Experiment 1. However, the coverage of ROOTH ET AL. is lower in the SYN PRIMARY setting, perhaps because the SEM PRIMARY setting smoothes rare verbs by grouping them in frames with other verbs. RESNIK also achieves better correlation in the SYN PRIMARY setting, but recall that it was trained on a subset of the BNC only to reduce noise in the training data—when trained on the whole BNC set, performance degrades to $\rho = 0.060$. The difference from the best EPP model remains numerically large. As for

Table 10
Comparison across models on Padó data. ***p < 0.001.

	SEM			SYN		
Model	Coverage	Spearman's ρ		Coverage	Spearman's ρ	
EPP (DEPSPACE Cosine)	100%	0.489	***	98%	**0.470**	***
EPP (DEPSPACE nGCM)	100%	0.393	***	98%	0.328	***
RESNIK	98%	0.230	***	97%	0.317	***
ROOTH ET AL.	88%	0.060	ns	58%	0.200	***
PADO ET AL.	97%	**0.515**	***	–	–	
EPP (WORDSPACE Lin)	100%	0.254	***	100%	0.056	ns
EPP (WORDSPACE nGCM)	100%	0.192	***	100%	0.078	ns
FREQ	32%	–0.041	ns	69%	0.090	ns

the McRae data set, the EPP WORDSPACE models show much worse performance than the DEPSPACE models, and do not significantly predict the human plausibility ratings.

The frequency baseline shows a considerably better coverage for this data set, but its correlations hover around zero, which underlines our intuition that verb–argument combinations can be plausible without being frequent in corpora. An example is the combination *(to) embarrass (an) official*, which is rated as highly plausible, but occurs only once each in the BNC and FrameNet.

6.6 Discussion

The McRae data set seems in general more difficult to account for than the Pado data set, as noted by Padó, Padó, and Erk (2007). They explain it by a general frequency effect in the BNC data (which are a superset of the FrameNet data): The median frequency of the hand-selected McRae nouns in the BNC is 1,356, as opposed to 8,184 for the corpus-derived Pado nouns.

Comparing all selectional preference models, we find that the RESNIK and the ROOTH ET AL. models generally do worse than EPP both in terms of coverage and quality of predictions. One notable exception is the excellent performance of the ROOTH ET AL. model on the McRae data in the SYN PRIMARY setting, which comes, however, with a low coverage of less than 50%. A closer inspection of the predictions showed that ROOTH ET AL. makes many predictions for verb–object pairs but abstains from subjects, thus reducing the complexity of the task. For only 20% of verbs, predictions are made for subjects and objects. As noted in Padó, Padó, and Erk (2007), the relatively poor performance of the RESNIK model may be explained by the fact that its ability to generalize is limited to the structure of WordNet, where some semantic distinctions are easier to make than others. For example, a fairly easy distinction to make for WordNet-based models is animate vs. inanimate. Because the Pado set contains a portion of inanimate arguments with animate counterparts, the RESNIK model does well on those. In contrast, in the McRae test set, all arguments are animates, and thus similar to one another in terms of WordNet.

The deep PADO ET AL. model achieves the best correlation with the human judgments on both data sets, but it is limited to the SEM PRIMARY setting. Although the best model is not always among the EPP DEPSPACE models, they consistently show a coverage of close to 100%, and are generally statistically indistinguishable from the best model. Unlike ROOTH ET AL. and RESNIK, whose performance varies widely between the SEM PRIMARY setting and the SYN PRIMARY setting, the correlation coefficients for the EPP models are generally similar across settings. We take this as evidence that EPP models can extract relevant information from deeper annotation on small corpora as well as from large, but noisy and shallow, training data.

Finally, we consider the different similarity measures for the EPP model evaluated on unreduced DEPSPACE. The picture differs somewhat between the two data sets, but the Cosine measure performs well overall, with Lin and nGCM generally in second and third place. So, the group of the three best similarity measures is the same as in Experiment 1, but Cosine shows better performance. One possible reason for this lies in the verb frequency, which is relatively high in both data sets: 68% of the McRae verbs and 83% of the Pado verbs have BNC frequencies of 1,000 and more, whereas Experiment 1 used an equal number of predicates from five frequency bins, the highest being 1,000 and more occurrences. In that highest predicate frequency bin, Cosine consistently performed as well as Lin or better in Experiment 1 (Figures 5 and 7).

7. Experiment 3: Inverse Selectional Preferences

The term **selectional preference** is typically used to describe the semantic constraints that predicates place on their arguments. In this section, we will investigate how *nominal arguments* place semantic constraints or expectations on the predicates with which they occur. Such expectations can be thought of as typical events that involve the given object. For example, a noun like *apple* could be said to have preferences about its **inverse subject position**, that is, the verbs that can take it as a plausible subject. Examples might be verbs like *grow* or *fall*; for its **inverse object position**, *apple* probably prefers verbs like *eat*, *cut*, or *plant*. We will use the term **inverse selectional preference** to refer to preferences of nouns for their predicates, distinguishing them from **regular selectional preferences**.

It is clear that not all verbs will be equally likely to occur with a given noun–role pair. Still, inverse selectional preferences warrant a closer look: To what extent do inverse selectional preferences differ from regular ones? And are the tasks of predicting regular and inverse selectional preferences equally difficult? We start in Section 7.2 with an exploratory data analysis of inverse selectional preferences, which shows that inverse selectional preferences show semantically coherent patterns like regular selectional preferences, but that, in contrast to most verbs, nouns tend to occur with multiple semantic groups of verbs. In Sections 7.3–7.5, we test the EPP model on a pseudo-disambiguation task for inverse selectional preferences.

7.1 Related Work

In computational linguistics, some approaches to characterizing selectional preferences have used the symmetric nature of their models to characterize nouns in terms of the verbs that they use (Hindle 1990; Rooth et al. 1999). However, they do not explicitly compare the two types of preferences. Also, there are approaches using selectional preference information, in particular for word sense disambiguation and related tasks, that could be characterized as using regular along with inverse selectional preferences (Dligach and Palmer 2008; Erk and Padó 2008; Nastase 2008). By comparing selectional preference model performance on the tasks of predicting inverse and regular selectional preferences in Sections 7.3–7.5, we hope to contribute to an understanding of what can be achieved by using inverse preferences in word sense analysis tasks.

At the same time, inverse selectional preferences have been the object of fruitful research in both psycholinguistics and theoretical linguistics. In psycholinguistics, a particularly plausible argument for the existence of expectations of nouns for their predicates in human language processing is head-final word order (as in Japanese or in German subordinate clauses), where hearers may encounter all objects before the head. It is likely that these objects are immediately integrated into a preliminary event structure with an assumed predicate instead of being stored in short-term memory until the predicate is encountered (Konieczny and Döring 2003; Nakatani and Gibson 2009). Another strand of work is McRae et al. (2001, 2005), who have studied priming of verbs from nouns. They found that a noun engenders priming of verbs for which it is a typical agent, patient, instrument, or location.

In theoretical linguistics, the idea of event knowledge being encoded in the lexical entries of nouns has been formulated in the context of Pustejovsky's generative lexicon (Pustejovsky 1995), where the qualia roles TELIC and AGENTIVE provide information about the typical use of an object (*book: read*) and its construction (*book: write*), respectively. Pustejovsky uses this knowledge to account, for example, for the interpretation of logical metonymy (*begin a book*). Although qualia roles are instantiated with individual predicates rather than characterizations of all possible events, construction and use are arguably two very salient events for an object. Through the data exploration in Section 7.2, we hope to contribute to a linguistic characterization of inverse selectional preferences.

7.2 Empirical Analysis of Inverse Selectional Preferences

The first question we ask concerns the **selectional preference strength** of regular and inverse selectional preferences, using the measure introduced by Resnik (1996) to determine the degree to which verbs select for nouns, and vice versa. As verb–role pairs, we re-use the same 100 pairs that were used for the pseudo-disambiguation task in Experiment 1. For the comparison, we randomly sample a total of 100 noun/inverse-role pairs from the BNC, using the same five frequency bands as for the verbs (50–100, 100–200, 200–500, 500–1,000, >1,000). The sample contains approximately the same number of (inverse) subject and object roles.

We adapt the selectional preference strength measure from Equation (1) to our case: Unlike Resnik, we compute KL divergence not on a distribution across WordNet synsets, but on a distribution across lemmas.

$$\mathsf{SelStr}(w_1, r) = D(P(w_2|w_1, r)||P(w_2|r)) \tag{7}$$

For regular selectional preferences, w_1 is a verb lemma, w_2 a noun lemma, and r a role. For inverse preferences, w_1 is a noun lemma, w_2 a verb lemma, and r an inverse role. $\mathsf{SelStr}(w_1, r)$ can be interpreted as a measure of the degree to which w_1 has selectional preferences concerning the role r. We induce the probability distributions through maximum likelihood estimation on the BNC.

We can expect to see the same overall tendency in regular and inverse selectional preference strength. It is not possible that inverse selectional preference strength would be uniform throughout if regular selectional preference strength varied between verbs. After all, if we fix the relation r for the time being, $P(v|n)$ and $P(n|v)$ are related through Bayes' formula. Instead, the questions we will ask are more specific. Are regular and inverse preference strengths similar in size? Are regular and inverse preference strengths similar by frequency band—that is, do frequent nouns behave similarly to frequent verbs? And what effects do we see of the prior distributions $P(n|r)$ and $P(v|r)$?

Table 11 shows the range of selectional preference strengths found in each frequency band for verbs and nouns. As expected, we see substantial strengths in both regular and inverse preferences. Both parts of speech show the same pattern of decreasing KL divergences for higher-frequency words, presumably because frequent words tend to be polysemous, and can combine with many different words. However, the strengths for inverse selectional preferences are in general lower than those for regular preferences.

One possible reason for this is that the number of nouns seen with a verb–role pair might differ, in general, from the number of verbs seen with each noun–role pair. However, we find that verbs and nouns occur with roughly the same number of associates in the frequency bands up to the 200–500 band. In the band 500–1,000, verbs appear with roughly one third more nouns than nouns appear with verbs, and in the band of 1,000 occurrences or more, verbs appear with twice as many nouns (on average)

Table 11
Minimal, median, and maximal selectional preference strength (measured in terms of KL divergence) in a sample of 100 verbs and 100 nouns (20 lemmas each per frequency band).

Band	Verbs			Nouns		
	min	median	max	min	median	max
50–100	4.5	7.4	8.8	3.7	4.8	6.3
100–200	3.9	5.8	7.6	2.7	3.8	5.0
200–500	3.3	5.2	6.9	2.4	3.3	4.7
500–1,000	2.4	4.4	5.9	1.9	2.9	4.1
1,000–	1.8	3.6	6.2	1.4	2.3	3.7

as nouns appear with verbs in this band (1,189 vs. 636). Incidentally, the fact that the highest-frequency verbs (which also tend to be the most ambiguous) appear in a much larger number of contexts than the highest-frequency nouns could be a contributing factor to the well-known problem that verbs are harder to disambiguate than nouns. For the lower frequency bands, number of associates is unlikely to be the reason for the weaker inverse preferences. Instead, a more likely reason for the overall weaker inverse preferences lies in the overall distributions of nouns and verbs in the BNC. Both show a Zipfian distribution, but there are 15,570 verbs as opposed to 455,173 nouns. Recall that KL divergence will be high whenever the individual terms $\frac{p(w_2|w_1,r)}{p(w_2|r)}$ to be summed are large. This, in turn, is the case when $p(w_2|r)$ is small. And $p(w_2|r)$ may be small when the distribution $p(w_2|r)$ ranges over a larger number of words w_2. For regular selectional preferences, the w_2 are nouns, and for inverse preferences the w_2 are verbs. Because there are many more nouns than verbs, the denominator $p(w_2|r)$ tends to be smaller for regular preferences.

To get a clearer understanding of how inverse selectional preferences compare to regular selectional preferences, we next do a qualitative analysis, looking at association strength SelAssoc for individual triples verb–role–noun and noun–inverse-role–verb. We adapt Equation (2) to the lexicon-free case and obtain

$$\mathsf{SelAssoc}(w_1, r, w_2) = \frac{1}{\mathsf{SelStr}(w_1, r)} P(w_2|w_1, r) \log \frac{P(w_2|w_1, r)}{P(w_2|r)} \tag{8}$$

Table 12 shows the five strongest associates for one verb–role pair and one noun–role pair from each frequency band. The associates on both sides of the table generally are semantically coherent and make intuitive sense. However, there is an interesting difference between the verbs and nouns: We find that the nouns' preferred verbs can often be grouped loosely into several meaning clusters, whereas the verbs' associates tend to group into one cluster per grammatical function. For example, predicates taking *wheat* as objects fall into those describing production (*grow, sow*) and those describing processing (*shred, grind, mill*). Similarly, the predicates found for *pill* either concern ingestion (*take, swallow, pop*), prescription, or idiomatic usage. In contrast, the objects of *rebut* describe different kinds of statements, and the objects of *celebrate* are anniversaries and other special events. Another observation that we can make in Table 12 is that the nouns' most preferred associates have a similarly large share in the nouns' overall selectional preference strength as the verbs' most preferred associates have in the verbs' selectional preference strength. This indicates that the distribution of selectional preferences is similarly skewed towards the most preferred associate for verbs and nouns.

Table 12
Examples of regular and inverse selectional preferences from different frequency bands for argument positions of nouns and verbs: overall selectional preference strength SelStr and most highly associated fillers with association strengths SelAssoc.

Band	Verbs		Nouns	
50–100	*rebut*–obj, SelStr(w)= 7.43		*wreckage*–obj^{-1}, SelStr(w)= 5.91	
	presumption	0.283	survey	0.126
	allegation	0.088	examine	0.089
	charge	0.082	sift	0.075
	criticism	0.049	clear	0.056
	claim	0.041	sight	0.051
100–200	*enunciate*–obj, SelStr(w)= 6.89		*wheat*–obj^{-1}, SelStr(w)= 5.00	
	principle	0.242	grow	0.184
	word	0.085	shred	0.049
	theory	0.034	grind	0.049
	philosophy	0.034	mill	0.042
	policy	0.029	sow	0.040
200–500	*break_with*–obj, SelStr(w)= 6.92		*pill*–obj^{-1}, SelStr(w)= 4.15	
	tradition	0.237	take	0.290
	past	0.054	swallow	0.165
	precedent	0.035	sweeten	0.070
	convention	0.035	prescribe	0.049
	Rome	0.022	pop	0.028
500–1,000	*commence*–obj, SelStr(w)= 5.92		*dividend*–obj^{-1}, SelStr(w)= 4.10	
	proceedings	0.185	pay	0.508
	action	0.051	declare	0.064
	work	0.043	receive	0.064
	proceeding	0.041	recommend	0.054
	operation	0.033	raise	0.023
1,000–	*celebrate*–obj, SelStr(w)= 6.23		*requirement*–obj^{-1}, SelStr(w)= 3.24	
	anniversary	0.177	meet	0.332
	birthday	0.170	satisfy	0.015
	centenary	0.046	comply_with	0.093
	victory	0.033	fulfill	0.061
	mass	0.028	impose	0.028

In sum, we find that inverse selectional preferences have weaker overall selectional preference strength than regular preferences, but that may be due more to specifics of the formula used rather than the skewness towards preferred role fillers. Two differences do emerge, though. First, noun selectional preferences show more semantic filler sets than verb preferences. Second, the highest frequency verbs appear with many more different associates than the highest frequency nouns.

7.3 Modeling Inverse Selectional Preferences

In the rest of this section, we test selectional preference models on the task of predicting inverse selectional preferences in a pseudo-disambiguation task, and compare the results to the performance on predicting regular preferences in Experiment 1. We do not repeat Experiment 2 even though it would have been technically possible to re-use the McRae and Pado data sets and predict plausibility judgments through inverse preferences. However, the data sets combine each verb with both plausible and implausible nouns, but they do not combine each noun with different verbs in a balanced fashion, so a repetition of Experiment 2 with inverse preferences would not be very informative.

For the pseudo-disambiguation experiment, we focus on the EPP model. Distributional models can, in general, be used straightforwardly to model both regular and inverse selectional preferences. This is different for models like RESNIK that use the WordNet noun hierarchy to represent regular selectional preferences. To model inverse preferences, it would be necessary to use the WordNet verb hierarchy. However, WordNet organizes verbs in a comparatively flat, unconnected hierarchy with a high branching factor formed by the hypernymy/troponymy ("type of") relation. This makes effective generalization difficult, in particular in conjunction with the marked variation in the set of preferred predicates that we observed for inverse selectional preferences in Section 7.2.

We adapt the formulation of the EPP model to the inverse selectional preference case as follows. Let a stand for a noun, r for an inverse argument position of this noun, and $\mathsf{Seenpreds}(r, a)$ for the set of predicates seen with noun a and role r. Then the selectional preference $\mathsf{Selpref}_{\mathsf{EPP}}$ of (r, a) for a verb v_0 is defined in parallel to Equation (6) as weighted average similarity to seen verbs:

$$\mathsf{Selpref}_{\mathsf{EPP}\,r,a}(v_0) = \sum_{v \in \mathsf{Seenpreds}(r,a)} \frac{wt_{r,a}(v)}{Z_{r,a}} \mathsf{sim}(v, v_0) \tag{9}$$

with $Z_{r,a} = \sum_{v \in \mathsf{Seenpreds}(r,a)} wt_{r,a}(v)$ as the normalization constant.

7.4 Pseudo-Disambiguation: Experimental Setup

We evaluate inverse selectional preferences on a pseudo-disambiguation task that is set up completely analogously to our experiments on regular preferences in Section 5: given a noun, an inverse argument position, one verb observed in this position, and a confounder verb, distinguish between the two verbs. We use the 100 nouns sampled across five frequency bands that we already used in Section 7.2. We experiment with both WORDSPACE and DEPSPACE models, but restrict our attention to DISCR weighting, which showed good results in Experiment 1.

In Section 5, we experimented on two different primary corpora, the BNC (SYN PRIMARY setting) and FrameNet (SEM PRIMARY setting). Subsequently, we will use the SYN PRIMARY setting again, but not the SEM PRIMARY setting. In the SEM PRIMARY setting, the roles are FrameNet frame elements (semantic roles). However, frame elements are specific to a single frame, for example, the frame element ROPE belongs to the frame ROPE_MANIPULATION.[15] It would thus be pointless to predict a verb frame given a noun and a frame element name, as the frame element already gives away the frame.

[15] It is possible for multiple frame elements to share a name, for example there are multiple frames with a frame element named THEME. However, conceptually, this is only a shared name, not a shared role across frames.

7.5 Pseudo-Disambiguation: Results and Discussion

Table 13 shows the results of testing the EPP model for inverse selectional preferences on pseudo-disambiguation. Coverage is very good for all model variants, similarly to Experiment 1. The error rates, as well, are close to those for the regular preferences in the SYN PRIMARY setting (cf. Table 4). The best model there (DEPSPACE, PCA, nGCM with DISCR weighting) achieved an error rate of 25.6%, and the best model for inverse preferences (WORDSPACE, Lin with DISCR weighting) reaches an error rate of 27.2% here. Lin shows the best error rates in all conditions, closely followed by nGCM (the difference is significant in WORDSPACE and the reduced DEPSPACE, but not significant in the unreduced DEPSPACE). The Hindle similarity measure again brings up the rear. In PCA-transformed spaces, the error rates are similar across all similarity measures except for Hindle, as in Experiment 1.

WORDSPACEs yield better results than DEPSPACEs here, in contrast to Experiment 1. The best WORDSPACE model (Lin without PCA) reaches significantly better error rates ($p \leq 0.01$) than the best DEPSPACE model (Lin with PCA). We think that the reason for this lies in the fact that for inverse selectional preferences, the true associate and the confounder that need to be distinguished in the pseudo-disambiguation task are verbs rather than nouns. A noun will probably have more other nouns in a bag-of-words context window than a verb would other verbs, which will make it easier to distinguish verbs in a WORDSPACE than to distinguish nouns. A DEPSPACE, in contrast, will bring out differences in the immediate syntactic neighborhood of nouns even if they occur in the same sentence.

8. Conclusion

In this article, we have presented a similarity-based model of selectional preferences, EPP. It computes the selectional fit of a candidate role filler as a weighted sum of semantic similarities to headwords observed in a corpus, in a straightforward implementation

Table 13
Pseudo-disambiguation results for inverse selectional preferences (BNC as primary and secondary corpus, DISCR weighting). ER = Error rate; Cov = Coverage.

Dimensions	Similarity	DEPSPACE		WORDSPACE	
		ER (%)	Cov (%)	ER (%)	Cov (%)
Original 2,000 dimensions	Cosine	37.4	99.0	34.0	99.1
	Dice	42.4	98.8	43.4	98.7
	nGCM	33.7	99.3	31.5	99.3
	Hindle	48.8	96.0	52.2	94.6
	Jaccard	36.7	99.4	44.9	98.7
	Lin	**32.8**	98.9	**27.2**	98.9
PCA 500 dimensions	Cosine	35.2	99.0	31.3	99.4
	Dice	35.0	99.6	32.9	99.8
	nGCM	32.4	99.2	30.3	99.6
	Hindle	44.2	99.0	48.7	99.1
	Jaccard	34.8	99.6	32.6	99.8
	Lin	**30.6**	99.8	**28.8**	99.8

of the intuition that plausibility judgments should generalize to fillers with similar meaning. Our model is simple and easy to compute. In common with other distributional models like Rooth et al. (1999), it does not depend on lexical resources. Our model derives additional flexibility from distinguishing between a primary corpus (for observing headwords) and a generalization corpus (for inducing semantic similarities). This allows it to use primary corpora with deeper semantic annotation that are too small as a basis for computing vector space representations.

We have evaluated the EPP model on two tasks, a pseudo-disambiguation task that can be viewed as an abstraction of both word sense disambiguation and semantic role labeling, as well as on the prediction of human plausibility judgments. The model achieves similar error rates to the semantic hierarchy–based RESNIK model, at considerably higher coverage, and it achieves lower error rates than the ROOTH ET AL. soft clustering model. The semantic role–based PADO ET AL. model, although highly accurate in its predictions, has much lower coverage and needs a semantically annotated corpus as a basis. We have also demonstrated that our model is able to meaningfully model inverse selectional preferences, that is, expectations of nouns about verbs for which they appear as arguments.

With respect to parameter settings of the EPP model, we find consistent patterns across the three tasks we have considered. nGCM, Lin, and Cosine are the best-performing similarity measures throughout. The good performance of the nGCM measure, an exponential similarity measure, is particularly noteworthy. We found it to work well on data sets that are sparse and not too noisy, whereas the Lin similarity measure achieved better performance when the data was noisy (see Section 5.4 for details). Dimensionality reduction (PCA) on the vector space raises the performance of the Jaccard and Dice similarity measures to a similar level as the best three. More importantly, PCA neutralizes a strong frequency bias that otherwise leads to a large performance drop on rare arguments. Concerning weighting schemes, we found that frequency-based weighting works well when the data is either clean or not too sparse. In the face of sparse noisy data, DISCR weighting (a variant of tf/idf) is helpful. Comparing bag-of-words–based and dependency-based vector spaces, DEPSPACEs are sparser but cleaner than WORDSPACEs. Accordingly, DEPSPACEs are at an advantage when many headwords are available, making efficient use of this information, whereas WORDSPACEs work better for predicates with few seen headwords because they are less affected by sparseness.

We conclude with two open questions. The first question concerns the appropriate representation of selectional preferences for polysemous verbs such as *address*, whose direct object can either be a person, or a problem. Polysemy leads to headwords with lower similarity among them than for non-polysemous verbs, which in turn can lead to artificially low plausibilities for all fillers. In the SEM PRIMARY setting, occurrences of polysemous verbs are separated into different frames. In future work, we hope to improve our SYN PRIMARY setting models by clustering the seen headwords, and then computing plausibility of new headwords relative to the nearest cluster.

A second question is the usefulness of inverse selectional preferences for the acquisition of fine-grained information about nouns. As we discussed in Section 7, the preferred verbs for a noun can often be grouped into meaning clusters. In future work, we plan to investigate whether there are groups of predicates that recur across similar nouns, and how they can be characterized. We expect some groups to correspond to Pustejovsky's qualia (Pustejovsky 1995), which constitute particularly salient events for an object, namely, their creation and typical use. However, we expect corpus data to yield a more complex picture of the events connected to a noun, which manifest themselves in the form of additional, more specific meaning clusters.

Acknowledgments

We would like to thank Detlef Prescher and Carsten Brockmann for their implementations of the ROOTH ET AL. and RESNIK models, respectively. We are also grateful to Nate Chambers and Yves Peirsman, as well as to the anonymous reviewers, for their comments and suggestions.

References

Abe, Naoki and Hang Li. 1996. Learning word association norms using tree cut pair models. In *Proceedings of the 10th International Conference on Machine Learning*, pages 3–11, Bari.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–222.

Bergsma, Shane, Dekang Lin, and Randy Goebel. 2008. Discriminative learning of selectional preference from unlabeled text. In *Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing*, pages 59–68, Honolulu, HI.

Best, John and D. E. Roberts. 1975. Algorithm AS 89: The upper tail probabilities of Spearman's Rho. *Applied Statistics*, 24:377–379.

Briscoe, Ted and Bran Boguraev, editors. 1989. *Computational Lexicography for Natural Language Processing*. Longman Publishing Group, New York.

Brockmann, Carsten and Mirella Lapata. 2003. Evaluating and combining approaches to selectional preference acquisition. In *Proceedings of the 16th Meeting of the European Chapter of the Association for Computational Linguistics*, pages 27–34, Budapest.

Burnard, Lou, 1995. *User's Guide for the British National Corpus*. British National Corpus Consortium, Oxford University Computing Services.

Ciaramita, Massimiliano and Mark Johnson. 2000. Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 187–193, Saarbrücken.

Clark, Stephen and David Weir. 2001. Class-based probability estimation using a semantic hierarchy. In *Proceedings of the 2nd Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 95–102, Pittsburgh, PA.

Collins, Michael. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 16–23, Madrid.

Curran, James. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, University of Edinburgh.

Daelemans, Walter and Antal van den Bosch. 2005. *Memory-Based Language Processing*. Cambridge University Press, Cambridge, UK.

Dagan, Ido, Lillian Lee, and Fernando C. N. Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34(1):34–69.

Dligach, Dmitriy and Martha Palmer. 2008. Novel semantic features for verb sense disambiguation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics:Human Language Technologies, Short Papers*, pages 29–32, Columbus, OH.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Efron, Bradley and Robert Tibshirani. 1994. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. Chapman & Hall, London.

Erk, Katrin. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 216–223, Prague.

Erk, Katrin and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing*, pages 897–906, Honolulu, HI.

Fillmore, Charles J., Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16:235–250.

Firth, John Rupert. 1957. A synopsis of linguistic theory, 1930–1955. In Philological Society, editor, *Studies in Linguistic Analysis*. Blackwell, Oxford, pages 1–32.

Garnsey, Susan, Neal Pearlmutter, Elizabeth Myers, and Melanie Lotocky. 1997. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 37:58–93.

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

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Dordrecht.

Grishman, Ralph and John Sterling. 1992. Acquisition of selectional patterns. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 658–664, Nantes.

Harris, Zellig. 1968. *Mathematical Structure of Language*. Wiley, New York.

Hay, Jennifer, Aaron Nolan, and Katie Drager. 2006. From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review*, 23(3):351–379.

Hindle, Donald. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh, PA.

Hindle, Donald and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

Katz, Jerrold J. and Jerry A. Fodor. 1963. The structure of a semantic theory. *Language*, 39(2):170–210.

Katz, Jerrold J. and Paul M. Postal. 1964. *An Integrated Theory of Linguistic Descriptions*. Research Monograph No. 26. MIT Press, Cambridge, MA.

Konieczny, Lars and Philipp Döring. 2003. Anticipation of clause-final heads. Evidence from eye-tracking and SRNs. In *Proceedings of the 4th International Conference on Cognitive Science*, pages 330–335, Sydney.

Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago, IL.

Landauer, Thomas and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Lee, Lillian. 1999. Measures of distributional similarity. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 25–32, College Park, MA.

Lin, Dekang. 1993. Principle-based parsing without overgeneration. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 112–120, Columbus, OH.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the Joint Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, pages 768–774, Montreal.

Lowe, Will. 2001. Towards a theory of semantic space. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 576–581, Edinburgh.

Lowe, Will and Scott McDonald. 2000. The direct route: Mediated priming in semantic space. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 675–680, Philadelphia, PA.

Lund, Kevin and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, 28:203–208.

McCarthy, Diana. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 256–263, Seattle, WA.

McCarthy, Diana and John Carroll. 2003. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, pages 279–286, Barcelona.

McCarthy, Diana, Sriram Venkatapathy, and Aravind K. Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *Proceedings of the 12th Joint Conference on Empirical Methods in Natural Language Processing and Conference on Natural Language Learning*, pages 369–379, Prague.

McDonald, Scott and Chris Brew. 2004. A distributional model of semantic context effects in lexical processing. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, pages 17–24, Barcelona.

McRae, Ken, Todd Ferretti, and Liane Amyote. 1997. Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, 12(2/3):137–176.

McRae, Ken, Mary Hare, Jeffrey Elman, and Todd Ferretti. 2005. A basis for generating expectancies for verbs from nouns. *Memory and Cognition*, 33(7):1174–1184.

McRae, Ken, Mary Hare, Todd Ferretti, and Jeffrey Elman. 2001. Activating verbs typical agents, patients, instruments, and locations via event schemas. In *Proceedings* *of the Twenty-Third Annual Conference of the Cognitive Science Society*, pages 617–622, Mahwah, NJ.

McRae, Ken, Michael Spivey-Knowlton, and Michael Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38:283–312.

Miller, George, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Five papers on WordNet. *International Journal of Lexicography*, 3(4):235–312.

Nakatani, Kentaro and Edward Gibson. 2009. An on-line study of Japanese nesting complexity. *Cognitive Science*, 1(34):94–112.

Nastase, Vivi. 2008. Unsupervised all-words word sense disambiguation with grammatical dependencies. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pages 757–762, Honolulu, HI.

Nosofsky, Robert. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39–57.

Padó, Sebastian and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Padó, Sebastian, Ulrike Padó, and Katrin Erk. 2007. Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of the 12th Joint Conference on Empirical Methods in Natural Language Processing and Conference on Natural Language Learning*, pages 400–409, Prague.

Padó, Ulrike. 2007. *The Integration of Syntax and Semantic Plausibility in a Wide-Coverage Model of Human Sentence Processing*. Ph.D. thesis, Saarland University, Saarbrücken, Germany.

Padó, Ulrike, Matthew W. Crocker, and Frank Keller. 2009. A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*, 33:794–838.

Pantel, Patrick, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning inferential selectional preferences. In *Proceedings of the Joint Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 564–571, Rochester, NY.

Pereira, Fernando, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Columbus, OH.

Pustejovsky, James. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

Resnik, Philip. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61:127–159.

Rooth, Mats, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, College Park, MA.

Salton, Gerard, Anita Wong, and Chung-Shu Yang. 1975. A vector-space model for information retrieval. *Journal of the American Society for Information Science*, 18:613–620.

Schulte im Walde, Sabine. 2010. Comparing computational approaches to selectional preferences: Second-order co-occurrence vs. latent semantic clusters. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 1381–1388, Valleta.

Schulte im Walde, Sabine, Christian Hying, Christian Scheible, and Helmut Schmid. 2008. Combining EM training and the MDL principle for an automatic verb classification incorporating selectional preferences. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 496–504, Columbus, OH.

Shepard, Roger. 1987. Towards a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323.

Stolcke, Andreas. 2002. SRILM—an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904, Denver, CO.

Toutanova, Kristina, Christoper D. Manning, Dan Flickinger, and Stephan Oepen. 2005. Stochastic HPSG parse selection using the Redwoods corpus. *Journal of Research on Language and Computation*, 3(1):83–105.

Trueswell, John, Michael Tanenhaus, and Susan Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33:285–318.

Vandekerckhove, Bram, Dominiek Sandra, and Walter Daelemans. 2009. A robust and extensible exemplar-based model of thematic fit. In *Proceedings of the 12th Meeting of the European Chapter of the Association for Computational Linguistics*, pages 826–834, Athens.

Wilks, Yorick. 1975. Preference semantics. In E. Keenan, editor, *Formal Semantics of Natural Language*. Cambridge University Press, Cambridge, UK, pages 329–350.

Yarowsky, David. 1993. One sense per collocation. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 266–271, Princeton, NJ.

Zanzotto, Fabio Massimo, Marco Pennacchiotti, and Maria Teresa Pazienza. 2006. Discovering asymmetric entailment relations between verbs using selectional preferences. In *Proceedings of the Joint Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, pages 849–856, Sydney.

Zapirain, Beñat, Eneko Agirre, and Lluís Màrquez. 2009. Generalizing over lexical features: Selectional preferences for semantic role classification. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, pages 73–76, Singapore.