A Probabilistic Model of Semantic Plausibility in Sentence Processing

A Probabilistic Model of Semantic Plausibility in Sentence Processing

Ulrike Padó and Matthew W. Crocker
Computational Linguistics, Saarland University

Frank Keller
School of Informatics, University of Edinburgh

Ulrike Padó: ulrike@coli.uni-sb.de, tel: +1-650-704-2710, fax: +1-650-723-5666,
Matthew W. Crocker: crocker@coli.uni-sb.de,
Frank Keller: keller@inf.ed.ac.uk

Abstract

Experimental research shows that human sentence processing uses information from different levels of linguistic analysis, for example lexical and syntactic preferences as well as semantic plausibility. Existing computational models of human sentence processing, however, have focused primarily on lexico-syntactic factors. Those models that do account for semantic plausibility effects lack a general model of human plausibility intuitions at the sentence level. Within a probabilistic framework, we propose a wide-coverage model that both assigns thematic roles to verb-argument pairs and determines a preferred interpretation by evaluating the plausibility of the resulting $(verb, role, argument)$ triples. The model is trained on a corpus of role-annotated language data. We also present a transparent integration of the semantic model with an incremental probabilistic parser. We demonstrate that both the semantic plausibility model and the combined syntax/semantics model predict judgment and reading time data from the experimental literature.

1. Introduction

Human language processing draws upon a range of information sources, as demonstrated by experimental results which reveal the rapid influence of lexical, structural, and

We would like to thank Brian Roark and Zvika Marx for kindly making their software available to us, and we thank three anonymous reviewers for their helpful comments. The research reported in this paper was carried out while the first author received a studentship within Deutsche Forschungsgemeinschaft IRTG 715 "Language Technology and Cognitive Systems". Work on this article was completed while she was a visiting scholar in the NLP group at Stanford University.

semantic factors on ambiguity resolution (e.g., Trueswell, Tanenhaus, & Kello, 1993; Stowe, 1989; Taraban & McClelland, 1988; MacDonald, 1994; Garnsey, Pearlmutter, Myers, & Lotocky, 1997). The sentence processing mechanism is capable of seamlessly integrating these diverse information sources, while remaining extremely fast, accurate and robust towards incorrect and noisy input.

Implemented computational models offer an opportunity to investigate the mechanisms underlying the processor's ability to integrate information from a variety of sources. Such models demand the precise specification of the hypotheses they implement and they can generate testable predictions. However, most existing sentence processing models have focused on lexico-syntactic factors only. Even models that do account for effects of semantic plausibility lack a general prediction mechanism for human plausibility intuitions on the sentence level. Furthermore, the human parser's *wide coverage*, i.e., its ability to handle a wide range of linguistic phenomena, and to cope with previously unseen material, remains a challenge for many models that are designed to cover only a small number of specific phenomena.

In this paper, we propose the SynSem-Integration model, which combines an incremental probabilistic parsing model with a new computational account of semantic plausibility. Semantic plausibility is a complex and multifaceted notion, which our model approximates as the thematic fit between a verb and its arguments, given the sense of the verb. The model implements a probabilistic notion of thematic fit and learns the relevant information from corpus data. The SynSem-Integration model is wide coverage, i.e., it is able to process material it has not encountered in the training corpus, and it is general enough to handle arbitrary linguistic phenomena, at least in principle.

The spectrum of existing computational models proposed to account for human sentence processing is large. There are models based on a small set of fixed parsing rules or principles (e.g., Frazier, 1978; Abney, 1989; Crocker, 1996), models focusing on memory constraints and other cognitive constraints (e.g., Gibson, 1991; Lewis & Vasishth, 2005), connectionist models (e.g., Rohde, 2002; Mayberry, 2003) and hybrid symbolic/connectionist accounts (e.g., Stevenson, 1994; Vosse & Kempen, 2000).

However, all of these models only provide restricted accounts of a property of the human sentence processor that is key to explaining its robustness and wide coverage (e.g., Jurafsky, 2003; Crocker, 2005; Chater & Manning, 2006), viz., the pervasiveness of frequency effects on different levels of processing. There is evidence for the effect of lexical category frequencies (e.g., Trueswell, 1996; Crocker & Corley, 2002), verb subcategorization frame frequencies (e.g., Trueswell et al., 1993; Garnsey et al., 1997), and structural frequencies (e.g., Cuetos, Mitchell, & Corley, 1996). Fully connectionist or hybrid connectionist/symbolic models such as the ones referenced above could in principle account for such frequency effects, and display considerable robustness to noisy input. However, these models require large amounts of training data, and many training iterations, which makes it difficult to scale them up to a realistically wide coverage of linguistic phenomena.

This problem is addressed by two classes of computational models that are explicitly probabilistic and use structural frequencies estimated from corpora: *probabilistic grammar-based models* and *constraint-based models*. Probabilistic grammar-based models have evolved from Jurafsky's (1996) proposal and subsequent work by Crocker and Brants (2000). This approach uses a probabilistic context-free grammar to encode information

about lexical and structural preferences. The model incrementally assigns each analysis a probability on the basis of the grammar rules applied, where rule probabilities are estimated from a training corpus. The human parser is assumed to entertain all possible analyses whose probability exceeds a certain threshold. Processing difficulty arises when an analysis that was previously dispreferred turns out to be correct based on subsequent input. Probabilistic grammar-based models thus account both for the generation of alternative analyses in case of an ambiguity and for processing difficulty that can arise from resolving such ambiguities. Their robustness and wide coverage stems from the fact that they use large, probabilistic grammars induced from a treebank, a syntactically annotated training corpus.

A variant of this type of approach is the *surprisal model* proposed by Hale (2001) (see also Levy, 2008). This model predicts processing difficulty by monitoring incremental changes in the probability distribution over all possible analyses of the input. It predicts increased processing load at the point where analyses with a large probability mass are disconfirmed (which indicates the integration of a word with high surprisal or information value). Surprisal-based models assume a wide-coverage grammar, which allows them to account for the human sentence processing system's robustness. They are also capable of predicting processing difficulty for non-ambiguous phenomena such as relative clause embedding. However, since they do not aim at directly predicting parsing preferences and ambiguity resolution processes, we will focus on the first type of probabilistic grammar-based models here.

A common shortcoming of all probabilistic-grammar based models is that they do not naturally integrate factors beyond the lexico-syntactic information encoded in a probabilistic context-free grammar. Specifically, they cannot account for semantic plausibility, as they have at best a syntactic representation of the relationship between a verb and its argument, and would require vast amounts of training data to reach sufficient coverage of such information to reliably predict plausibility effects.

The second class of explicitly probabilistic models includes *constraint integration models*. Accounts like that of Spivey-Knowlton (1996) or Narayanan and Jurafsky (2002) explicitly focus on the integration of a wide range of probabilistic constraints on linguistic processing. They select the preferred analysis from a pool of pre-specified possible structures for an ambiguous input, using competition for activation (in the case of Spivey-Knowlton, 1996), or Bayesian reasoning (in the case of Narayanan & Jurafsky, 2002). Difficulty is predicted in the same way as for probabilistic grammar-based models by Narayanan and Jurafsky's (2002) approach, while competition-based models link processing difficulty to the time the system takes to settle on a preferred analysis (Spivey-Knowlton, 1996); it converges quickly if all constraints prefer the same analysis and slowly if there is conflicting evidence.

Constraint integration models are well suited to model the influence of semantic plausibility, which they can achieve by simply introducing additional constraints. A disadvantage of these models, however, is that they have no theoretically motivated way of determining the values of such constraints; they are typically instantiated from semantic plausibility judgments. Another disadvantage of constraint-integration models is that they require constraints to be specified by hand and separately for every phenomenon; it is therefore difficult to achieve a wide coverage of phenomena, and to deal with unseen input. Furthermore, by looking only at a small number of pre-specified alternatives, these models leave

aside the non-trivial question of how syntactic analyses are constructed in the first place. They also assume an unrealistically low level of ambiguity: probabilistic grammar models demonstrate that even seemingly unambiguous sentences or sentence fragments can have hundreds or thousands of analyses, while constraint-based models typically only deal with two or three pre-selected alternatives for ambiguous fragments.

In our discussion of both types of explicitly probabilistic models, it has become clear that one basic difficulty for computational models of sentence processing models lies in accounting for human semantic plausibility intuitions. Existing models are forced to either consider lexico-syntactic factors only, or to use costly-to-obtain human judgments to capture the influence of plausibility on processing. While the latter solution allows the representation of plausibility constraints, it does not actually model the factors that underlie them.

To address this problem, and the ensuing shortcomings of existing probabilistic models, this paper proposes:

- a probabilistic model of human plausibility intuitions that approximates plausibility as the thematic fit between a verb and its arguments and is trained on verb-argument-role triples extracted from semantic-role-annotated corpora;
- the SynSem-Integration model, an architecture that integrates the plausibility model with a probabilistic grammar-based model to capture the construction of syntactic structures and the resolution of ambiguities using lexical, syntactic and semantic information, while being able to handle a wide range of linguistic phenomena, and to cope with previously unseen material.

In the following, we will discuss these two proposals in turn. We first introduce and evaluate the semantic plausibility model. We then go on to describe the architecture of the SynSem-Integration model and evaluate its predictions against empirical findings.

2. A model of semantic plausibility

Our first contribution is a general model of human intuitions about the plausibility of events. We represent aspects of events as a verb and argument in a specific relation, breaking down an event like *The pirate terrorizes the Seven Seas* into *pirate is the agent in a terrorizing event* and *Seven Seas is the patient in a terrorizing event*. We describe the semantic relation between a verb and its argument by the thematic role which the verb assigns to the argument. This representation follows both the neo-Davidsonian approach to event description in semantics (e.g., Parsons, 1990; Carlson, 1984) and the status of thematic roles in psycholinguistics as a pivotal link between syntactic and semantic processing, for example as a type of low-cost, preliminary semantic analysis (Carlson & Tanenhaus, 1988). The verb-argument-role representation of sentence semantics encodes basic information about the events referred to in a sentence, while avoiding complex issues like quantifier scope and verb tense and aspect.

In experimental psycholinguistics, plausibility is typically manipulated using thematic fit, which can be achieved by varying the argument of a verb-argument-relation triple. Such a plausibility manipulation on the thematic fit level was carried out for example in McRae, Spivey-Knowlton, and Tanenhaus (1998). Their study investigated the influence of thematic fit information on the processing of the main clause/reduced relative (MC/RR) ambiguity in sentences like

(1) a. The pirate terrorized by his captors was freed quickly.
b. The victim terrorized by his captors was freed quickly.

During incremental processing of sentences like (1-a), the prefix *The pirate terrorized* is ambiguous between the more frequent main clause continuation (e.g., as *The pirate terrorized the Seven Seas*) and a less frequent reduced relative continuation as shown in (1-a), where *terrorized* heads a relative clause that modifies *pirate*. The subsequent *by*-phrase provides strong evidence towards the reduced relative reading and the main verb region *was freed* completely disambiguates.

Evidence from experimental work shows that readers initially prefer the main clause interpretation over the reduced relative, but that this preference can be modulated by other factors (e.g., Rayner, Carlson, & Frazier, 1983; Trueswell, 1996; Crain & Steedman, 1985). McRae et al. showed that good thematic fit of the first NP as an object of the verb in the case of *victim* in (1-b) allowed readers to partially overcome the main clause preference and more easily adopt the dispreferred reduced relative interpretation, which makes the first NP the object of the verb (as opposed to the main clause reading, where it is a subject). Reading time effects both on the ambiguous verb and in the disambiguating region suggest that the thematic fit of the first NP and the verb rapidly influences the human sentence processor's preference for the two candidate structures.

To account for the thematic fit information in items like sentences (1-a) and (1-b) above, a model has to solve two tasks: It has to identify the semantic relation that holds between pairs of verb and argument like *terrorize-pirate*. These pairs can be extracted from a syntactic analysis of the input fragment *The pirate terrorized ...* Given the pair *terrorize-pirate* (and the corresponding grammatical function), a model should predict, for example, the *agent* role, and not the *experiencer* or the *means* roles.[1] However, identifying the role intended by the speaker does not necessarily allow conclusions about the real-world plausibility of the verb-argument-role triple (cf. the syntactically straightforward, but semantically implausible assignments for *The victim terrorized the pirate*). The model therefore also needs to predict the plausibility of the event described by the verb-argument-role triple. In the case of *terrorize-pirate-agent*, this plausibility estimate should be high, whereas it should be lower for *terrorize-victim-agent*.

The first task is similar to that of a semantic role labeling model in computational linguistics. There has been considerable interest in this topic starting with work by Gildea and Jurafsky (2002). Influential work by Surdeanu, Harabagiu, Williams, and Aarseth (2003) and Xue and Palmer (2004) has established useful features and modeling procedures, and a wide range of models has been proposed due to the adoption of semantic role labeling as a shared task in the Senseval-III competition (Litkowski, 2004) and at the CoNLL-2004 and 2005 conferences (Carreras & Márquez, 2005). We propose our own model here, however, because semantic role labeling models do not explicitly address the second modeling task, the prediction of human plausibility ratings. We have explored the possibility of using a role labeling model for plausibility prediction, but have found that it did not succeed because the standard labeling features rely heavily on syntactic information to assign labels and lack the semantic information that is crucial here (Padó, Crocker, & Keller, 2006). The model we propose here is specifically designed to assign both roles and plausibility predictions.

[1] Roles are given as defined by FrameNet 1.2 for the Cause_to_experience frame.

In parallel to probabilistic parser models for syntax, we choose a probabilistic model formulation based on frequency information for linguistic utterances. Instead of using corpora with purely syntactic annotation, as for syntax models, we rely on corpora that are (additionally) annotated with thematic information, such as FrameNet (Baker, Fillmore, & Lowe, 1998) or PropBank (Palmer, Gildea, & Kingsbury, 2005). FrameNet annotates a subset of the British National Corpus with Frame Semantics (Fillmore, 1982). PropBank adds a layer of thematic role annotation to the Wall Street Journal section of the Penn Treebank (Marcus, Santorini, & Marcinkiewicz, 1994). We use the FrameNet (release 1.2) corpus to derive the probabilistic model, since it has been shown to offer a better basis for modeling plausibility data than PropBank (Padó et al., 2006). The fundamental assumption of the probabilistic approach is that the plausibility of real-world events can be modeled using the frequency of the events' descriptions in linguistic utterances. We discuss this issue further in the General Discussion below.

The probabilistic formulation of the semantic model equates the plausibility of a verb-argument-role triple with the probability of seeing the thematic role with the verb-argument pair in a large corpus of annotated language data. This is parallel to the syntactic modeling practice of equating the preferredness of a structure with the probability of encountering it in an annotated corpus. The semantic model estimates the plausibility of a verb-role-argument triple as the joint probability of five variables: These are the identity of the verb v, argument a and thematic role r, the verb's sense s and the grammatical function gf of the argument. The verb's sense is relevant because it determines the set of applicable thematic roles, while the grammatical function linking verb and argument (e.g., *syntactic subject* or *syntactic object*) carries information about the thematic role intended by the speaker. The semantic model equation is given in Equation 1.

$$Plausibility_{v,r,a} = P(v, s, gf, r, a) \tag{1}$$

The joint probability formulation makes the model an instance of a *generative model*. This type of model attempts to estimate the joint probability distribution that is most likely to generate the observed co-occurrence of the input variables (here, the verb and argument as well as the verb sense and grammatical function) and the output variable (the thematic role). On the basis of the estimated distribution, generative models can predict the most likely instantiation for missing input or output values. This property allows the model to naturally solve its dual task of identifying the correct role that links a given verb and argument, and making a plausibility prediction for the triple: It predicts the preferred thematic role for a verb-argument pair by generating the most probable instantiation for the role, as shown in Equation 2.

$$\hat{r}_{v,a} = \underset{r}{argmax} \;\; P(v, s, gf, r, a) \tag{2}$$

If necessary, the verb sense and grammatical function can also be generated. The probability assigned to the resulting combination of variable instantiations is the model's plausibility prediction for the verb-argument pair and chosen role. If all variables are known, the generation and maximization steps are unnecessary and the plausibility prediction is made directly.

An equivalent, decomposed version of Equation 1 (derived using the chain rule) allows a more intuitive understanding of the linguistically relevant information about the verb-argument pair used by the model.

$$\begin{aligned} \mathit{Plausibility}_{v,r,a} &= P(v,s,gf,r,a) \\ &= P(v)\cdot P(s|v)\cdot P(gf|v,s)\cdot P(r|v,s,gf)\cdot P(a|v,s,gf,r) \end{aligned} \tag{3}$$

The decomposed formulation contains $P(s|v)$, which denotes the sense distribution of a polysemous verb. The $P(gf|v,s)$ term captures information about the verb's syntactic subcategorization preferences when used in sense s: It reflects the probability of the verb's co-occurrence with dependents in any of the possible grammatical functions. The $P(r|v,s,gf)$ term shows how the verb prefers to realize its thematic role fillers syntactically. Finally, the $P(a|v,s,gf,r)$ term is similar to the term estimated by selectional preference models in cognitive science and computational linguistics (Resnik, 1996; Clark & Weir, 2002) which determine a verb's preference for certain argument types and estimate the fit of a verb and argument in a given role.[2]

Given the above model of plausibility for individual arguments, we now define the computation of the plausibility of a sentence or sentence-initial fragment with several arguments. We determine the plausibility of a completed or incremental syntactic analysis by multiplying the plausibility estimates for all verb-argument pairs it contains. This constitutes an independence assumption that ignores the existing dependencies between thematic roles assigned to different arguments of the same verb. It is however necessary because data sparseness in the training data makes it impossible to model these dependencies explicitly. We augment our approach to mitigate two problems:

- To approximate the dependencies betweens arguments of the same verb, we posit the constraint that each role can be assigned only once by the same verb and determine the optimal set of role predictions given this constraint. Equation 4 demonstrates the case of a verb with two arguments, using the joint model formulation for the sake of brevity. The role assignments by different verbs in the same sentence or fragment are treated as independent.

$$\mathit{Plausibility}_s = P(v,s,gf_1,\hat{r}_1,a_1)\cdot P(v,s,gf_2,\hat{r}_2,a_2) \tag{4}$$

where

$$(\hat{r}_1,\hat{r}_2) = \underset{\{r_1,r_2|r_1\neq r_2\}}{argmax}\; P(v,s,gf_1,r_1,a_1)\cdot P(v,s,gf_2,r_2,a_2) \tag{5}$$

This approach allows the assignment of semantically dispreferred roles where a more plausible role filler is available for the same verb. Note that Equation 4 indicates that the computation of plausibility requires the joint maximization of $\hat{r}_1$ and $\hat{r}_2$. However, this is a tractable problem, as the number of roles to consider is small and finite (and so is the number of verb senses s and grammatical functions gf, should these be unknown). Hence a complete search of the problem space is possible to perform the maximization.

- Computing the overall probability of multiple role assignments as the product of the individual probabilities causes a preference for analyses with small sets of role assignments

[2] Evaluation against selectional preference models on the plausibility prediction task shows that our model outperforms the existing approaches (Padó, 2007).

per verb. This leads to unexpected semantic rankings when we compare the semantic plausibilities of various syntactic analyses. We improve the predictions by using the geometric mean over the role assignment probabilities for each role set (mitigating the influence of the number of roles). We also weight the role sets by how well they correspond to the verbs' preferred role assignment patterns in the training data (see Padó (2007) for details).

2.1. Model estimation

The semantic model can be estimated from any language corpus with semantic role annotation. Two corpora with such markup are currently available: FrameNet (Fillmore, Johnson, & Petruck, 2003) and PropBank (Palmer et al., 2005). PropBank is the larger of the two, but our experiments indicate that the syntax-oriented quality of the PropBank semantic annotation allows less semantic generalization than the FrameNet role labels and is less suited to our task (Padó et al., 2006). We therefore estimate the semantic model from the FrameNet corpus.

The FrameNet annotation project groups verbs with similar meanings together into *frames* (i.e., descriptions of prototypical situations). Each frame introduces a set of frame-specific roles for typical participants in these situations, for example an *agent* and an *experiencer* in the *Cause_to_experience* frame. Frames can also introduce non-core roles like *time* or *means* that are the same across all frames and that generally apply to adjuncts. The annotated sentences are manually selected from the British National Corpus (BNC, Burnard, 1995), a corpus of English drawn from a variety of genres and containing written as well as spoken data. The FrameNet resource (release 1.2) contains c. 57,000 verbal propositions and c. 2,000 verbs. The resource aims to present instances of each verb with all its roles and in all syntactic diatheses, which generally allows good coverage of roles, despite the relatively small size of the corpus.

The sampling method however implies that the corpus is not a representative sample of English. Therefore, when trained on the FrameNet corpus, our model relies on probability estimates that are not necessarily representative of every day language use. Our model is still able to make meaningful predictions because co-occurrence information for specific verbs and arguments is usually very sparse even in larger corpora, so that any probabilistic model essentially classifies seen and unseen events. This classification represents a very high baseline in semantically influenced tasks (see, e.g., its successful use in early work on prepositional phrase attachment by Hindle and Rooth (1993)). If a larger corpus with FrameNet-style annotation were available, our model would gain more coverage of specific verb-argument pairs and a finer-grained estimate of co-occurrence frequencies, both of which we expect to improve its predictions. In the absence of such a resource, we rely on the information available in the corpus and use smoothing techniques to generalize to unseen cases.

2.1.1. Smoothing.

To estimate the semantic probability model, we can use maximum likelihood estimation on word-co-occurrences in our training corpus. However, we encounter a serious *sparse data problem*: For instance, if we use the data from McRae et al. (1998) as a test set (see below), only 6% of all verb-argument-role triples are attested in the FrameNet corpus. For the remaining 94% of data points, the model would predict a probability of 0. A model

induced by maximum likelihood estimation alone therefore underestimates the plausibility of data points unseen in the training data.

We apply class-based smoothing (CB), a standard method used in computational linguistics to approach this problem. Class-based smoothing pools similar observations in the training data to arrive at a more robust probability estimate for each class member. In experience-based models of syntax (and probabilistic parsers in computational linguistics), abstract categories like parts of speech are used as classes. We make semantic generalizations instead by employing semantic verb and noun classes. The method therefore serves not only to avoid the problems of sparse data, but also to base the model's predictions on semantic generalizations rather than pure word co-occurrence. From a cognitive perspective, semantic categories are a much-researched basic tool for human reasoning about the world (see, e.g., Medin & Aguilar, 1999), and there is evidence for the existence of semantic classes as an organizational principle of the human mental lexicon (see, e.g., Aitchison, 2003). Class-based smoothing as inference about the plausibility of events based on semantic class membership therefore appears to be a plausible modeling tool.

Technically, when applying class-based smoothing to the semantic model, we estimate a joint probability distribution over semantic classes $P(cl_v, gf, r, cl_a)$ instead of over individual words $P(v, s, gf, r, a)$ and thereby base our estimate on a much larger set of relevant data points. Given a semantic noun class that contains *pirate* and *buccaneer* and a semantic verb class with *terrorize* and *terrify*, class-based smoothing allows us to count observations of *terrify-pirate-agent* and *terrorize-buccaneer-agent* to estimate the plausibility of *terrorize-pirate-agent*. This method is therefore especially well-suited to making reliable plausibility predictions even for unseen verb-argument combinations.

In the semantic plausibility model, we use class-based smoothing for both nouns and verbs. WordNet's synonym sets serve as noun classes (Fellbaum, 1998). These very fine-grained classes ensure valid generalizations and perform better than the coarse-grained set of WordNet unique-beginner (top-level) classes (Padó, 2007). However, fine-grained noun sets can contribute only relatively little smoothing power exactly because their generalizations are very specific. Most of the generalizations are in fact made by the verb classes, which we induce from the FrameNet training data (Padó et al., 2006). Our induced verb classes outperform hand-crafted classes such as VerbNet (Kipper, Dang, & Palmer, 2000) or WordNet because they are optimized for the task and the training set (Padó, 2007).

Verbs are clustered according to which roles they assign to their arguments, and how they realize them syntactically. We use an implementation of two soft clustering algorithms (Marx, 2004) derived from Information Theory: the Information Distortion (ID) (Gedeon, Parker, & Dimitrov, 2003) and Information Bottleneck (IB) (Tishby, Pereira, & Bialek, 1999) methods. Soft clustering allows us to identify and use verb polysemy, which is often characterized by different patterns of syntactic behavior for each verb meaning. Features for the clustering algorithms were the lemmas of the argument head of the verb, the syntactic configuration of verb and argument (as a path through a parse tree), the verb's sense (i.e., its FrameNet frame), the role assigned to each argument and a combined feature of role and syntactic configuration.

To choose the optimal values for the parameters *clustering algorithm* and *number of clusters*, we evaluated different parameter instantiations by comparing the quality of the semantic model's plausibility predictions when using the resulting clusters for smoothing.

Table 1: Example Clusters: Top ten verbs from two induced clusters.

Cluster 1	Cluster 2
resent	cycle
envy	follow
dislike	travel
like	lead
hate	chase
prove	accompany
delight	escort
want	usher
argue	pursue
regret	trail

Evaluation was done on a development data set with 60 human ratings for verb-argument-role triples (a subset of the plausibility norming data from McRae et al., 1998, see below). For the FrameNet data, the ID algorithm performed best, and a set of 13 clusters proved optimal. Note that this is much fewer than the c. 300 verbal frames specified in the training data. Our verb classes thus constitute a compact, task-specific generalization of the information available in FrameNet. For a more detailed discussion of the clustering process, see Padó et al. (2006) and Padó (2007). Table 1 shows the top ten members of two of our induced clusters, sorted by their probability of cluster membership (all probability values > 0.84). Cluster 1, which also includes *terrorize*, has the common theme of experiencing (*like, dislike*) or causing emotion (*delight*). Cluster 2 contains verbs of motion. Other cluster topics include perception, modes of communication, or verbs of increase and change (e.g., *increase, soar*).

To broaden coverage in cases where CB smoothing does not return estimates, we also employ Good-Turing (GT) smoothing (see, e.g., Good, 1953; Manning & Schütze, 1999). This method re-estimates the model's probability distribution and assigns a uniform, small amount of probability mass to all events that are unseen in the training data (and thus receive a zero probability prediction in the unsmoothed model). Re-estimation of the training distribution also makes estimates for rare events (such as hapax legomena) more robust.

2.1.2. The smoothed model.

We combine CB and GT smoothing using a back-off strategy. Equation 6 illustrates our combination method using the decomposed model formulation. GT smoothing is always applied to the first four model terms, which are the least sparse. Since in these four terms we do not allow predictions for events that are unseen, to avoid overgeneration of inconsistent verb-sense-role combinations, GT smoothing of these terms mainly serves to smooth the counts for events that only appear once in the training data, because these are prone to

noise.

$$\begin{aligned} \mathit{Plausibility}_{v,r,a} &= P_{GT}(v) \cdot P_{GT}(v|s) \cdot P_{GT}(gf|v,s) \cdot P_{GT}(r|v,s,gf) \cdot \\ & \quad P_{BO}(a|v,s,gf,r) \end{aligned} \tag{6}$$

The final, sparsest model term $P_{BO}(a|v,s,gf,r)$ is estimated in a series of back-off steps (see Katz, 1987), given in Equation 7. Here, cl_v denotes the class of a verb, and cl_a denotes the class of an argument induced by the class-based smoothing algorithm.

$$P_{BO}(a|v,s,gf,r) = \begin{cases} P_{CB}(cl_a|cl_v,gf,r) & \text{if } f_{CB}(cl_a,cl_v,gf,r) > 0 \\ P_{CB}(cl_a|cl_v,r) & \text{if } f_{CB}(cl_a,cl_v,gf,r) = 0 \\ & \text{and } f_{CB}(cl_a,cl_v,r) > 0 \\ P_{GT}(cl_a|cl_v,r) & \text{else} \end{cases} \tag{7}$$

First, we try to estimate $P(a|v,s,gf,r)$ using class-based smoothing. Note that while the verb's sense s does not appear in the CB formula, the model generates the sense value that maximizes the plausibility equation while being compatible with the predicted role. If a combination of classes, grammatical function and role is unseen even after generalization, we apply class-based smoothing again, but remove the grammatical function term. While the grammatical function information may yield useful hints about the intended role if it is present, it is not central to determining the plausibility of a verb-argument-role triple. If class-based smoothing fails entirely, we back off to a GT estimate of seeing an unknown combination of classes.

In cases where the model has to rely on GT smoothing only, there is an advantage to using the decomposed formulation over the joint formulation. In the decomposed formulation, the less sparse first four model terms contribute information about the verb's preferred syntactic and semantic realization of its arguments that is lost if the joint probability model is smoothed with a uniform estimate for all unseen combinations of the five model variables. We therefore use the decomposed model formulation below.

Note also that Equation 7 is simplified for ease of exposition. In order to ensure that a probability distribution is returned by the back-off sequence, the back-off terms have to be weighted appropriately: The total probability mass returned by each back-off step has to be scaled to take up only the mass assigned to unseen events by the previous step (see, e.g., Dagan, Pereira, & Lee, 1994, for a suitable scaling factor).

2.2. Experimental evaluation

The semantic model's appropriateness for its task can be tested by using it in isolation to predict human plausibility intuitions. We investigate the performance of the smoothing methods and demonstrate the quality of the smoothed model's predictions and its wide coverage of unseen input data.

Four example test data points from McRae et al. (1998) are presented in Table 2. Each triple of verb, argument and role is associated with an average human plausibility rating on a 1–7 scale. The ratings were collected by asking participants to answer questions like *How common is it for a pirate to terrorize someone?* (probing the agent relation between *pirate* and *terrorize*) with the rating that seemed appropriate. The experiencer relation between

Table 2: Test item: verb-argument-role triples with plausibility ratings from McRae et al. (1998); scale ranges from 1 (implausible) to 7 (plausible).

Verb	Argument	Role	Rating
terrorize	pirate	agent	6.5
terrorize	pirate	experiencer	2.2
terrorize	victim	agent	1.4
terrorize	victim	experiencer	6.6

pirate and *terrorize* was probed by asking *How common is it for a pirate to be terrorized by someone?*

The model's task is to predict the human rating given the verb, argument and role. We correlate the plausibility values predicted by the model (probabilities ranging between 0 and 1) and the human judgments (average ratings ranging between 1 and 7). Since the judgment data is not normally distributed, we use Spearman's ρ (a non-parametric rank-order test); ρ ranges between 0 and 1, where a value of 1 indicates a perfect correlation.

2.2.1. Training and test data.

We train the model on the FrameNet corpus, release 1.2, and present results from two test sets. The first is a set of norming data from the literature. We use the data for 25 randomly chosen verbs (corresponding to 100 data points) out of the 160 data points reported in McRae et al. (1998) (the remainder serves as a development set for parameter optimization). Recall that in this data set, each verb is paired with two arguments and two roles each so that each verb-argument pair is plausible in one role and implausible in the other, as shown in Table 2. The balancing of plausible and implausible verb-argument-role triples means that the semantic model can only correctly predict the judgments if it indeed uses semantic plausibility information (rather than just relying on general syntactic role preferences). The judgment prediction task is very hard to solve if the verb is unseen during training, since its identity determines the set of applicable thematic roles.[3] We therefore exclude items with unseen verbs from the test data, retaining 64 of the original 100 data points.

The second test set, from Padó et al. (2006), allows us to explore the semantic model's performance on items which were extracted from corpus data, namely the FrameNet and the PropBank corpora. We chose 18 verbs that occur in both FrameNet and PropBank according to the roles they assign in VerbNet: Six experiencer verbs like *hear*, six patient verbs like *hit* and six communication verbs like *tell*. For each verb, we extracted six arguments from each corpus: The three most frequent arguments in the preferred subject role and the three most frequent arguments in the preferred object role. We constructed verb-role-argument triples by combining each verb-argument pair with both roles, obtaining 24 verb-role-argument triples per verb, and elicited ratings on a seven-point scale for each triple in a web-based study. In all, there are 414 verb-role-argument triples instead of the

[3]While it is conceivable to set up the model to induce the closest FrameNet frame for an unseen verb, this is an ambitious research project that is complicated by the problem of having seen only one instance of the unknown verb.

Table 3: Semantic Model Performance: Test set size, coverage and correlation strength for McRae and Padó test sets using different smoothing regimes.

	McRae				Padó			
Smoothing	N	Coverage	Spearman's ρ^a		N	Coverage	Spearman's ρ^a	
None	64	6%	-0.316,	ns	414	27%	0.364,	***
GT	64	88%	0.032,	ns	414	99%	0.170,	***
CB+GT	64	88%	0.415,	**	414	99%	0.522,	***

[a]ns: not significant, $**: \ p < 0.01$, $***: \ p < 0.001$

full $24 \times 18 = 432$, because some arguments were seen in both corpora. This approach weakens the balancing seen in the McRae data, where each argument is highly plausible in one role and highly implausible in the other, but there remains a clear tendency towards this behavior in the data.

By definition, all the verbs in this test set are covered by FrameNet, and roughly one quarter of the verb-argument-role triples are present in the FrameNet training data. This allows the investigation of the model's performance when the sparse data problem is less pressing and when the test vocabulary is more similar to the training data than when using hand-crafted items.

2.2.2. Results and discussion.

Table 3 reports the semantic model's coverage of the test set items and the correlation between predicted and observed human judgments. We also present results for the GT smoothing method and the unsmoothed model for comparison.

The unsmoothed results highlight the severity of the sparse data problem: For the 64-data-point McRae data set, predictions can be made for only 6% of all data points, and the correlation is negative and non-significant. The Padó data set was designed to contain more seen data points. The predictions for the 414 Padó data points are significantly correlated to the human judgments, demonstrating that a probabilistic corpus-based model is capable of making accurate predictions for seen triples.

GT smoothing alone allows only poor predictions, especially for the almost completely unseen McRae data set. While the decomposition of the model into separate, less sparse subterms supplies some verb-specific preferences, the smoothing method does not make argument-specific plausibility predictions. Therefore, it cannot capture the crucial thematic fit variations in the data sets. However, coverage has increased significantly, of course. Some items still remain uncovered due to a restriction which we have placed on the model to ensure consistency of the role predictions: Only thematic roles that have been seen with the verb during training may be predicted. This includes cases where the verb was observed in a different sense from the one probed by the test data, so that the correct role cannot be predicted given the training data. The correct role is unseen with the verb more often for the McRae data set, which differs in genre from the training data more than the Padó data set.

Adding the CB smoothing method to the GT smoothed model finally supplies argument specific smoothing information. In contrast to the first two results, the fully smoothed

semantic model achieves significant correlations with the human data with realistic coverage. For the McRae data set, this is owed almost completely to the semantic generalizations made in CB smoothing, since virtually all data points are unseen and GT smoothing alone did not succeed. For the Padó test set, the application of CB smoothing even increases the correlation coefficient noticeably over that for the seen data points only, at almost perfect coverage. To interpret the coefficients, human performance can serve as a point of comparison. A human rater's judgments predict the average of the other raters' judgments at about $\rho = 0.7$ (Padó, 2007). While the model performs below this level, its performance is still substantial in comparison.

These results suggest that our smoothing methods are appropriate and allow the model good performance on a test set of almost completely unseen data points. Not surprisingly for a probabilistic approach, the model performs best on a test set that is more similar to the training data and contains some seen data points. This evaluation demonstrates that the semantic model is capable of predicting human judgments for new data sets. This makes it a key component of the SynSem-Integration Model, which we now go on to discuss.

3. The Syntax-Semantics Integration model

The model of semantic plausibility introduced above allows us to integrate semantic information with an existing approach to modeling syntactic preferences. The resulting SynSem-Integration model of human sentence processing reliably predicts sentence processing difficulty observed in experimental studies and is capable of processing unrestricted input data, thus displaying wide coverage of language data.

The SynSem-Integration model is derived from a probabilistic grammar-based model in the tradition of Jurafsky (1996) and Crocker and Brants (2000) because this type of model explains the creation of syntactic analyses as well as the resolution of ambiguities. As mentioned above, grammar-based models cannot easily account for semantic effects directly, as the information about word co-occurrence they can capture is at the syntactic level only and extremely sparse. Therefore, we add a dedicated semantic model. The existence of separate syntactic and semantic models should not be taken as a claim about cognitive reality, but rather serves to improve the transparency of the combined model and to allow the separate evaluation of each component.

Fig. 1 illustrates the architecture of the SynSem-Integration model: The syntax model incrementally computes all possible analyses of the input. The semantic model's task is to evaluate the resulting structures with respect to the plausibility of the verb-argument pairs they contain. Both models simultaneously rank the candidate structures: The syntax model ranks them by parse probability, and the semantic model by the plausibility of the verb-argument relations contained in the structures. The two rankings are interpolated into a *global* ranking which allows the prediction of a humanly preferred structure, as in a grammar-based model. Depending on the interpolation parameter for the global ranking, either source of information can dominate the preferred structure prediction.

Difficulty is predicted with respect to the global ranking and the two local rankings, by taking up elements of the difficulty prediction strategies in both probabilistic grammar-based and constraint-integration models. As in a competition-based constraint-integration model, difficulty is predicted if the information sources disagree in their support for the globally

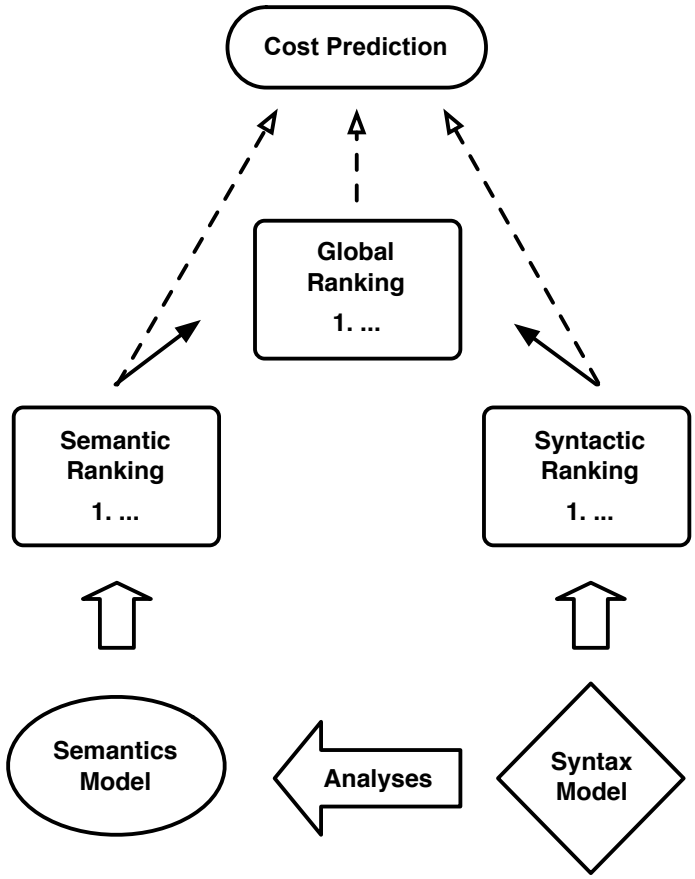

Figure 1. The architecture of the SynSem-Integration model.

preferred structure. This means that even if one model dominates the global ranking, the other model's preferences are still vital for difficulty prediction. As in both Jurafsky-style grammar-based models and constraint-based models, difficulty is also predicted if new evidence leads to the abandoning of the globally preferred structure in favor of another one.

In the following, we first describe the implementation of the syntactic model. We then go on to discuss the difficulty prediction strategies of our model and existing probabilistic approaches. This leads us to describe the parameter space for cost prediction in the SynSem-Integration model, and the setting of these parameters on a held-out set of observed patterns of human processing difficulty. Finally, we present an evaluation of the SynSem-Integration model against experimental data on four locally ambiguous constructions, from a total of eight experimental studies.

3.1. The syntactic model

The SynSem-Integration model incorporates a probabilistic grammar-based model as a source of information about lexical and syntactic preferences. As in Jurafsky's (1996) approach, the grammar-based model proposes analyses of the input based on a probabilistic context-free grammar (PCFG). Fig. 2 gives an example of PCFG rules of the form $N \rightarrow \zeta$

1.	S	→	NP VP	1.0	*6.*	V	→	terrorized	.8
2.	NP	→	DT N	1.0	*7.*	V	→	slept	.2
3.	VP	→	V NP	.9	*8.*	N	→	pirate	.5
4.	VP	→	V	.1	*9.*	N	→	sea	.5
5.	DT	→	the	1.0					

Figure 2. Example of a PCFG fragment: Numbered $N \rightarrow \zeta$ rules annotated with rule probabilities.

1.	S	→	NP VP[terrorize]	.8	*6.*	V[sleep]	→	sleeps	0.5
2.	S	→	NP VP[sleep]	.2	*7.*	V[sleep]	→	slept	0.5
3.	VP[terrorize]	→	V[terrorize]	.2	*8.*	V[terrorize]	→	terrorized	0.7
4.	VP[terrorize]	→	V[terrorize] NP	.8	*9.*	V[terrorize]	→	terrorizes	0.3
5.	VP[sleep]	→	V[sleep]	1.0					

Figure 3. Example of a partially lexicalized PCFG fragment: Numbered $N \rightarrow \zeta$ rules annotated with rule probabilities.

(N rewrites as ζ) with rule probability $P(N \rightarrow \zeta)$. This grammar covers sentences like *The pirate slept* or *The pirate terrorized the sea*. The probability of a syntactic structure T for an input sentence can be computed by multiplying the probabilities of the grammar rules involved in constructing T, as expressed in Equation 8:

$$P(T) = \prod_{(N \rightarrow \zeta) \in T} P(N \rightarrow \zeta) \tag{8}$$

The probability of analyzing *The pirate slept* as a sentence composed of a noun phrase and a verb phrase that is a single verb is thus 0.01 (using rules 1, 2, 4, 5, 7 and 8).

Like Crocker and Brants (2000), we use a wide-coverage grammar induced from a large corpus of syntactically annotated data. This grammar is able to account for all syntactic phenomena encountered in the corpus and can thus make correct structural predictions also for input that was not encountered during training. This allows our syntactic model wide coverage of phenomena and the ability correctly process unseen input.

We use a *lexicalized* model that contains not only purely structural information, but also preferences associated with single lexical items, such as lexical category preferences or verb subcategorization preferences (Jelinek, Laerty, Magerman, & Roukos, 1994; Collins, 1996). As shown in Fig. 3, a lexicalized grammar not only contains information about the internal structure of phrasal categories, but also about the lexical heads involved. This information allows the grammar to capture structural preferences that are specific to given lexical heads. The grammar fragment in Fig. 3 for example encodes verbal subcategorization information: Rule 5 states that *sleep* is an intransitive verb, always forming a VP without a noun phrase argument, while according to rule 6 *terrorize* is preferably transitive. In contrast to the semantic model, the lexicalized grammar does not distinguish between verb senses, since no sense information is annotated in the training corpus (if sense information were given, it would be possible to distinguish, e.g., between the preferred argument patterns of different verb senses).

Fig. 3 shows a *head-lexicalized* grammar with lexicalization for the head of each phrase. It is possible to include further information about lexical heads observed together in some

Table 4: Bracketing recall and precision, F-score and coverage of a lexicalized and a fully head-head lexicalized parser on WSJ Section 23.

Parser	Recall	Precision	F	Cov.
Head-Head Lexicalization	86.47	86.65	86.49	100%
Head Lexicalization	86.17	86.31	86.29	100%

syntactic relation, for example as a verb and its argument. Such a *head-head lexicalized* grammar could use this co-occurrence information to differentiate between syntactic analyses with different verb-argument configurations.

However, in practice, this approach demands larger amounts of syntactically annotated training data than are available today. Results by Gildea (2001) and Bikel (2004) suggest that the relevant head-head lexical information is so sparse that it is rarely available in the parsing of unseen text using standard training corpora like the Penn Treebank (Marcus et al., 1994). This is especially true if the domain of the training data differs from that of the test data, as is the case for a probabilistic grammar-based model trained on the standard newspaper corpora and used to analyze experimental items. Therefore, we expect that head-head lexicalization will not improve parsing performance on unseen test data much, and also that a head-head lexicalized grammar-based model will not be able to distinguish between possible syntactic analyses on the basis of the available head-head co-occurrence information.

Evaluation We test the assumption that head-head lexicalization does not improve parsing for our purposes by analyzing the parsing performance of a head- and a head-head lexicalized grammar. We use the incremental top-down parser proposed by Roark (2001) as a parsing engine. We derive the two lexicalized grammars from the standard training data for syntactic parsers, sections 2–21 of the Wall Street Journal section of the Penn Treebank (WSJ Marcus et al., 1994). We add the data from section 24, to gain as much lexically-specific information as possible, and retain section 22 as a development set. Section 23 is the standard test set for probabilistic parsers. We slightly modify this training data to distinguish between adverbial PPs and agent PPs in passive constructions by introducing a new phrase label for agent PPs.

We present evaluation results both for the head lexicalized syntactic model and the head-head lexicalized version. Table 4 summarizes the results obtained on the WSJ section 23. We report the standard measures coverage and parsing F score, based on bracketing precision and recall across the best parses. Precision measures how many of the proposed syntactic nodes are correct, punishing predictions with incorrect nodes. Recall gives the proportion of correctly proposed tree nodes over the number of nodes in the target tree, punishing predictions with missing nodes. F score is the harmonic mean of precision and recall, $F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$.

The results in Table 4 demonstrate first the wide coverage of both models, which are capable of assigning structure to all sentences in the unseen test data. Further, both models' structural predictions match the human annotations on the test data to a large degree, which allows us to assume that the predictions are mostly reliable. Finally, we observe

that, as expected, both models perform very similarly. The additional information present in the head-head lexicalized grammar does improve parsing decisions, but not by a great margin, because for most structures, the relevant head-head co-occurrence information does not exist. We will therefore use the simpler head lexicalization technique for the syntactic model's grammar.

The head lexicalized parser proposes a large number of analyses for each input, many of which are very unlikely. To restrict the amount of analyses that have to be processed in the SynSem-Integration model, we follow Jurafsky (1996) in introducing a search beam which contains only analyses within a certain probability range. [4] We base difficulty prediction only on analyses with probabilities up to two orders of magnitude away from the best parse's probability. The introduction of a search beam avoids the need to assume full syntactic parallelism in human sentence processing and takes into account the existence of memory limitations.

3.2. Difficulty prediction

The SynSem-Integration model predicts processing difficulty on the basis of semantic and syntactic preferences determined by the semantic plausibility model and the syntactic parser model introduced above. This section discusses difficulty prediction in the SynSem-Integration model in relation to the strategies used in other approaches. We base our discussion on the observation that in isolated sentences with local syntactic ambiguities, human processing difficulty may be observed in two regions: During the processing of an ambiguous region, there may be conflicting evidence from different information sources, and at the point of disambiguation towards one of the alternative analyses, a previously preferred analysis may have to be abandoned in favor of a previously dispreferred one. We term these situations *Conflict* and *Revision*.

Take again the Main Clause/Reduced Relative (MC/RR) ambiguity as an example. We repeat Sentence (1-b) from above as Sentence (2):

(2) The victim terrorized by his captors was freed quickly.

Recall that the ambiguous region up to *terrorized* has two possible interpretations: A main clause continuation, and the reduced relative continuation as in (2). In the main clause analysis, the victim is the semantic subject of the terrorizing event, while in the reduced relative analysis, it is the semantic object. During this ambiguous region, a *Conflict* situation arises if there is conflicting evidence for which of the two analyses to prefer. In sentence (2), the main clause analysis is syntactically more likely, being much more frequent. However, semantically, the *victim* is much more likely to be the theme rather than the subject of the terrorizing action. The conflicting syntactic and semantic preferences cause processing difficulty.

The ambiguity continues until the prepositional phrase *by his captors* provides a strong syntactic bias towards the reduced relative interpretation. *Revision* difficulty may be observed if the processor initially preferred the main clause interpretation, but now abandons it. The main verb cluster, *was freed*, disambiguates completely: Only the reduced

[4] The search beam limits the amount of analyses used in predicting difficulty; the parser's internal representations of partial parses are not affected.

relative interpretation is syntactically plausible now. Even readers who held on to the main verb interpretation until this point are forced to revise and may encounter difficulty.

A successful model of human sentence comprehension needs a means of predicting difficulty in both Conflict and Revision situations. The SynSem-Integration model bases its predictions on two cost functions specific to these situations. We discuss the cost functions employed in existing models and compare them to the ones used in our model.

Conflict during the processing of an ambiguous region is handled naturally by competition-based constraint-integration models, where difficulty is predicted by competition of strong opposing constraints which delay the identification of a preferred interpretation. Grammar-based models in the Jurafsky tradition, on the other hand, use a difficulty prediction function that only reacts to a change in the proposed preferred syntactic structure. Since a Conflict situation does not necessarily lead to such a change (the most probable syntactic analysis of the input may remain the same despite conflicting preferences), these models do not account consistently for this source of processing difficulty.

In the SynSem-Integration model, difficulty due to Conflict is predicted if either the syntactic or the semantic model does not agree with the globally preferred structure. This is equivalent to a conflict between the preferences of the syntactic and semantic models, since the globally preferred structure is based on an interpolation of both models' rankings. In the Conflict situation, the SynSem-Integration model thus relies on a similar mechanism as competition-based models.

Revision occurs if a reader gives up a previously preferred analysis for a different one. Probabilistic grammar-based models easily capture Revision situations as they predict processing difficulty if the preferred syntactic structure changes. This cost function can be seen as an abstraction of the process employed by competition-based constraint-integration models, which predict processing difficulty in Revision situations due to the competition between the well-supported previously preferred analysis and the strong activation from new evidence received by the other analysis. Both existing proposals for cost prediction thus capture the complexity involved in abandoning one interpretation of the input in favor of another.

The SynSem-Integration model uses a similar prediction function to that of a probabilistic grammar-based model. A conflict-based account of the Revision situation, as in constraint-based models, is not open to the SynSem-Integration model on technical grounds, because it operates strictly on the set of possible syntactic analyses of the current input. If syntactic disambiguation completely rules out the preferred analyses of the previous time step, its semantic interpretation is no longer available to compete with the interpretation of the confirmed alternative analysis. Therefore, the SynSem-Integration model detects a Revision situation by tracking the preferred structure at each point in processing, like probabilistic grammar-based models.

Together, the Conflict and Revision cost functions guarantee that the SynSem-Integration model can make difficulty predictions during the whole course of ambiguity processing. The total cost predicted by the model is the sum of all Conflict and Revision cost incurred in a region (it is possible for both cost types to be incurred simultaneously,

if the globally preferred analysis has changed, but another analysis is semantically more plausible).

Granularity of Predictions A further aspect of cost prediction that is worth comparing is the granularity of predictions. Models have a choice between three levels of granularity for difficulty predictions: We call predictions that are binary flags for the existence of difficulty *qualitative* predictions, predictions that specify the relative size of processing difficulty *relative-quantitative* predictions, and predictions that directly link a model's output to reading times in milliseconds *absolute-quantitative* predictions. Absolute reading times are known to depend also on factors like word length, word frequency and predictability (Just & Carpenter, 1980; McDonald & Shillcock, 2003), which are not considered in any of the models discussed here (see, e.g., Demberg and Keller (2008); Boston, Hale, Kliegl, Patil, and Vasishth (2008) for models of absolute reading times for newspaper text).

The models introduced above fall into different classes on this scale. Probabilistic grammar-based models using the cost function introduced by Crocker and Brants (2000) make qualitative predictions by announcing the presence of difficulty if a change in preferred structure takes place. This type of prediction is quite imprecise, since it does not give an indication of the relative difficulty encountered in the region in comparison to other regions. The settling time of competition-based constraint-integration models, on the other hand, predicts relative processing difficulty and therefore constitutes a relative-quantitative prediction.

The SynSem-Integration model's per-condition predictions are also relative-quantitative. Recall that we define the final cost prediction for the processing of an input region as the average cost predicted over all stimuli. Cost predictions therefore depend not only on the amount of difficulty predicted for individual stimuli and the granularity of those predictions, but also on the number of stimuli for which difficulty is predicted. The model's predictions thus reflect the relative processing ease for a condition with many easy stimuli in comparison to one with many difficult ones. The granularity of the model's per-item predictions depends on the cost function used. We will discuss cost functions of different granularity in the next section. We will show that the most reliable per-condition predictions are made by binary or coarse-grained relative-quantitative cost functions, which are most resistant to noise.

3.3. Parameters of the model

Having discussed the component models and the cost prediction mechanism of the SynSem-Integration model, we now conclude the description of the model by discussing the setting of the cost prediction parameters. There are two types of parameters: The first is the interpolation factor used to compute the global preference score. The other is the implementation of the two cost functions that predict difficulty. We introduce both types of parameters and then describe the parameter selection process, during which the SynSem-Integration model's performance on a development set is optimized.

3.3.1. The interpolation factor. The interpolation factor f is used to compute the global preference score for the candidate analyses a_i. The global score of the analyses determines the globally preferred syntactic structure, which has to be known for cost prediction.

The interpolation factor f determines the respective influence of the syntactic and semantic scores predicted by the two model components, as shown in Equation 9. *Syn* is the probability of the syntactic analysis assigned to interpretation i by the parser and *Sem* is the semantic plausibility score assigned by the semantic plausibility model.

$$Global\ score(a_i) = f \cdot Syn(a_i) \cdot (1 - f)\ Sem(a_i) \tag{9}$$

The interpolation factor f ranges between 0 and 1. The larger this factor, the more the syntactic probability of an analysis dominates its global score (i.e., the more similar the global ranking of analyses becomes like the ranking based on the syntax score).

3.3.2. The cost functions.

The second type of model parameter is the exact formulation of the cost functions used for difficulty prediction. Recall that the SynSem-Integration model employs a combination of two cost functions tailored to the Conflict and Revision situations in human sentence processing identified above. Since each of the cost functions applies to only one source of difficulty, their output is simply added to predict overall difficulty for an incremental processing step.

Conflict cost quantifies the processing difficulty incurred in situations where the input yields conflicting evidence for which analysis to prefer, while Revision cost accounts for the processing difficulty caused by abandoning a preferred interpretation of the input and replacing it with another. Cost prediction in either of these situations can be instantiated by cost functions with different granularity of prediction. We define three alternatives each for computing Conflict and Revision cost and evaluate their appropriateness during parameter setting. Recall that the granularity of the cost functions only affects the grain size of the SynSem-Integration model's per-item predictions, not that of its per-condition predictions (see Section 3.2).

Conflict cost is predicted on the basis of the insight from competition-based models that processing difficulty can be explained by a conflict between strong disagreeing constraints. The conflict cost functions in the SynSem-Integration model therefore are sensitive to differing structural preferences in the two information sources. Take $rank_{syn}$ and $rank_{sem}$ to denote the syntactic and semantic rank[5] of the globally preferred analysis gp. We define three cost functions, presented here in the order of increasing fineness of granularity.

1. *Fixed Cost*: $cost_{conflict} = \begin{cases} 1 & \text{if } rank_{syn}(gp) \neq rank_{sem}(gp) \\ 0 & \text{else} \end{cases}$

Fixed Cost is a qualitative measure which predicts binary difficulty by assigning a cost of 1 if the rank of the globally preferred analysis differs in the syntactic and semantic models. This is the simplest possible way of modeling a Conflict situation in the SynSem-Integration model.

2. *Rank Cost*: $cost_{conflict} = abs(rank_{syn}(gp) - rank_{sem}(gp))$

Rank cost computes Conflict cost as the difference between the ranks assigned to the globally preferred analysis by the two models. For this function, no cost is incurred if the globally

[5]Note that analyses with identical scores are assumed to share a rank, so there can be two equally preferred analyses. In these cases, as long as one of the equally preferred analyses corresponds to the globally preferred one, no difficulty is predicted.

preferred analysis is ranked first in both models, and growing amounts of cost are assigned the lower the globally preferred analysis is ranked in a disagreeing model. This cost function is motivated by the intuition that more cost should be incurred in a Conflict situation if the rankings of the syntactic and semantic model differ widely than if they differ by only one rank position. Since it captures the strength of the disagreement between the models, it allows relative-quantitative predictions.

3. *Ratio Cost*: $cost_{conflict} = \begin{cases} \frac{p_{syn}(lp)}{p_{syn}(gp)} & \text{if } rank_{sem}(gp) > rank_{syn}(gp) \\ \frac{p_{sem}(lp)}{p_{sem}(gp)} & \text{if } rank_{syn}(gp) > rank_{sem}(gp) \\ 0 & \text{else} \end{cases}$

Ratio cost, the most fine-grained relative-quantitative measure, considers the probability ratio between the locally preferred (lp) analysis put forward by the disagreeing model and the value that this model assigns to the globally preferred (gp) analysis (the one that is ranked highest in the overall ranking). This function is a more graded implementation of Rank cost, such that a structure that is dispreferred in the disagreeing model by a small margin incurs less cost than one that is much less likely than the highest-ranked analysis. Predicted cost larger than zero is scaled by the logistic function $\frac{1}{1+e^{-cost}}$ to values between 0.5 and 1 to avoid an explosion of cost if the locally preferred analysis is much more likely than the globally preferred analysis.

Revision cost We also identify three Revision cost functions that apply when the semantic interpretation of the globally preferred analysis changes from the last processing step. We take this to be the case when the set of verb-argument pairs in the current semantic interpretation is not equal to or a monotonic extension of the set derived from the preferred semantic analysis at the last time step.[6] Here, $set(gp_t)$ denotes the set of verb-argument pairs associated with the globally preferred syntactic structure gp at time step t, and $p_{sem}(gp_t)$ denotes the semantic plausibility of gp at t. Again, we present the three cost functions in order of increasing fineness of granularity.

1. *Fixed Cost*: $cost_{revision} = \begin{cases} 1 & \text{if } set(gp_t) \not\supseteq set(gp_{t-1}) \\ 0 & \text{else} \end{cases}$

Fixed cost as a qualitative cost function assigns a fixed penalty of 1 if the set of verb-argument pairs in the globally preferred parse at t is not a monotonic extension of the semantic representation of the globally preferred parse from the previous time step. This is the cost function used in non-surprisal probabilistic grammar-based models since Crocker and Brants (2000).

2. *If-Worse Cost*: $cost_{revision} = \begin{cases} 1 & \text{if } set(gp_t) \not\supseteq set(gp_{t-1}) \\ & \text{and } p_{sem}(gp_t) < p_{sem}(gp_{t-1}) \\ 0 & \text{else} \end{cases}$

The If-Worse function is a qualitative modification of the Fixed cost function. It only assigns a fixed Revision cost if the set of verb-argument pairs in the globally preferred structure

[6]Note that we do not pay attention to the roles assigned to the verb-argument pairs, because role reassignment does not appear to incur cost as long as the syntactic structure remains the same (e.g., *He loaded the truck*$_{Goal}$, which is easily reanalyzed into *He loaded the truck*$_{Theme}$ *onto the boat*$_{Goal}$, upon encountering *onto the boat* Pritchett, 1992).

has changed *and* the semantic analysis of the globally preferred parse is less probable than the preferred one at the last time step. The intuition behind this modification is that a semantically equal or more acceptable interpretation should be adopted more readily than one that is less satisfying to the comprehender than the previously preferred one.

3. *Ratio Cost*: $cost_{revision} = \begin{cases} \frac{p_{sem}(gp_{t-1})}{p_{sem}(gp_t)} & \text{if } set(gp_t) \not\supseteq set(gp_{t-1}) \\ & \text{and } p_{sem}(gp_t) < p_{sem}(gp_{t-1}) \\ 0 & \text{else} \end{cases}$

The Ratio cost function is a relative-quantitative version of the If-Worse function. It assigns the ratio of the semantic probabilities of the last preferred analysis and the current preferred analysis, capturing the difference in semantic preferredness between the two instead of assigning a fixed penalty. Cost is then scaled by the logistic function $\frac{1}{1+e^{-cost}}$, as for the Ratio Conflict cost function, to avoid an explosion of cost if the current best analysis is much less likely than the last preferred analysis.

3.3.3. Parameter setting.

The model parameters (interpolation factor f and cost functions) are chosen so that the model predicts an experimentally observed pattern of human processing difficulty with maximal accuracy. As a development set, we use a data set from the Garnsey et al. (1997) reading time study, namely the reading times for equibiased verbs. This data set was chosen because it shows statistically significant effects and yields a relatively large number of stimuli for processing by the SynSem-Integration model, and because there was a sufficient number of other data sets for the same phenomenon (the NP/S ambiguity, see Materials below) available for testing.

The data set contains a total of four reading time measurements, taken during two critical regions in two conditions. The SynSem-Integration model's task is to process the original experimental items and to predict the observed pattern of difficulty as closely as possible from them. We use the results for the total-time measure, since the model's predictions do not extend to the level of early versus late effects. The total time measure sums all fixations on the region in question and reflects the total time spent inspecting the region, be it during early or later processing.

The experimental observations and the predictions of the SynSem-Integration model are scaled to indicate the percentage of difficulty contributed by each region as proposed in Narayanan and Jurafsky (2005). This is more appropriate than using unscaled predictions and observations, since the model does not intend to directly predict reading times or reading time differences, but the occurrence of relative difficulty due to processing mechanisms. We scale separately for each condition by normalizing each region's observed or predicted difficulty by the total difficulty observed or predicted across all regions.

We evaluate a range of different parameter values according to the quality of predictions that they allow the SynSem-Integration model to make. Parameter settings that cause the model's predictions to exhibit a different pattern from the observed data are rejected, and settings that emulate the observed pattern as closely as possible are preferred. We further differentiate between the parameter settings that lead to qualitative acceptable predictions by the size of the correlation coefficient between predictions and observations (although we do not report the significance level for the correlation, since only four data points are available).

Table 5: Best-performing interpolation factors for different cost function combinations.

		f Range [a]	
Conflict Cost	Revision Cost	$r > 0.95$	$r > 0.99$
Fixed	Fixed	–	–
Fixed	If-Worse	0.7–1	–
Rank	Fixed	–	–
Rank	If-Worse	0.7–0.8	0.9–1
Ratio	Ratio	0.9–1	–

[a] 1: syntax only, 0: semantics only, –: No correct predictions

We evaluate ten values for the weighting parameter f (in 0.1 steps from 0 to 1) for each of five combinations of Conflict and Revision cost functions (we do not combine the Ratio cost functions with any of the others due to their vastly different granularity).

Results and Discussion Table 5 gives an overview over the parameter values that allow good qualitative predictions of the pattern of difficulty in the development data. The Conflict and Revision cost functions introduced above are reported with the range of values for the interpolation factor f that lead to qualitatively correct predictions. All reported values of f lead to a correlation coefficient of Pearson's $r \geq 0.95$ between the predicted and observed data points. The Rank/If-Worse combination with $f > 0.8$ leads to especially good predictions (Pearson's $r > 0.99$). We make several observations:

• For all successful model parametrizations, predictions become more like the observed development data the larger the interpolation factor is, that is, the more the syntax model determines the global ranking. Recall that the semantic ranking is always used for Conflict cost prediction, no matter what the global ranking is, so the resulting model is not equal to using a syntax-only model. The observation that "extreme is better" may be at least in part due to the fact that syntax and semantics are pitched against each other in the development data, leaving the constraints either in perfect agreement or exactly at odds. However, the range of f for which the non-probabilistic functions qualitatively predict the experimental observations is relatively wide. This indicates that the model is quite robust as long as the syntactic model has more weight in deciding the global ranking.

• The probability ratio approach, though appealing due to its fine grain size, does not allow us to predict the correct distribution of difficulty as well and across as broad a range of f values as the coarser-grained approaches. This is probably due to noise present in the two probabilistic component models.

• Only models using the probabilistic or If-Worse Revision cost function make qualitatively correct predictions. These cost functions postulate Revision cost only if the new globally preferred analysis is less plausible than the old one was.

In the evaluation of the SynSem-Integration model, we will primarily refer to the predictions of the best-performing Rank/If-Worse combination of cost functions with $f = 1$. To show that the model's predictions are robust across a range of model parametrizations, we will also report numerical evaluation results for the other two successful parametrization, Fixed/If-Worse with $f = 1$ and Ratio/Ratio with $f = 1$. Choosing $f = 1$ from the range

of possible values seems justified for two reasons: First, model performance increases with higher values of f, and second, this choice simplifies the model, as it reduces the global ranking of analyses to the syntactic ranking, effectively eliminating one of the three separate rankings required in the general case. Conflict can now be identified by directly comparing the syntactic to the semantic ranking, and Revision by tracking the preferred analysis in the syntactic ranking.

3.4. An example: the MC/RR ambiguity

We now present an example of the difficulty prediction process in the SynSem-Integration model, presenting the actual system output for the input sentence *The victim terrorized by his captors was freed quickly.* Fig. 4 gives a schematic overview over the four processing steps that we consider: The ambiguous verb, the beginning of the disambiguating *by*-phrase, the completion of the *by*-phrase and the main verb.

In the figure, each row in a table represents one possible syntactic analysis, characterized as the main clause (MC) or reduced relative (RR) interpretation. We also list the syntactic model's probability prediction (normalized over all analyses in the search beam) and the resulting syntactic ranking. This is complemented by the semantic model's ranking, normalized probability prediction and the underlying role assignment. We show data from the Rank/If-Worse, $f = 1$ model, so these two rankings are enough to determine Conflict and Revision cost at our chosen parameter settings. For the sake of brevity, we only list the relevant parses. The syntactic parser proposes several additional analyses, most of which differ on the level of part-of-speech labels (e.g., singular noun versus plural noun). Where there are real syntactic alternatives beyond the MC and RR interpretations, we mention them explicitly below.

At the first processing step, the ambiguous region, the main clause analysis is clearly syntactically preferred - its normalized probability is almost 0.9. However, this analysis implies that the *victim* is the semantic agent of the *terrorizing* event, which is highly unlikely. The semantic model markedly prefers to rank the main clause and reduced relative analyses in the opposite order for this item. The conflict between the syntactic and semantic ranking causes a prediction of processing difficulty in this region.

At the preposition *by*, the semantic ranking remains the same (*victim* is preferred to be the *experiencer* in syntactic object position), but the syntactic ranking changes. A reduced relative construction with an agent PP is now more likely than the main clause reading, where the PP has to be interpreted as an adverbial. In addition to these two analyses, the syntactic model also proposes a reduced relative analysis (not shown in the figure) that interprets the PP as an adverbial, as in *The victim$_{obj}$ terrorized (PP-Adv by the seaside) was freed quickly.* The change in preferred analysis from the main clause to the reduced relative interpretation prompts no Revision cost in the If-Worse cost function presented here, because the newly-preferred analysis is semantically more likely than the abandoned one. If we were using the Fixed Revision cost function, difficulty would be predicted. Note that a prediction of "no difficulty" on the item level does not mean that the region as a whole is predicted to show no processing difficulty, since the predictions over individual items are averaged for the region prediction, and noise in items and model will cause a non-zero difficulty prediction on average.

At the next time step, an explicit agent of the terrorizing event is processed. This does

Ambiguous Verb					
Syntactic Model					Semantic Model
MC: The victim$_{subj}$ terrorized	0.898	1.	2.	0.001	*terrorize-victim-agent*
RR: The victim$_{obj}$ terrorized	0.074	2.	1.	0.999	*terrorize-victim-experiencer*
by					
Syntactic Model					Semantic Model
RR: The victim$_{obj}$ terrorized (PP-Agt by)	0.787	1.	1.	0.999	*terrorize-victim-experiencer*
MC: The victim$_{subj}$ terrorized (PP-Adv by)	0.002	2.	2.	0.001	*terrorize-victim-agent*
Agent PP					
Syntactic Model					Semantic Model
RR: The victim$_{obj}$ terrorized (PP-Agt by his captors)	0.808	1.	1.	0.999	*terrorize-victim-experiencer/ terrorize-captors-agent*
MC: The victim$_{subj}$ terrorized (PP-Adv by his captors)	0.046	3.	2.	0.001	*terrorize-victim-agent/ terrorize-captors-means*
Main Verb					
Syntactic Model					Semantic Model
RR: The victim terrorized (PP-Agt by his captors) was freed	0.784	1.	1.	0.5	*terrorize-victim-experiencer/ terrorize-captors-agent/ victim-free-unk*

Figure 4. Processing an experimental item: Analyses, predicted normalized probabilities and rankings by the syntactic and semantic models.

not affect the syntactic or semantic ranking in comparison to the previous time step: Both models continue to prefer the reduced relative interpretation. The main clause analysis is unlikely both syntactically and semantically: The semantic model's interpretations assumes that the *captors* are the means by which the *victim* carries out the terrorizing event, which does not serve to increase the likelihood of *victim* as an agent. As for the previous time step, we do not show the reduced relative analysis that interprets the PP as an adverbial. Since both models agree in their ranking and no change in preferred analysis has taken place, no cost is predicted for this region.

Finally, on the main verb, only the reduced relative interpretation remains syntactically viable. The syntactic parser proposes to interpret *freed* either as a verb or as an adjective, resulting in two syntactic analyses with different main verbs, namely *freed* and *was*. Neither main verb is present in the semantic model's training data (cf. the role prediction of *unknown*), so the analyses are equally likely semantically and tied for first rank. In this case, no Conflict or Revision cost is predicted.

3.5. Evaluation of the SynSem-Integration model

We now turn to evaluating the SynSem-Integration model. We present the model's predictions of processing difficulty for four ambiguity phenomena: The Main Clause/Reduced Relative (MC/RR) ambiguity, NP object/Sentential Complement (NP/S) ambiguity, NP object/Clause Boundary (NP/0) ambiguity and PP-Attachment ambiguity. For each phenomenon, the model's predictions for two experimental reading-time studies are computed based on the original materials used in the studies. We present a qualitative evaluation for one study on each of the four phenomena to illustrate the SynSem-Integration model's predictions.

As a further step to evaluate the SynSem-Integration model as objectively as possible, we correlate its predictions with the processing difficulty observed in all eight studies (computed as the reading time difference between ambiguous and control conditions). This tests how the model performs over a range of studies, and assesses the relative difference predicted between all the observations.

3.5.1. Method.

As for parameter setting, we compare the SynSem-Integration model's predictions to the results reported for self-paced reading times or, in eye-tracking studies, for the total-time measure, which collects all fixations on the region in question and thereby reflects all effects of reading and re-reading visible in fixation durations. We use the results for the total-time measure since the model's predictions do not extend to the level of early versus late effects.

We create predictions for all critical regions (up to and including the disambiguation region) measured in the experimental data used for evaluation. The SynSem-Integration model's difficulty prediction for a region is the sum of the Conflict and Revision cost predicted in this region for all items, normalized by the number of items processed. We use the best-performing parameters determined on the development set, namely the Rank/If-Worse combination of cost functions and $f = 1$.

We base our predictions on all the items from any one study that can be processed by the SynSem-Integration model. This excludes items that cannot be parsed correctly. A

correctly parsed item is one where the preferred analysis at each point in processing is one of the alternative analyses that the experimenters assumed for the ambiguity. The syntactic model correctly parses between 32% and 83% of items across the eight studies, with a median of 57%. From these items, we further exclude items that cannot be processed by the semantic plausibility model because the target verb is unseen in training. Final coverage is between 27% and 75% of all items, with a median of 42%. For 80% of these items, the semantic model prefers one of the syntactic analyses assumed by the experimenters. For the remainder, it supports alternative analyses that either were not assumed present by the experimenters or are syntactic misparses.

In addition to the predictions by the SynSem-Integration model, we also report the predictions made by a head-head lexicalized probabilistic-grammar based model. This model serves as an informed baseline for the SynSem-Integration model's performance. It has the same syntactic information as the SynSem-Integration model's syntax model and can also use information on the co-occurrence of lexical heads in syntactic configurations to evaluate alternative parses. We use the head-head lexicalized grammar derived from the Penn Treebank that is described in Section 3.1 above. This model predicts difficulty whenever the best syntactic parse at the current time step is not a monotonic extension of the best parse at the last time step.

The experimental observations and the predictions of the models are again scaled as described in Section 3.3 above to reflect the proportion of overall processing difficulty contributed by each region. For each condition, we sum the observed or predicted difficulty over all regions and normalize each region's difficulty by the total. In the case of negative observed difficulty, we first move all observations for the affected condition into positive space by adding a constant value chosen to bring the lowest negative value to 1. This transformation preserves the relative position of the data points and allows us to apply the standard scaling procedure.

We evaluate both models' predictions by correlating the predicted and observed patterns of difficulty using Spearman's ρ, since the use of a parametric correlation test is not justified for all data sets.

3.5.2. The MC/RR ambiguity.

The influence of thematic fit on the processing of this ambiguity, introduced above in Section 3.2, was investigated, among others, by MacDonald (1994) and McRae et al. (1998). Both studies manipulated the thematic fit of the first NP with the verb as an agent or patient (varying *pirate* in the sentence *The pirate terrorized by his captors was freed quickly* with *victim*), testing whether a good agent like *pirate* biases readers towards the ultimately wrong main clause interpretation, while a good patient like *victim* might bias them towards the reduced relative reading.

MacDonald (1994), in her Experiment 2, also varied the number of possible analyses in the ambiguous region through the amount of disambiguating information present in post-verbal constituents Sentences (3-a) to (3-d) show a complete item set with all manipulations.

(3) a. The news stated that the microfilm concealed inside the secret passageway was discovered. (Good Patient/Early Disambiguation)
b. The news stated that the microfilm concealed most of the night was discovered. (Good Patient/Late Disambiguation)

c. The news stated that the spy concealed inside the secret passageway was discovered. (Poor Patient/Early Disambiguation)
d. The news stated that the spy concealed most of the night was discovered. (Poor Patient/Late Disambiguation)

The manipulation of post-verbal material consisted of varying the point at which the post-verbal phrases excluded a transitive main clause continuation of the sentences, thereby promoting the reduced relative meaning. Early Disambiguation materials as in (3-a) and (3-c) made this obvious at the first post-verbal word. Late Disambiguation materials as in (3-b) and (3-d) reliably excluded the transitive main clause only at the third or fourth word (*most of the* could still be continued to be a direct object, for example as *most of the documents*), giving the reader more time to entertain the initially preferred main clause hypothesis.

MacDonald (1994) found that a combination of good patient first NP and early disambiguation post-verbal material (both pointing towards the reduced relative) eliminated the difficulty at the disambiguating main verb. When the two information sources pointed towards different interpretations, she found some indication of difficulty at the disambiguation. When both information sources pointed towards a main clause, readers had significant difficulty at the disambiguating main verb.

McRae et al. (1998) used agentive *by*-phrases as post-verbal material, which corresponds to MacDonald's Early Disambiguation condition. They presented two words at a time and measured self-paced reading. They also found an influence of thematic fit: Readers found it harder to process ambiguous sentences with good patients at the verb+by region, where the good patients are implausible agents in the preferred main clause interpretation, but at the main verb, which disambiguates towards the dispreferred reduced relative interpretation, the good agent sentences were harder. We present the modeling results for this study below.

Qualitative Analysis We present modeling results for the McRae et al. data set, our running example in this paper. The reading time data was measured on the regions *verb+by*, *agent NP* and *main verb*. We make predictions for the *verb* and *by* separately, since both words contain cues for the processing system. The other regions are retained without modification. We plot the observed data both with the SynSem-Integration model's predictions (in Fig. 5) and with the baseline model predictions (in Fig. 6).

The SynSem-Integration model (gray lines in Fig. 5) predicts that stimuli with good patients should be harder to read at the verb than stimuli with good agents, because good patients introduce a conflict between the syntactic preference for the main clause reading and the semantic preference for the reduced relative. At *by*, both conditions are predicted to be similarly difficult, and in the agent NP region, our model predicts more difficulty for the good agent sentences than for the good patients. This reflects the revision of the previously well-supported main clause readings as the disambiguating region unfolds. At the main verb, our model predicts equally low difficulty for both conditions, most of the revision having taken place in the previous two regions.

We find these predictions mirrored in the experimental results, but one region late. Recall that the first experimental region combines the first and second region for which our model makes predictions (verb+*by*). In this long region, we see the difficulty with good

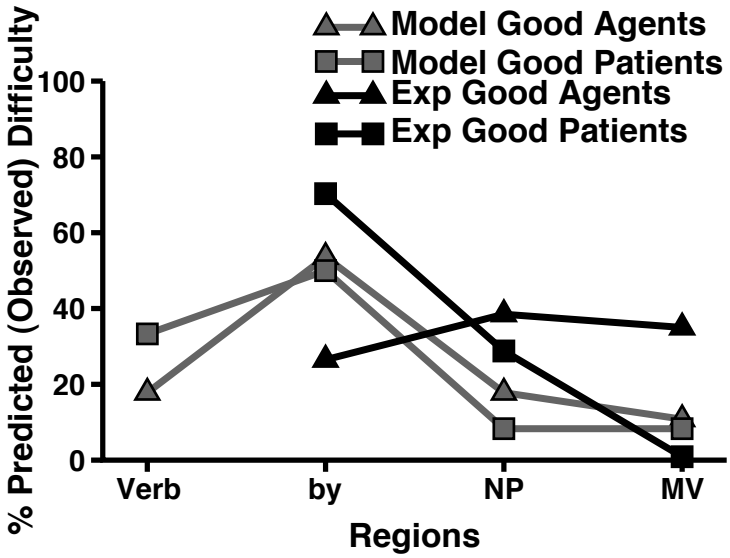

Figure 5. McRae et al. 1998: Experimental results and model predictions for the MC/RR ambiguity. GA: Good agent first NP, GP: Good patient first NP.

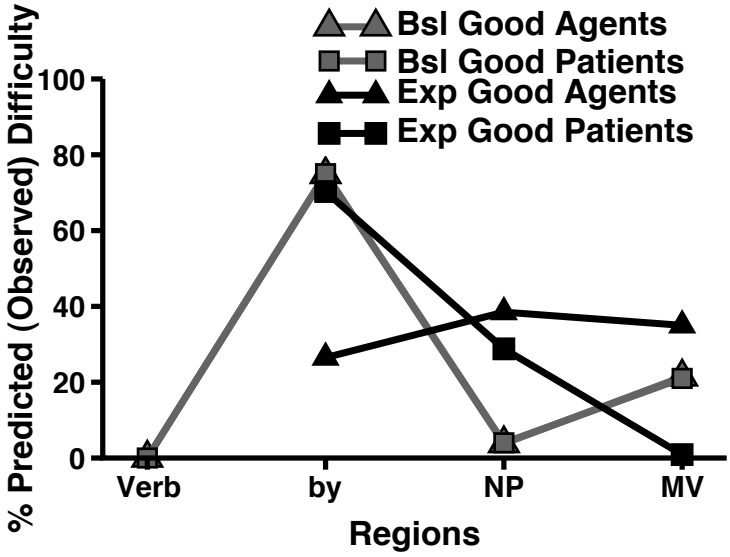

Figure 6. McRae et al. 1998: Experimental results and baseline predictions for the MC/RR ambiguity. GA: Good agent first NP, GP: Good patient first NP

patient sentences that was predicted by the model to be encountered at the verb. In the next region, difficulty for good agent and good patient sentences is relatively similar (the difference is not significant in the experimental results). Finally, good agent sentences prove to be significantly harder than good patient sentences. The discrepancy in timing between the model predictions and the observed data are presumably caused by two factors: First, the conflation of verb and *by* in the measurements, which makes it hard to exactly identify the onset of the difficulty with good agents, and second a *spillover effect* (Just, Carpenter, & Woolley, 1982), a phenomenon frequently found with self-paced reading data, where effects show or linger a region or two after their hypothesized onset.

The predictions of the syntactic baseline (see Fig. 6) are notably dissimilar from the observed data. The baseline model makes exactly the same predictions for both plausibility conditions, which is to be expected given our observations about the sparseness of head-head co-occurrence information that could yield clues to semantic plausibility. The model predicts a large amount of difficulty at the *by*-phrase followed by a smaller amount at the main verb. This distribution clearly reflects the difficulty encountered in purely syntactic processing: After an initial preference for the more frequent main clause interpretation, most stimuli are analyzed as containing a reduced relative at *by*, and the remainder switches the preferred analysis towards a reduced relative at the disambiguating main verb. The SynSem-Integration model's modulation of this general pattern by thematic fit effects more closely reflects the observed human behavior.

3.6. The NP/S ambiguity

The NP/S ambiguity results from the possibility to interpret a post-verbal NP as a direct object or as the subject of an embedded sentence complement, as in the example sentences (4-a) and (4-b) (from Pickering, Traxler, & Crocker, 2000).

(4) a. The criminal confessed his sins and reformed.
b. The criminal confessed his sins harmed too many people.

In sentence (4-a), *his sins* is a direct object in a main clause, but in the sentence complement reading shown in (4-b), the NP is part of an embedded sentence complement. Disambiguation towards the sentence complement reading follows immediately at the next word after the NP. In this ambiguity, readers usually initially interpret the second NP as the direct object of the main verb and show difficulty at a disambiguation towards the sentential complement interpretation.

Pickering and Traxler (1998) varied the thematic fit of the ambiguous NP as a direct object of the verb. Their eye-tracking study found an influence of thematic fit both on the ambiguous NP and at the disambiguation. Ambiguous NPs that made implausible direct objects were harder to read than plausible ones, and the disambiguation was harder to read after seeing a plausible ambiguous NP (that biases towards the ultimately incorrect object interpretation) than after seeing an implausible one.

Garnsey et al. (1997) varied the plausibility of the ambiguous NP as well as the subcategorization preference of the verb. They used verbs that prefer a sentential complement (SC verbs), verbs that prefer an NP argument (DO verbs) and verbs that are equibiased (EQ verbs, our development set). Sentences (5-a) and (5-b) are an example of DO and SC

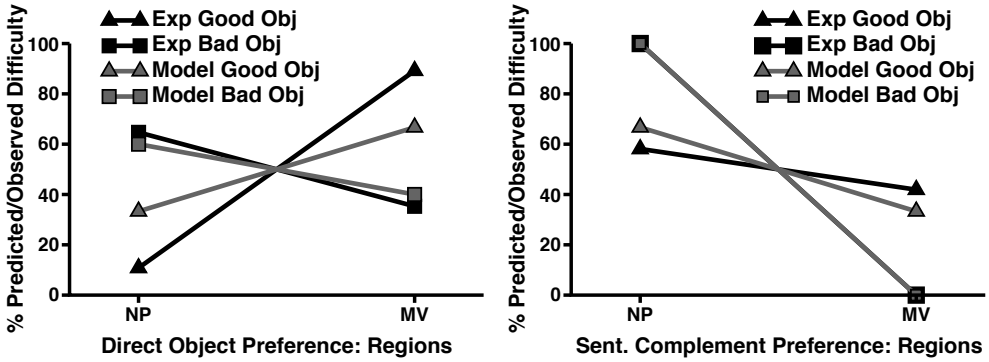

Figure 7. Garnsey et al. 1997: Experimental results and model predictions for the NP/S ambiguity. Left: Direct Object preference, right: Sentential Complement preference. Bad Obj: Bad NP object, Good Obj: Good NP object.

bias stimuli, which we model for evaluation.

(5) a. The director confirmed the rumor should have been stopped earlier. (Good object, DO-preferring verb)
b. The director confirmed the money should have been managed better. (Bad object, DO-preferring verb)
c. The agent admitted the mistake had been careless. (Good object, SC-preferring verb)
d. The agent admitted the airplane had been late taking off. (Bad object, SC-preferring verb)

Garnsey et al.'s eye tracking study found no significant effect of plausibility on SC-biased verbs for the total time measure we model, but there was some indication of difficulty when participants read the disambiguation region in the DO condition for stimuli with plausible object NPs. These NPs initially support the direct object hypothesis which is contradicted at the disambiguation.

Qualitative Analysis Fig. 7 shows our model's predictions for Garnsey et al.'s direct object and sentential complement conditions. For the direct object preference condition (on the left), our model predicts that stimuli with NPs that are implausible direct objects should be hard to process at the NP, but much easier at the main verb, which shows them not to be direct objects of the first verb at all. Inversely, good direct object stimuli should be easy to process at the NP, but harder at the disambiguation.

For the sentential complement condition (Fig. 7, right), the SynSem-Integration model predicts a similar interaction, with an especially extreme distribution of difficulty for the implausible object NPs. For both conditions, the observations follow a very similar pattern to the predictions.

The baseline model's predictions are shown in Fig. 8. They verify again that this model lacks sufficient thematic fit information: The baseline model predicts no difference between the conditions for the direct object bias verbs, and the small predicted difference

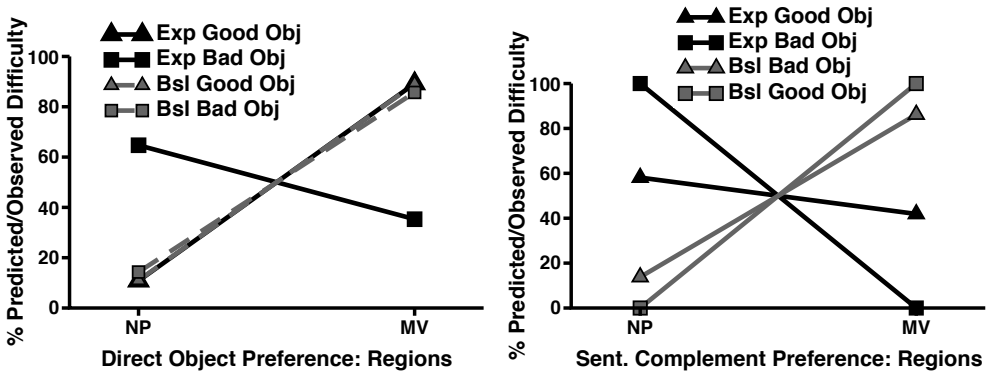

Figure 8. Garnsey et al. 1997: Experimental results and baseline predictions for the NP/S ambiguity. Left: Direct Object preference, right: Sentential Complement preference. Bad Obj: Bad NP object, Good Obj: Good NP object.

for the sentential complement verbs trends in the wrong direction.

3.7. The NP/0 ambiguity

The NP/0 ambiguity also centers around the interpretation of an ambiguous NP. This NP can either serve as a direct object to a verb in an adverbial clause which precedes a main clause, as in (6-a) (the *NP* alternative), or as the subject of the main clause, as in (6-b), where it stands in no relation to the verb in the adverbial clause (the *0* case, from Pickering & Traxler, 1998).

(6) a. While the woman was editing the magazine it started to rain.
b. While the woman was editing the magazine amused the reporters.

When processing this ambiguity, readers usually interpret the ambiguous NP as the direct object of the verb and show difficulty when it is disambiguated towards being the subject of the main clause.

Pickering and Traxler (1998) manipulated the thematic fit of the ambiguous NP as a direct object of the verb. Their eye-tracking study found a clear influence of thematic fit. For the total reading time measure, significant effects were found both on the ambiguous NP and at the disambiguation, such that implausible ambiguous NPs were harder to read than plausible ones, but caused less processing difficulty than plausible NPs at the disambiguation towards the 0 alternative.

Pickering et al. (2000) investigated the case of optionally transitive verbs with a strong intransitive bias in addition to manipulating thematic fit, using stimuli like (7-a) and (7-b).

(7) a. While the pilot was flying the plane stood over by the fence.
b. While the pilot was flying the horse stood over by the fence.

The total time findings for each region from their eye-tracking study show that reading time was longer on the NP for implausible object stimuli, while on the verb, reading time was shorter for these stimuli.

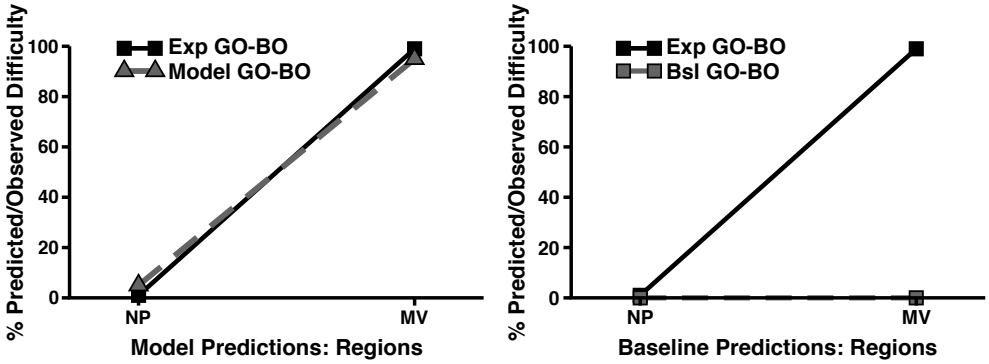

Figure 9. Pickering, Traxler and Crocker 2000: Left: Experimental results and model predictions for the NP/0 ambiguity. Right: Experimental results and baseline predictions. Bad object reading times minus Good object reading times.

Qualitative Analysis We present results for the Pickering et al. (2000) study. Since there is no way to construct a syntactically unambiguous control condition for the NP/0 ambiguity, Pickering et al. compare the reading times for the good object conditions to the reading times for the bad object conditions. The plots of observed and predicted difficulty in Fig. 9 therefore represent the relative difficulty of good objects as opposed to bad objects. They show the scaled difference between the reading times for good object sentences from the reading times for bad object sentences.

Our model correctly predicts that good objects should be easy to read in comparison to bad objects at the ambiguous NP (Fig. 9, left-hand side), and that bad objects in contrast should be hard to read in comparison with good objects at the disambiguation. The syntactic baseline again predicts no difference in difficulty between the semantic conditions (Fig. 9, right-hand side). This manifests as a straight line on the abscissa on the right-hand graph in Fig. 9.

3.8. The PP-attachment ambiguity

A PP-Attachment ambiguity usually arises in utterances like (8-a) and (8-b) from Rayner et al. (1983), where the attachment of the prepositional phrase *with binoculars* or *with a revolver* is possible both to the main verb (*see with binoculars*) and to the object NP (*cop with binoculars*).

(8) a. The spy saw the cop with binoculars.
b. The spy saw the crook with a revolver.

The PP-Attachment ambiguity is syntactically a global ambiguity: There is no way of unambiguously specifying the attachment site. However, semantic plausibility disambiguates the attachment of *with a revolver* to *the crook* in (8-b) and makes the attachment of *with binoculars* to *see* vastly more plausible than to *cop*. This means has been used to investigate the preferred initial attachment in the processing of this ambiguity.

Rayner et al. (1983) assumed that the verb attachment alternative is the syntactically simpler one and, following the parsing principle of *Minimal Attachment* (Frazier, 1978),

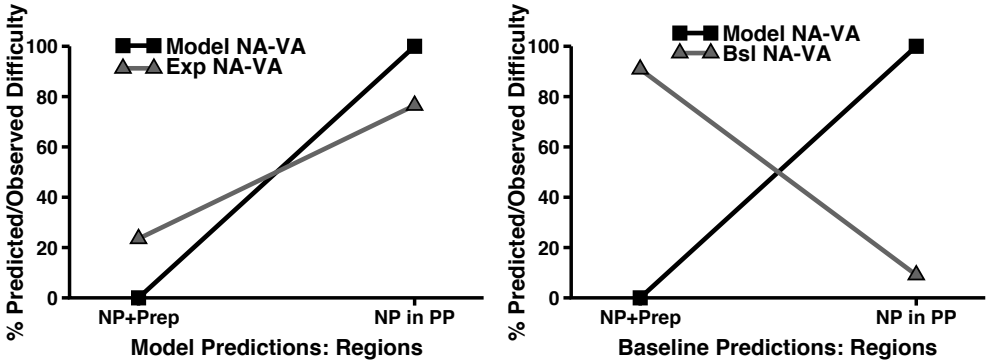

Figure 10. Rayner and Frazier 1983: Left: Experimental results and model predictions for the PP-Attachment ambiguity. Right: Experimental results and baseline predictions. Noun attachment reading times minus verb attachment reading times.

hypothesized a global attachment preference to the verb. The total reading time measure recorded in their eye tracking study indeed shows that readers took longer to read the noun in the PP if it was biased towards NP attachment rather than verb attachment.

Taraban and McClelland (1988) assumed the existence of a verb-specific attachment bias rather than a global parsing principle. They identified a verb bias for PP attachment in the Rayner et al. stimuli, and added an equal number of stimuli with verbs biased against PP attachment. We modeled self-paced reading times from Experiment 1A, where the findings from Rayner et al. were replicated for their stimuli, while the new Taraban and McClelland items showed the opposite pattern, supporting the assumption that attachment preferences are verb-specific.

We present results for the Rayner et al. (1983) study because the Taraban and McClelland (1988) study yields only two data points in a single region.

Qualitative Analysis Fig. 10 shows modeling results for the Rayner et al. (1983) study. Since no syntactically unambiguous controls can be constructed for the PP-Attachment ambiguity, we again use the difference between the attachment conditions as an indication of relative difficulty with the conditions. The plots in Fig. 10 show the scaled difference between predicted or observed difficulty in the NP attachment condition and predicted or observed difficulty in the verb attachment condition.

Rayner et al. (1983) measured reading difficulty in two regions: On the NP+preposition (*the crook with*), and on the NP that completes the prepositional phrase (*a revolver*). The SynSem-Integration model predicts that there should be little difference in difficulty between the conditions on the NP+preposition material that is identical in both conditions. At the noun in the PP, the model predicts that the NP attachment condition should cause more difficulty than the verb attachment condition, as indicated by the positive direction of the plotted predictions. The SynSem-Integration model's predictions correspond almost exactly to the pattern found in the data.

The syntactic baseline model predicts that when the NP within in the PP is read, NP attachment will be much easier than verb attachment, leading to a large negative difference in difficulty. This prediction is due to chance noise: The parser only predicts difficulty for

Table 6: Correlations between model predictions and observations (Spearman's ρ).

	All Data			No Garnsey et al.		
Model	N	ρ^a		N	ρ^a	
Baseline	36	-0.246,	ns	28	-0.276,	ns
Rank/If-Worse	36	0.714,	***	28	0.704,	***
Fixed/If-Worse	36	0.743,	***	28	0.694,	***
Ratio/Ratio	36	0.551,	**	28	0.412,	*

[a]ns: not significant, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

a single stimulus in a single region.

3.9. Quantitative evaluation

Our quantitative evaluation of the SynSem-Integration model was carried out against observations from the above-mentioned eight reading-time studies investigating four phenomena. The model's predictions were computed as described in Section 3.5. Evaluation was done by correlation analysis (Spearman's ρ) between the predicted and observed data points for each study. Table 6 shows an analysis across the pooled data from all modeled studies. We present the baseline results and the performance of the Rank/If-Worse model, which uses the best parametrization on the development set, as well as the other two well-performing parametrizations to demonstrate the model's robustness across parametrizations.

The correlation analysis is significant with a correlation coefficient of about 0.7 for the Rank/If-Worse model. The coarser-grained Fixed/If-Worse cost functions even do slightly better than this, while the finer-grained Ratio/Ratio cost functions prove to be very sensitive to the noise inherent in our probabilistic models at a correlation coefficient of $\rho = 0.551$. In contrast, the syntactic baseline model does not achieve a significant correlation with the observed data.

One reservation about the pooled analysis as a measure of the model's general performance might be that it includes the two NP/S data sets from Garnsey et al., the study that furnished the development set. One might argue that optimizing on one data subset from a study makes it likely that the other data subsets from this study will also be optimized indirectly. The right section of Table 6 presents the correlation results for the overall analysis without using the Garnsey et al. data sets. The difference in correlation coefficients is not statistically significant for any of the models (all $p > 0.4$, two-tailed, using Raghunathan's (2003) test which allows for missing values).

3.9.1. Discussion.

The quantitative and qualitative analyses of the SynSem-Integration model's predictions have demonstrated its reliability. The model clearly outperformed a lexicalized syntax-only model, which, presumably due to sparse data problems, failed to predict the influence of thematic fit on human sentence processing. This result highlights the importance of the explicit, independently motivated model of semantic plausibility employed in the SynSem-Integration model.

The SynSem-Integration model is able to predict the patterns of human processing difficulty for four well-studied phenomena with unchanged parameter settings and without per-phenomenon adaptations. The SynSem-Integration model completely eliminates the problem of hand-selecting and hand-setting constraints for individual phenomena. Its component models, especially the syntactic model, account for a large amount of constraints typically used in constraint-integration models, for example word form or sentence structure preferences. This information is incorporated in a single comprehensive model of lexical and syntactic frequencies that is trained once on a single data set. This model has the advantage of being general enough to contain the relevant information for a large number of phenomena. At the same time, it ensures that no potentially important preference information is neglected.

The quantitative evaluation of three different combinations of cost functions has demonstrated the SynSem-Integration model's robustness given per-item predictions of different grain size. In the face of noise in the model and the data, the least fine-grained cost functions performed best. Importantly, all three variants of the model reliably predict patterns of processing difficulty, and clearly outperform the baseline model.

4. General discussion

We have presented the SynSem-Integration model of human sentence processing. This model extends the standard probabilistic grammar-based account of syntactic processing with a model of human thematic plausibility intuitions. The model is therefore able to account for syntactic *and* semantic effects in human sentence processing, while retaining the main advantages of probabilistic grammar-based models, namely their ability to naturally account for frequency effects and their wide coverage of syntactic phenomena and unseen input. The model is to a large extent derived automatically from training data, which obviates the need for experimenter intervention and grounds the model's predictions in naturalistic language data. This is an advantage of our model over constraint-based accounts, where the set of relevant constraints has to be specified by hand for each new phenomenon to be modeled. Note that a large number of constraints used in constraint-based accounts, such as structural and lexical preferences, are covered by the probabilistic grammar in the syntactic model in a unified and homogeneous way. Further, the SynSem-Integration model is the first to employ a model of human plausibility intuitions (instantiated as verb-argument thematic fit), which allows wide coverage of unseen input.

Our evaluation has shown that both the plausibility model that we have proposed and the SynSem-Integration model reliably predict human data. The plausibility model predicts human verb-argument-role plausibility judgments, showing wide coverage of unseen verb-argument-role triples and reliable predictions for both seen and unseen data points. The SynSem-Integration model's predictions have been evaluated against results from eight experimental studies and across four ambiguity phenomena. We have presented qualitative results for each phenomenon and have shown that the model's predictions are significantly correlated with observed human processing difficulty across all phenomena. This demonstrates the model's generality and robustness.

We now turn to discussing the theoretical implications of our model's implementation. The model consists of a syntactic and a semantic model, which co-operate to determine a globally preferred analysis of the input. The semantic model is assumed to operate on the

analyses created by the syntactic model. This modular architecture is an implementational choice, and we do not make any specific claims with regard to its cognitive plausibility. Note especially that our model is not a syntax-first approach, as it does not assume a temporal disjunction between purely syntactic, lexical and semantic processing: The syntactic component immediately integrates lexically-specific information (e.g., verb subcategorization and word class preferences) and the semantic model processes and ranks the input within the same time step as the syntactic model.

A second point concerns the implementation of the semantic plausibility model. We have demonstrated that a probabilistic model enhanced with knowledge about semantic generalizations can predict human semantic judgments. This model relies on descriptions of events in a corpus to assess event plausibility. Human beings learn a lot about event plausibility by observation, and not necessarily in verbal contexts. Hence using language data to model plausibility is an indirect route. We use it in the absence of any other kind of training data for event plausibilities, and with the additional justification that there is plausibility knowledge that is learned through the medium of language. For example, many people would confirm that wizards are plausible agents of a jinxing event, even though it is unlikely that they have directly witnessed such an event.

We operate on the assumption that there is a link between the plausibility of the (partial) event denoted by a verb-argument-role triple and the frequency with which it is encountered in a corpus. Of course, we cannot assume that corpus-based plausibility estimates will be perfect, because humans usually make utterances with the goal of communicating information to a hearer. Corpus frequencies may be distorted for example if commonplace events are not deemed worthy of explicit discussion, or if infrequent events are perceived as more informative or interesting, and therefore are discussed more often than they are experienced. In addition, data sparseness often turns a frequency estimate into a seen-unseen classification in practice. However, we observed that verb-argument-relation triples encountered in corpora were rated as significantly more plausible than unseen triples in a previous norming study (Padó, 2007), indicating that events described in a corpus are generally plausible. Our class-based smoothing approach attempts to distinguish between events that are unseen, yet plausible, and those that are unseen and implausible. We take the performance of our implemented semantic model as an indication that corpus data yields sufficient information about verb-argument-role plausibility for successful modeling.

Finally, the parameter setting process for the SynSem-Integration model yielded two interesting observations. Both are relevant to the prediction of Revision cost. The first is that the model's performance improves as the influence of the syntactic ranking on the global ranking grows stronger. Which model dominates the global ranking does not influence the Conflict cost function, as it only registers disagreement between the two models. However, the predictions of the Revision cost function depend on which analysis is preferred initially. If the preferred analysis is determined by syntactic preferences, the SynSem-Integration model makes correct predictions about difficulty due to Revision. If the preferred analysis is determined by semantic plausibility, the model's predictions do not match the observed difficulty. This appears to imply that plausibility information modulates, but does not strongly determine, the preferred syntactic structure in processing. Studies investigating the influence of thematic fit on parsing indeed regularly find that thematic fit information weakens, but does not eliminate Revision effects at disambiguation (e.g., Ferreira & Clifton,

1986; McRae et al., 1998; Clifton et al., 2003).

The weakening of Revision difficulty due to thematic fit information is implemented in the SynSem-Integration model by the If-Worse Revision cost function that only predicts difficulty when the new interpretation is less semantically plausible than the revised interpretation. This means that no difficulty is predicted on the item level if the change of preferred interpretation makes semantic sense. Our exploration of the parameter space showed that only models using this cost function were able to predict the correct pattern of difficulty in the experimental data. Note that the preference for this cost function does not mean that the model assigns no difficulty at all in a condition with a semantic bias towards the disambiguated reading. Due to noise in the items and in the semantic model, this cost function results in a reduced, but not a zero difficulty prediction.

Taken together, the cost functions of the SynSem-Integration model thus predict a situation in which semantic information is used to continually (and simultaneously) evaluate syntactic decisions, but in which it does not immediately determine the syntactic analysis of the input that the processor entertains. This description is realistic given empirical findings of both semantic effects during the processing of an ambiguity (Trueswell, Tanenhaus, & Garnsey, 1994; McRae et al., 1998) and the observation that thematic fit does not necessarily suffice to cancel out Revision effects at disambiguation.

One possible limitation to our model is the combination of two cost functions for difficulty prediction, where competition-based models such as Spivey-Knowlton's (1996) use only one. This is due to our decision to extend Jurafsky-style probabilistic grammar-based models, which, unlike constraint-based models, explain the construction of syntactic analyses as well as the resolution of ambiguities. The difficulty prediction mechanism in these models covers only Revision situations and cannot be easily adapted to also account for Conflict situations. Similarly, the cost prediction mechanism from competition-based models does not completely carry over to probabilistic grammar-based models. Note that while we propose two cost functions, both ultimately compute the support that the globally preferred parse has from previous linguistic experience (the component models) and assumptions based on earlier processing stages.

We have presented a wide-coverage probabilistic model of thematic role assignment and plausibility which is transparently integrated with a probabilistic lexico-syntactic processor. While this model is able to account for a range of relevant judgment and reading time data relating to semantic plausibility, there remain of course many dimensions of semantic processing to be modeled. These include the role of discourse context for the resolution of ambiguous references (e.g., Altmann & Steedman, 1988; Spivey & Tanenhaus, 1998), the accommodation of definite versus indefinite NPs (Crain & Steedman, 1985; Spivey-Knowlton & Sedivy, 1995), and the resolution of quantifier scope (Kurtzmann & MacDonald, 1993). We leave it to future work to extend the model to further semantic phenomena, and explore the scalability of the architecture.

References

Abney, S. (1989). A computational model of human parsing. *Journal of Psycholinguistic Research*, *18*(1), 129–144.

Aitchison, J. (2003). *Words in the mind: An introduction to the mental lexicon* (third ed.). Oxford and New York: Basil Blackwell.

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*, 191–238.

Baker, C., Fillmore, C., & Lowe, J. (1998). The Berkeley FrameNet project. In *Proceedings of the joint international conference on computational linguistics and annual meeting of the association for computational linguistics (COLING/ACL)* (pp. 86–90). East Stroudsburg, PA: Association for Computational Linguistics.

Bikel, D. (2004). Intricacies of Collins' parsing model. *Computational Linguistics*, *30*(4), 479–511.

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of parsing difficulty: An evaluation using the Potsdam sentence corpus. *Journal of Eye Movement Research*. (To appear.)

Burnard, L. (1995). User's guide for the British National Corpus [Computer software manual]. Oxford.

Carlson, G. (1984). Thematic roles and their role in semantic interpretation. *Linguistics*, *22*, 259–279.

Carlson, G., & Tanenhaus, M. (1988). Thematic roles and language comprehension. In W. Wilkins (Ed.), *Thematic relations* (Vol. 21). New York: Academic Press.

Carreras, X., & Márquez, L. (2005). Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the conference on computational natural language learning (CoNLL)* (pp. 152–164). East Stroudsburg, PA: Association for Computational Linguistics.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Science*, *10*(7), 335–344.

Clark, S., & Weir, D. (2002). Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, *28*(2), 187–206.

Clifton, C., Traxler, M., Mohamed, M. T., Williams, R., Morris, R., & Rayner, K. (2003). The use of thematic role information in parsing: Syntactic autonomy revisited. *Journal of Memory and Language*, *49*, 317–334.

Collins, M. (1996). A new statistical parser based on bigram lexical dependencies. In *Proceedings of the annual meeting of the association for computational linguistics (ACL)* (pp. 184–191). East Stroudsburg, PA: Association for Computational Linguistics.

Crain, S., & Steedman, M. (1985). On not being led up the garden path: The use of context by the psychological syntax processor. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language parsing* (pp. 320–358). Cambridge, UK: Cambridge University Press.

Crocker, M. W. (1996). *Computational psycholinguistics: An interdisciplinary approach to the study of language*. Dordrecht: Kluwer Academic Publishers.

Crocker, M. W. (2005). Rational models of comprehension: Addressing the performance paradox. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 363–380). London: Lawrence Erlbaum Associates.

Crocker, M. W., & Brants, T. (2000). Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, *29*(6), 647–669.

Crocker, M. W., & Corley, S. (2002). Modular architectures and statistical mechanisms: The case from lexical category disambiguation. In P. Merlo & S. Stevenson (Eds.), *The lexical basis of sentence processing*. Amsterdam: John Benjamins.

Cuetos, F., Mitchell, D., & Corley, M. (1996). Parsing in different languages. In M. Carreiras, J. García-Albea, & N. Sebastían-Gallés (Eds.), *Language processing in Spanish* (pp. 156–187). Hillsdale, NJ: Lawrence Erlbaum.

Dagan, I., Pereira, F., & Lee, L. (1994). Similarity-based estimation of word cooccurrence probabilities. In *Proceedings of the annual meeting of the association for computational linguistics (ACL)* (pp. 272–278). East Stroudsburg, PA: Association for Computational Linguistics.

Demberg, V., & Keller, F. (2008). Data from eye tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*. (To appear.)

Fellbaum, C. (Ed.). (1998). *Wordnet – An electronic lexical database*. Cambridge, MA: MIT Press.

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*, 348–368.

Fillmore, C. (1982). Frame semantics. In *Linguistics in the morning calm* (pp. 111–137). Seoul, South Korea: Hanshin Publishing Co.

Fillmore, C., Johnson, C., & Petruck, M. (2003). Background to FrameNet. *International Journal of Lexicography*, *16*, 235–250.

Frazier, L. (1978). *On comprehending sentences: Syntactic parsing strategies.* Bloomington, IN: Indiana University Linguistics Club.

Garnsey, S., Pearlmutter, N., Myers, E., & Lotocky, M. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, *37*, 58–93.

Gedeon, T., Parker, A., & Dimitrov, A. (2003). Information distortion and neural coding. *Canadian Applied Mathematics Quarterly*, *10*(1), 33–70.

Gibson, T. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown.* Doctoral dissertation, Carnegie Mellon University. (UMI: AAT 9126944)

Gildea, D. (2001). Corpus variation and parser performance. In *Proceedings of the conference on empirical methods in natural language Processing (EMNLP)* (pp. 167–202). East Stroudsburg, PA: Association for Computational Linguistics: SIGDAT.

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

Good, I. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*, 237–264.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the meeting of the North American chapter of the association for computational linguistics (NAACL)* (pp. 168–196). East Stroudsburg, PA: Association for Computational Linguistics.

Hindle, D., & Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, *19*, 103 – 120.

Jelinek, F., Laerty, J., Magerman, D., & Roukos, S. (1994). Decision tree parsing using a hidden derivation model. In *Proceedings of the 1994 human language technology workshop* (pp. 272–277). San Francisco, CA: Morgan Kaufmann.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*, 137–194.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics.* Cambridge, MA: MIT Press.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329–354.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *3*(2), 228-238.

Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *35*(3), 400–401.

Kipper, K., Dang, H. T., & Palmer, M. (2000). Class-based construction of a verb lexicon. In *Proceedings of the national conference on artificial intelligence (AAAI)* (pp. 691 – 696). Cambridge, MA: AAAI Press / The MIT Press.

Kurtzmann, H., & MacDonald, M. (1993). Resolution of quantifier scope ambiguities. *Cognition*, *48*, 243–279.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 1–45.

Litkowski, K. (2004). Senseval-3 task: Automatic labeling of semantic roles. In *Proceedings of Senseval-3: The third international workshop on the evaluation of systems for the semantic analysis of text* (pp. 9–12). East Stroudsburg, PA: Association for Computational Linguistics: SIGLEX.

MacDonald, M. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, *9*(2), 157–201.

Manning, C., & Schütze, H. (1999). *Foundations of statistical language processing*. Cambridge, MA: MIT Press.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Marx, Z. (2004). *Structure-based computational aspects of similarity and analogy in natural language*. Doctoral dissertation, Hebrew University, Jerusalem.

Mayberry, M. R. (2003). *Incremental nonmonotonic parsing through semantic self-organization*. Doctoral dissertation, University of Texas at Austin. (UMI: AAT 3116385)

McDonald, S., & Shillcock, R. (2003). Eye movements reveal the on-line computation of lexical probabilities. *Psychological Science*, *14*, 648—652.

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

Medin, D. L., & Aguilar, C. (1999). Categorization. In R. A. Wilson & F. C. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences* (pp. 104–105). Cambridge, MA: MIT Press.

Narayanan, S., & Jurafsky, D. (2002). A Bayesian model predicts human parse preference and reading time in sentence processing. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14* (pp. 59–65). Cambridge, MA: MIT Press.

Narayanan, S., & Jurafsky, D. (2005). *A Bayesian model of human sentence processing.* (MS, available at: `http://www.icsi.berkeley.edu/~snarayan/newcog.pdf`. Accessed February 2006)

Padó, U. (2007). *The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing*. Doctoral dissertation, Saarland University. (URN: urn:nbn:de:bsz:291-scidok-11381)

Padó, U., Crocker, M. W., & Keller, F. (2006). Modelling semantic role plausibility in human sentence processing. In *Proceedings of the meeting of the European chapter of the association for computational lingustics (EACL)* (pp. 345–352). East Stroudsburg, PA: Association for Computational Linguistics.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*(1), 71–105.

Parsons, T. (1990). *Events in the semantics of English: A study in subatomic semantics*. Cambridge, MA: MIT Press.

Pickering, M., & Traxler, M. (1998). Plausibility and recovery from garden paths. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24*.

Pickering, M., Traxler, M., & Crocker, M. W. (2000). Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. *Journal of Memory and Language*, *43*, 447–475.

Pritchett, B. (1992). *Grammatical competence and parsing performance*. Chicago, IL: The University of Chicago Press.

Raghunathan, T. (2003). An approximate test for homogeneity of correlated correlations. *Quality and Quantity*, *37*, 99-110.

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behaviour*, *22*, 358–374.

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

Roark, B. (2001). *Robust probabilistic predictive syntactic processing: Motivations, models, and applications*. Doctoral dissertation, Brown University. (UMI: AAT 3006783)

Rohde, D. (2002). *A connectionist model of sentence comprehension and production*. Doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA. (UMI: AAT 3051010)

Spivey, M., & Tanenhaus, M. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24*(6), 1521–1543.

Spivey-Knowlton, M. (1996). *Integration of visual and linguistic information: Human data and model simulations*. Doctoral dissertation, University of Rochester. (UMI: AAT 9074332)

Spivey-Knowlton, M., & Sedivy, J. (1995). Parsing attachment ambiguities with multiple constraints. *Cognition*, *55*, 227–267.

Stevenson, S. (1994). Competition and recency in a hybrid network model of syntactic disambiguation. *Journal of Psycholinguistic Research*, *23*(4), 295–322.

Stowe, L. (1989). Thematic structures and sentence comprehension. In G. Carlson & M. Tanenhaus (Eds.), *Linguistic structure in language processing* (pp. 319–357). Dordrecht; Boston: Kluwer Academic Publishers.

Surdeanu, M., Harabagiu, S., Williams, J., & Aarseth, P. (2003). Using predicate-argument structures for information extraction. In *Proceedings of the annual meeting of the association for computational linguistics (ACL)* (p. 9-16). East Stroudsburg, PA: Association for Computational Linguistics.

Taraban, R., & McClelland, J. (1988). Constituent attachment and thematic role assignment in sentence processing: Influences of content-based expectations. *Journal of Memory and Language*, *27*, 597–632.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The Information Bottleneck method. In *Proceedings of the annual Allerton conference on communication, control and computing* (pp. 368–377). Urbana-Champaign, IL: University of Illinois.

Trueswell, J. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, *35*, 566–585.

Trueswell, J., Tanenhaus, M., & Garnsey, S. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, *33*, 285–318.

Trueswell, J., Tanenhaus, M., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*(3), 528–553.

Vosse, T., & Kempen, G. (2000). Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and a lexicalist grammar. *Cognition*, *75*, 105–143.

Xue, N., & Palmer, M. (2004). Calibrating features for semantic role labeling. In *Proceedings of the joint human language technology conference and conference on empirical methods in natural language processing (HLT/EMNLP)* (pp. 88–94). East Stroudsburg, PA: Association for Computational Linguistics.