Combining Syntax and Thematic Fit in a Probabilistic Model of Sentence Processing

Ulrike Padó (ulrike@coli.uni-sb.de)
Computational Linguistics
Saarland University, 66041 Saarbrücken

Frank Keller (keller@inf.ed.ac.uk)
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, UK

Matthew Crocker (crocker@coli.uni-sb.de)
Computational Linguistics
Saarland University, 66041 Saarbrücken

Abstract

We present a model of human sentence processing that extends a standard probabilistic grammar model with a semantic module which computes the thematic fit of verbs and arguments in a cognitively plausible way. Our model differs from existing probabilistic accounts (e.g., Jurafsky, 1996) by capturing both syntactic and semantic influences in human sentence processing. It also overcomes limitations of constraint-based models (Spivey-Knowlton, 1996; Narayanan and Jurafsky, 2002), as its parameters can be acquired automatically from corpus data, and no hand-coding of constraints is required. We evaluate our semantic module against human ratings of thematic fit, and also test the complete model's performance for two well-studied ambiguities from the sentence processing literature.

Introduction

In the investigation of human sentence processing, the central importance of frequency is a recurring theme. A wide range of frequencies seem to be used by the human sentence processor, including verb frame frequencies (e.g., Garnsey et al., 1997), frequencies of morphological forms (e.g., Trueswell, 1996), lexical category frequencies (e.g., Crocker and Corley, 2002) and structural frequencies (e.g., Brysbaert and Mitchell, 1996). These findings have led to the formulation of a range of probabilistic models that account for frequency effects in human sentence processing.

In this paper, we review two approaches that have been put forward in the area of sentence processing: Models based on probabilistic grammars (Jurafsky, 1996; Crocker and Brants, 2000) and models which integrate a number of constraints into an activation-based model (Spivey-Knowlton, 1996) or a Bayesian belief net (Narayanan and Jurafsky, 2002).

Both approaches have drawbacks. Models based on probabilistic grammars are not designed to take into account the influence of semantic processing. Constraint-based models, in contrast, require both manual specification of a different set of constraints for each phenomenon and the compilation of parameters from many, often heterogeneous, sources.

In this paper, we introduce a new model of sentence processing which extends probabilistic grammar-based models. We add a semantic component which evaluates the plausibility of each structure the parser generates, on the basis of the thematic fit between a verb and its arguments. The model identifies one syntactically most likely and one semantically most plausible structure. If syntactic and semantic preferences conflict, we predict processing difficulty, as reanalysis is required. Both the syntactic and the semantic component of our model are automatically trained on annotated corpus data and require no hand-tuning of constraints. We successfully model the influence of thematic fit on processing the Main Clause/Reduced Relative (MC/RR) ambiguity (which cannot be modelled by standard probabilistic grammar models) and the NP/Sentence Complement (NP/S) ambiguity.

We begin by introducing probabilistic grammar models and demonstrate how they fail to capture semantic processing effects in the MC/RR ambiguity. We then review two alternative constraint-based models that account successfully for this ambiguity by integrating syntactic and semantic information from different sources. However, methodological and practical problems exist with these approaches. We then introduce our own model in detail, focusing on our main innovation, the semantic module, and evaluate its cognitive plausibility. Finally, we demonstrate that our model correctly captures processing data for the MC/RR and the NP/S ambiguity.

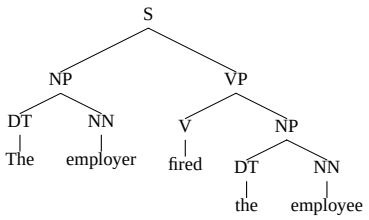
Previous Models

Probabilistic Parsers

A standard account of frequency effects in language processing is provided by models based on probabilistic grammars (Jurafsky, 1996; Crocker and Brants, 2000). Typically, these models use a Probabilistic Context-Free Grammar (PCFG) to compute the probability of each possible structure given the input sentence. A PCFG consists of a set of context-free rules which define the daughter nodes licensed by a mother node in a phrase structure tree. Each rule is annotated with a probability which represents the likelihood of expanding the mother category into the daughter categories. The probability of a syntactic structure (parse tree) T is defined as the product of the probabilities of all rules applied in generating T. An example PCFG is given in Fig. 1. This grammar can be used to generate the parse tree shown underneath, with probability $P(T) = .105$. Algorithms exist to efficiently compute PCFG parse probabilities on a word-by-word basis (Stolcke, 1995).

In order to incrementally predict processing difficulty, complexity measures can be defined based on the probabilities of all parses licensed by a PCFG. The *Surprisal* account

S	→	NP VP	1.0
NP	→	DT NN	1.0
VP	→	V NP	.6
VP	→	V	.4
V	→	fired	.7
V	→	left	.3
DT	→	the	1.0
NN	→	employer	.5
NN	→	employee	.5



$P(T) = 1.0 \cdot 1.0 \cdot 1.0 \cdot .5 \cdot .6 \cdot .7 \cdot 1.0 \cdot 1.0 \cdot .5 = .105$

Figure 1: Example of a PCFG and one tree it generates

(Hale, 2001) monitors the incremental changes in the probability distribution over all parses to predict cognitive load, assuming fully parallel processing and making no predictions about preferred structures. Alternatively, PCFG probabilities can be used to rank the parses (*Ranking approach*), assuming that the most likely structure is the one preferred by humans. Processing difficulty is linked specifically to the processing effort made when a previously preferred analysis suddenly becomes dispreferred (Jurafsky, 1996; Crocker and Brants, 2000). Human memory limitations are modelled by a *search beam* containing only the most likely analyses, since the number of possible structures rises with the size of the grammar.

PCFG-based models account elegantly for frequency effects in lexical category and morphological form ambiguities through probabilistic lexicon entries, while structural preferences are covered by probabilistic grammar rules.

These models also account for processing failure in difficult garden path sentences. In the famous example, *the horse raced past the barn fell*, the ultimately correct reduced relative analysis corresponding to *the horse that was raced past the barn* is assigned only a small probability because it is infrequent overall and *raced* is biased towards the intransitive, active interpretation. In the Ranking approach, the analysis drops out of the beam of accessible parses and cannot be retrieved any more when *fell* is encountered, which causes parsing to fail for this sentence. Alternatively, the Surprisal approach uses the fact that the very likely main clause parse becomes impossible at *fell* to correctly predict difficulty.

The most important difference among PCFG-based models is the way the grammar is induced. The Jurafsky (1996) model uses a set of hand-selected rules, with probabilities extracted from a corpus of structurally annotated sentences. The Crocker and Brants (2000) model, on the other hand, induced both grammar rules and probabilities from a corpus. This gives their model broad coverage of syntactic constructions and allows it to correctly analyse unseen text, which is an important characteristic of human sentence processing.

Restriction to Syntax

While PCFG-based models have been shown to account for a variety of syntactic phenomena in sentence processing, they suffer from a major restriction: They are unable to take semantic information into account. Consider sentence (1):

Input	Structure	Flip
The	MC	no
employer/employee	MC	no
fired	MC	no
by	RR	yes

Figure 2: Preferred structure as predicted incrementally by a PCFG-based model for an item from McRae et al. (1998).

(1) *The employer fired by the owner was jobless*

At the verb *fired*, both a main clause continuation (e.g. as *the employer fired the employee*) and the ultimately correct reduced relative continuation are possible.

Fig. 2 shows the structural predictions made for this sentence by an incremental PCFG-based parser (Roark, 2001). The parser's (unlexicalised) grammar and lexicon are derived from sections 2-21+24 of the Wall Street Journal corpus (Marcus et al., 1994). The parser predicts the main clause (MC) structure at *fired*, and then switches to the reduced relative (RR) structure at *by*. This "flip" in preferred structures predicts processing difficulty at *by*. The parser makes the same structural predictions for both *employer* and *employee*.

However, the results of reading time studies by McRae et al. (1998) and Trueswell et al. (1994) demonstrate that readers use the thematic fit of the first NP and verb in processing this ambiguity. When the first NP is a plausible agent of *firing*, like *the employer*, readers prefer the main clause interpretation and show difficulty during the disambiguating *by*-phrase, as predicted by the parser. However, when the first NP is a bad agent, but a good patient of the verb, like *the employee*, readers reanalyse towards the reduced relative reading right away because it allows them to interpret the first NP as a patient of the verb. In this case, they show difficulty in the verb region, and not during the *by*-phrase.

In sum, unlike human readers, the PCFG-based model makes the wrong prediction for the good patient (*employee*) case because it does not take the semantics of the first NP into account. This general problem is common to all PCFG-based models, be they Ranking or Surprisal approaches.

Constraint Integration Models

To account for both syntactic and semantic influences in sentence processing, McRae et al. (1998) use a constraint-based model, the Competition-Integration model (Spivey-Knowlton, 1996). In contrast to the PCFG-based models, this model does not create the structural alternatives itself, but only decides between them. It uses weighted constraints which provide support for the analyses by activating them to a greater or lesser degree. The activation of each analysis is computed iteratively, and an analysis is chosen when its activation exceeds a threshold. If all constraints point towards the same analysis, the model needs few iterations to settle than if constraints conflict. The number of iterations until settling can be used to predict processing difficulty.

To model the MC/RR ambiguity, McRae et al. (1998) used the following constraints in the verb and *by*-region that were estimated from a variety of different sources: Thematic fit of first NP and verb (from a rating study), tense/voice preferences of the first verb (from a corpus study), a bias for the

reduced relative interpretation when reading *by* and a general bias for the main clause analysis over the reduced relative (from a corpus study). After disambiguation, two more constraints supported the relative clause interpretation. The weights for the constraints were set by fitting to off-line completion data. The resulting model successfully predicts human processing data for the MC/RR ambiguity.

A second constraint-integrating model that accounts for the data is described by Narayanan and Jurafsky (2002). It extends Jurafsky's original model by proposing a combination of Bayesian belief nets (a formalism for reasoning about events based on partial probabilistic information). The proposed architecture can incrementally integrate a parsing model with any number of constraints from other sources in a mathematically clean and consistent way. The parser is cast as a belief net which computes the syntactic probability of each parse, while a second belief net integrates thematic fit and lexical (verb tense/voice and valence) probabilities. The predictions of the nets are combined into a single probability value for each structure. Again, the most probable analysis is taken to be the preferred one, and flips predict difficulty.

However, there are two main drawbacks to both types of constraint-integrating model. First, a specific set of constraints has to be chosen for each ambiguity (e.g., Tanenhaus et al., 2000). Consequently, the models are unlikely to generalise to new constructions without further changes.

A second problem is that the constraint weights often are estimated from a diverse set of sources, for example various corpus studies as well as rating and completion studies (e.g., Narayanan and Jurafsky, 2002). This is at least inelegant and can be costly if rating studies have to be run. It may even be problematic if sources (e.g., corpora) with different or even conflicting biases are used (see Roland and Jurafsky, 1998).

A PCFG-Based Model with Semantics

We introduce a probabilistic processing model based on the Ranking approach that overcomes the limitations of PCFG-based models by integrating a semantic module. At the same time, our model does not require the stipulation of arbitrary semantic constraints, and its parameters need not be set by hand, but are learnt automatically from corpus data. Learning from corpora also gives our model broad coverage both of structures that are processed effortlessly as well as those that cause interesting disruption. This allows the model to cover different phenomena without requiring modifications.

Our model computes the plausibility of the verb-argument relations in each structure that the parser constructs and uses the plausibility score to complement the syntactic probability computed by the parser. When the role assignment that the semantic module prefers is incompatible with the preferred syntactic interpretation, we predict difficulty due to the processing effort made to solve the conflict.

We first introduce the semantic module in detail. We then test the cognitive plausibility of the semantic module's predictions, and finally review the complete model's handling of the MC/RR and NP/S ambiguities.

Semantic Module

The task of our semantic module is to compute the plausibility of a verb-argument relation in terms of thematic fit. Thematic fit is influenced (at least) by the verb (or, more specifically, its current sense), the argument head, the thematic role, and the grammatical function of the argument. We equate the plausibility of a verb-role-argument triple with its probability, which we compute as the joint probability of the verb's sense v_s[1], the role r, the argument head a, and the grammatical function gf of a:

$$Plausibility_{v,r,a} = P(v_s, r, a, gf)$$

This joint probability cannot be reliably estimated from co-occurrence counts due to lack of data. But we can decompose this term into a number of subterms that approximate intuitively important information such as syntactic subcategorisation ($P(gf|v_s)$), the syntactic realisation of a semantic role ($P(r|v_s, gf)$) and selectional preferences ($P(a|v_s, gf, r)$):

$$\begin{aligned} Plausibility_{v,r,a} &= P(v_s, r, a, gf) = \\ & P(v_s) \cdot P(gf|v_s) \cdot P(r|v_s, gf) \cdot P(a|v_s, gf, r) \end{aligned}$$

This formulation allows us to estimate each of the subterms from training data with semantic role annotation. However, we still need to smooth our estimates, especially as the counts needed for estimating the $P(a|v_s, gf, r)$ term remain sparse.

We use two complementary approaches to smoothing sparse training data. One, Good-Turing smoothing, approaches the problem of unseen data points by assigning them a small probability. This method relies on re-estimating the probability of seen and unseen events based on knowledge about more frequent events. We apply it to all estimates that are 0 or 1 to avoid not being able to make predictions at all.

The other method, class-based smoothing, attempts to arrive at semantic generalisations for words. These serve to identify equivalent verb-argument pairs that furnish additional counts for the estimation of $P(a|v_s, gf, r)$. For example, if {*boss, employer, chief*} forms a synonym group of nouns, class-based smoothing allows us to share counts for *boss-fire-agent* and *employer-fire-agent*. While little is known about the cognitive plausibility of smoothing, the kind of generalisations we use seem intuitively not far removed from human reasoning. We employ noun classes as well as verb classes. Our noun classes are the lowest class level from WordNet (Miller et al., 1990), the synonym sets. Our verb classes are induced from the training data by unsupervised soft clustering methods (see Padó et al., 2006). Soft clustering allows different verb senses to be distinguished by a verb's membership in different clusters.

Evaluation of the Semantic Module

We first establish that our semantic module reliably captures human intuitions. The model's task is to correctly predict human thematic fit judgements for verb-role-argument triples. We take a significant positive correlation of the predictions to the human judgements to indicate reliable performance.

Training and Test Data To train our model, we need language data with thematic role annotation. To date, there are two main efforts to semantically annotate corpora: PropBank (PB, Palmer et al., 2005) and FrameNet (FN, Baker et al.,

[1] Since the correct verb sense is unknown, we compute plausibility for all senses and choose the most plausible one.

fire.01 [The employer $_{Arg0}$] fired [the employee $_{Arg1}$]
Firing [The employer $_{Employer}$] fired [the employee $_{Employee}$]

Figure 3: Example annotation: PropBank (above) and FrameNet (below).

Verb	Noun	Role	Rating
fire	employer	agent	6.1
fire	employer	patient	2.4
fire	employee	agent	1.9
fire	employee	patient	6.4

Table 1: Test items: Verb-noun pairs with ratings for the agent and patient role using a 7 point scale (McRae et al., 1998).

2003). The PB corpus (c. 120,000 propositions, c. 3,000 verbs) adds semantic annotation to the Wall Street Journal corpus, the same data our parser is trained on. Arguments and adjuncts are annotated for every verbal proposition in the corpus. A common set of argument labels *Arg0* to *Arg5* and *ArgM* (for adjuncts) is used, and interpreted in a verb-specific way. Some consistency in mapping has been achieved, so that agents are generally *Arg0* and patients/themes *Arg1*.

The FN corpus groups verbs with similar meanings together into frames (i.e., descriptions of situations), and assumes a set of frame-specific roles for the participants (e.g., an *Employer* and *Employee* in the Firing frame). Fig. 3 gives an example of PB and FN style annotation. The FN resource is about half as large as PB at 57,000 propositions (c. 2,000 verbs). Since corpus annotation is frame-driven, only some senses of a verb may be present and word frequencies in the FN corpus may not be representative of English. The PB approach annotates running text, which makes it more reliable in this respect. However, both the definition of frames as semantic verb classes and the semantic characterisation of frame-specific roles introduce information to FN annotation that is not present in PB. We train on both corpora and compare the results below.

Our test data consists of two sets of verb-argument pairs with plausibility ratings for two roles each. For both sets, raters answered questions like *How common is it for an employer to fire someone?* with a rating from 1 (*very uncommon*) to 7 (*very common*), as for the example item in Table 1. One test set comprises 100 out of the 160 verb-argument pairs from McRae et al. (1998) (the remaining 60 were used as a development set for parameter setting). Not all of the verbs and nouns used in this study were seen in our training corpora. For example, for only 64 test set items the verb had been seen in FN, while PB covers the verbs from 92 items. Nouns are even sparser. This is largely due to vocabulary differences between our training corpora and the items.

We gathered a second, larger test set ourselves with the goal of obtaining human judgement data that is more similar in vocabulary to the training data for a fairer evaluation. To ensure that all the verbs in the new test set are covered, we used 18 verbs that appear in both FN and PB. We extracted the three most frequent arguments seen as subjects and objects in each corpus, so that for each verb, there were usually six arguments from each corpus (some overlap could not be avoided). We constructed 414 verb-role-argument triples using for each verb-argument pair the roles that the verb typically assigns to its subject and object. Our semantic module has more information about the items in this set, but its task remains non-trivial, since half of the verb-argument pairs still have not been seen together in each training set.

Test	Train	Coverage	Correlation (ρ)
McRae	PB	88.0%	0.130, ns
	FN	56.0%	0.368, **
Own	Upper Bound	100%	0.68
	PB	100%	0.272, ***
	FN	98.6%	0.532, ***
	FN Seen	97.7%	0.593, ***
	FN Unseen	99%	0.428, ***

Table 2: Coverage and correlation strength for PB and FN data on McRae and own test sets. ns: not significant, **: $p < 0.01$, ***: $p < 0.001$

We collected ratings on the World Wide Web (using the WebExp package, `www.webexp.info`). To avoid participants rating the same item in both the agent and patient interpretation and due to the large number of items, we presented four separate lists of items that were assigned randomly to participants. Participation in the experiment was voluntary, but restricted to native speakers of English. The raters were recruited through postings to mailing lists and Usenet.

106 raters completed the experiment. We excluded five participants because they did not supply a valid email address (which we took as a sign of participation in earnest) and one non-native speaker. From the remaining 100 (25 participants per sub-experiment), we excluded one more participant who had rated only one item. We further excluded ratings that were more than 2 points from the item median. The average number of ratings per item was 21.

Results We trained our model on both the PB and FN corpora and created predictions for both test sets, which we then correlated with the human judgements. Since the data are not normally distributed, we used Spearman's ρ, a non-parametric rank-order test.

Table 2 gives an overview of the results. Due to the sparseness of verbs from the McRae et al. items in our training set, the coverage of the FN model is relatively low. Nonetheless, only its predictions are correlated significantly with the human data, despite the better coverage of the PB model. As intended, FN coverage rises when we use our own test set, and both models' predictions are significantly correlated to human judgements on the $\alpha = 0.001$ level. The FN model's ρ value is however still much higher than the PB model's.

To obtain an upper bound for model performance we computed inter-rater agreement, i.e., the degree of consensus about how a role should be rated. This can be estimated by correlating the ratings of a single participant with the average ratings of all remaining participants, and repeating for all participants. The resulting upper bound of $\rho = 0.68$ shows that our model performs reasonably well by achieving a maximum of $\rho = 0.532$. (The inter-rater agreement for the McRae et al. items is presumably similar.)

These results, however, raise the question why performance is so much better on our data than on the McRae et al.

Input	Syn	Sem	Conflict	Correct
The	MC	–	no	yes
employer	MC	–	no	yes
fired	MC	MC	no	yes
by	RR	MC	**yes**	**yes**

Input	Syn	Sem	Conflict	Correct
The	MC	–	no	yes
employee	MC	–	no	yes
fired	MC	RR	**yes**	**yes**
by	RR	RR	no	yes

Figure 4: MC/RR ambiguity: Preferred structure as predicted incrementally by our combined model consisting of a PCFG-based parser (Syn) and a semantic module (Sem) with % of conflict over all items. Good Agent first NP (left) vs Good Patient first NP (right).

items. A closer look at the seen and unseen verb-argument pairs in our own test set reveals that the model's predictions are better when more is known about the verb-argument pair, while the model still makes reliable predictions for unseen combinations. This explains the performance gap between our data and the McRae et al. items: Virtually all of the verb-argument pairs in the literature items are unseen, leading to predictions that are reliable, but worse than for our data.

In sum, using the FN corpus, we are able to correctly model items from the literature, as well as data from our own study. Our semantic module can therefore be used as a model of human semantic intuitions about verb-argument-role triples.

The MC/RR Ambiguity Revisited

We now consider our combined model and its predictions for the MC/RR ambiguity. The syntactic predictions are made by the same parser as in the Restriction to Syntax section above (see Fig. 2), and on their own would fail to account for both sentences. However, our semantic module is able to counterbalance the syntactic preferences: We assume that processing difficulty occurs if the semantic module prefers a different syntactic structure than the parser. We define the semantic module's preferred structure as the one that is consistent with the semantic module's preferred role. For example, if the model prefers an agentive role for *employer* given *the employer fired*, it thereby prefers the main clause reading.

As shown in Fig. 4, the semantic module provides predictions as soon as the first verb-argument pair is seen (at *fired*). For the good agents (*employer-fired*), it prefers the role that is consistent with the main clause reading. The parser and the semantic module agree, so no difficulty is predicted. Seeing *by* makes the main clause reading of *fired* unlikely because no direct object was seen (and none can follow now), and the parser switches to the reduced relative. However, the semantic module continues to prefer the role which indicates the main clause reading. This conflict between the syntactic and semantic modules predicts difficulty. This example was previously correctly accounted for by the parser alone, and we make the right predictions again using the semantic module.

For the good patients, the semantic module disagrees with the parser's preference already at *fire* by preferring the patient role for *employee-fired*, which is consistent only with the reduced relative structure. This conflict correctly predicts processing difficulty at the verb. As the sentence unfolds, the parser changes its preferred interpretation so that it agrees with the semantic module's, and no more difficulty is predicted at *by*. Using the semantic module in combination with the PCFG-based model allows us to make a correct prediction that the parser alone could not make.

The NP/S Ambiguity

We now turn to a second phenomenon, the so-called *NP/S complement ambiguity*. Note that, unlike the constraint-integration models, our model requires no changes to account for new phenomena. Consider sentence (2):

(2) *The man realised his goals were out of reach.*

At *goals*, it is unclear whether the NP is a direct object of *realise* or the subject of a sentence complement (as *were* later indicates). Pickering et al. (2000) investigated this phenomenon using nouns which are plausible and implausible objects of the first verb *realise*. Despite a verb preference for the sentence complement, they found evidence that readers initially prefer the NP interpretation: Their results show robust effects of difficulty in the noun region (*his goals*) when the noun is an implausible object of the verb. This indicates that readers initially construct the object reading of the noun phrase and reanalyse if this interpretation is implausible. Reanalysis is restricted to the noun region. At the disambiguating verb (*were*), there are indications of difficulty for sentences with plausible object nouns only. While this effect is weaker, Pickering et al. conclude that readers now reanalyse their initially plausible object interpretation of the NP.

Fig. 5 shows our model's predictions for the example item from Pickering et al. (2000). To see what a pure PCFG-based model would do, consider the parser's predictions in column Syn only. At *realised*, the predicted structure is the same for both interpretations. At *his* and *goals/shoes*, the parser predicts the object interpretation (NP), which is correct according to Pickering et al.'s results. [2] The parser fails to predict difficulty at *shoes* for implausible direct object nouns. Instead, it always predicts difficulty at the disambiguating verb by switching to the sentence complement (S) interpretation.

We now turn to the predictions of the combined model. For this ambiguity, the semantic model either prefers the direct object interpretation of *goals/shoes* by assigning a role licensed by *admit*, or it prefers to assume that a role will be assigned to the NP by an upcoming, unseen verb in the embedded sentence reading.

For the plausible object case (Fig. 5, left), *goals* is assigned a role: The semantic module agrees with the parser in preferring the object interpretation (Sem column). At *were*, it continues to do so, conflicting with the syntactic parser and thus correctly predicting difficulty. Later in the sentence, the main

[2] Note that this preference for the object interpretation is due to the small tree bias inherent in PCFG-based models, where fewer rule applications mean higher tree probabilities. Following Crocker and Brants (2000), we propose to interpret this bias as implementing a preference for simple structures.

Input	Syn	Sem	Conflict	Correct
realised	–	–	no	yes
his	NP	–	no	yes
goals	NP	NP	no	yes
were	S	NP	**yes**	**yes**

Input	Syn	Sem	Conflict	Correct
realised	–	–	no	yes
his	NP	–	no	yes
shoes	NP	S	**yes**	**yes**
were	S	S	no	yes

Figure 5: NP/S ambiguity: Preferred structure as predicted incrementally by our combined model consisting of a PCFG-based parser (Syn) and a semantic module (Sem). Plausible Object NP (left) vs Implausible Object NP (right).

verb of the embedded clause will be available to assign a role to *goals*, which can reverse the structural preference.

For the implausible object case (Fig. 5, right), the semantic module immediately prefers not to interpret *shoes* as a direct object of *realised*. This causes a conflict with the parser, which correctly predicts difficulty at the noun. This conflict is overcome when the parser at the next word switches to preferring the embedded sentence interpretation, too. Our model thus accounts correctly for human preferences that a pure PCFG-based approach cannot mode.

Conclusions

We reviewed two approaches to modelling human sentence processing. The first approach is based on probabilistic grammars, and has the advantage of being automatically trainable on a single data source. However, it does not incorporate semantic information, which means the influence of thematic information in human sentence processing cannot be captured. The second approach integrates a wide variety of constraints (Competition-Integration, Bayes Nets) and is able to correctly account also for thematic effects in the human data. However, the constraints have to be manually specified for each construction, and constraints have to be derived from a range of diverse data sources.

We presented an alternative model which builds on probabilistic grammar, but integrates a semantic module that assigns thematic roles for the structures generated by the parser. The parameters of the complete model can be automatically acquired from corpus data, which affords broad coverage and allows the model to cover different phenomena without requiring hand-tuning. We verified the cognitive plausibility of our semantic module by successfully correlating its predictions to human thematic fit ratings. We also demonstrated that the complete model correctly predicts processing difficulty for two classic ambiguities in the psycholinguistic literature: the MC/RR and the NP/S ambiguity.

In future work, we want to derive a quantitative measure of processing difficulty, as evidenced, e.g., by reading times. We will explore two different strategies: One is to continue exploiting the disagreement between the modules by quantifying the strength of disagreement. The other is to quantify the preference for the best analysis given a combined syntactic and semantic score.

References

Baker, C., Fillmore, C., and Cronin, B. (2003). The structure of the Framenet database. *International Journal of Lexicography*, 16(3):281–269.

Brysbaert, M. and Mitchell, D. C. (1996). Modifier attachment in sentence parsing: Evidence from Dutch. *Quarterly Journal of Experimental Psychology*, 49A(3):664–695.

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

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

Garnsey, S., Pearlmutter, N., Myers, E., and 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.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL*.

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

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

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

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Narayanan, S. and Jurafsky, D. (2002). A Bayesian model predicts human parse preference and reading time in sentence processing. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*. MIT Press.

Padó, U., Crocker, M., and Keller, F. (2006). Modelling semantic role plausibility in human sentence processing. In *Proceedings of EACL*.

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

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

Roark, B. (2001). *Robust Probabilistic Predictive Syntactic Processing: Motivations, Models, and Applications*. PhD thesis, Brown University.

Roland, D. and Jurafsky, D. (1998). How verb subcategorization frequencies are affected by corpus choice. In *Proceedings of COLING/ACL*.

Spivey-Knowlton, M. (1996). *Integration of visual and linguistic information: Human data and model simulations*. PhD thesis, University of Rochester.

Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2):165–201.

Tanenhaus, M., Spivey-Knowlton, M., and Hanna, J. (2000). Modeling thematic and discourse context effects with a multiple constraints approach: Implications for the architecture of the language comprehension system. In Crocker, M., Pickering, M., and Clifton, C., editors, *Architectures and Mechanisms for Language Processing*. Cambridge University Press.

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