The Integration of Syntax and Semantic Plausibility in a Wide-Coverage Model of Human Sentence Processing

Dissertation
zur Erlangung des akademischen Grades
eines Doktors der Philosophie
der Philosophischen Fakultäten
der Universität des Saarlandes

vorgelegt von Ulrike Padó
aus Detmold

Saarbrücken, 2007

Dekan:	Prof. Dr. Ulrike Demske
Berichterstatter:	Prof. Dr. Matthew W. Crocker
	Dr. Frank Keller

Tag der letzten Prüfungsleistung: 21.5.2007

Abstract

Models of human sentence processing have paid much attention to three key characteristics of the sentence processor: Its robust and accurate processing of unseen input (wide coverage), its immediate, incremental interpretation of partial input and its sensitivity to structural frequencies in previous language experience. In this thesis, we propose a model of human sentence processing that accounts for these three characteristics and also models a fourth key characteristic, namely the influence of semantic plausibility on sentence processing.

The precondition for such a sentence processing model is a general model of human plausibility intuitions. We therefore begin by presenting a probabilistic model of the plausibility of verb-argument relations, which we estimate as the probability of encountering a verb-argument pair in the relation specified by a thematic role in a role-annotated training corpus. This model faces a significant sparse data problem, which we alleviate by combining two orthogonal smoothing methods. We show that the smoothed model's predictions are significantly correlated to human plausibility judgements for a range of test sets. We also demonstrate that our semantic plausibility model outperforms selectional preference models and a standard role labeller, which solve tasks from computational linguistics that are related to the prediction of human judgements.

We then integrate this semantic plausibility model with an incremental, wide-coverage, probabilistic model of syntactic processing to form the Syntax/Semantics (SynSem) Integration model of sentence processing. The SynSem-Integration model combines preferences for candidate syntactic structures from two sources: Syntactic probability estimates from a probabilistic parser and our semantic plausibility model's estimates of the verb-argument relations in each syntactic analysis. The model uses these preferences to determine a globally preferred structure and predicts difficulty in human sentence processing either if syntactic and semantic preferences conflict, or if the interpretation of the preferred analysis changes non-monotonically. In a thorough evaluation against the patterns of processing difficulty found for four ambiguity phenomena in eight reading-time studies, we demonstrate that the SynSem-Integration model reliably predicts human reading time behaviour.

Zusammenfassung

Menschen verstehen Sprache in den allermeisten Situationen schnell und korrekt. Manchmal kommt es jedoch durch Verarbeitungsschwierigkeiten zu Verzögerungen oder sogar zum Scheitern der Verarbeitung, so daß überhaupt keine Analyse für das Gehörte oder Gelesene gefunden wird. Diese Dissertation behandelt die Modellierung des menschlichen Sprachverstehens auf der Ebene einzelner Sätze. Modelle des menschlichen Sprachverstehens sollen helfen zu erklären, wann und warum es zu Verarbeitungsschwierigkeiten kommt und wie diese überwunden werden. Während sich bereits existierende Modelle hauptsächlich mit syntaktischen Prozessen befassen, liegt unser Schwerpunkt darauf, ein Modell für die semantische Plausibilität von Äußerungen in ein Satzverarbeitungsmodell zu integrieren.

Vier wichtige Eigenschaften des Sprachverstehens bestimmen die Konstruktion unseres Modells: *Inkrementelle Verarbeitung*, eine *erfahrungsbasierte* Architektur, *breite Abdeckung* von Äußerungen, und die *Integration von semantischer Plausibilität*. Alle diese Eigenschaften sind zentrale Voraussetzungen für die menschliche Fähigkeit, Sprache schnell und korrekt zu verstehen. Sprachverarbeitung geschieht inkrementell, das heißt, jedes gehörte oder gelesene Wort wird sofort in die Interpretation der gesamten Äußerung integriert. Bei inkrementeller Verarbeitung müssen im Falle einer lokalen Ambiguität allerdings oft strukturelle Entscheidungen getroffen werden, bevor desambiguierendes Material erreichbar ist. In solchen Situationen zeigt es sich, daß das menschliche Sprachverstehen erfahrungsbasiert ist, also die partielle Äußerung so interpretiert, wie es der häufigsten Analyse in der bisherigen Spracherfahrung entspricht. Dieses Verhalten findet sich auf vielen Ebenen der Sprachverarbeitung, von der strukturellen bis zur lexikalischen, und hat zur Popularität probabilistischer Modelle beigetragen, die die Präferenzen aus vorhergehender Spracherfahrung durch Estimierung aus großen Textkorpora simulieren. Probabilistische Modelle erlauben auch eine breite Abdeckung ungesehener Äußerungen, die ein besonders auffälliges Merkmal menschlichen Sprachverstehens ist. Anhand der Verarbeitung von Ambiguitätsphänomenen zeigt sich schließlich auch, daß die Plausibilität der alternativen Interpretationen die Verarbeitung ebenfalls beeinflußt.

Während die Eigenschaften Inkrementalität, Erfahrungsbasiertheit und breite Abdeckung von vielen Modellen aufgegriffen wurden, gibt es kein Modell, das außerdem auch Plausibilität einbezieht. Das Fehlen solcher Modelle läßt sich zum Großteil darauf zurückführen, daß kein generelles Modell für menschliche Plausibilitätsbewertungen existiert. Daher behelfen sich viele Modelle mit Ansätzen, die sich nicht für breite

Abdeckung verallgemeinern lassen. In dieser Dissertation stellen wir deshalb ein generelles Plausibilitätsmodell vor, um es dann mit einem inkrementellen, probabilistischen Satzverarbeitungsmodell mit breiter Abdeckung zu einem Modell mit allen vier angestrebten Eigenschaften zu integrieren.

Unser Plausibilitätsmodell sagt menschliche Plausibilitätsbewertungen für Verb--Argumentpaare in verschiedenen Relationen (z.B. Agens oder Patiens) voraus. Das Modell estimiert die Plausibilität eines Verb-Argumentpaars in einer spezifischen, durch eine thematische Rolle angegebenen Relation als die Wahrscheinlichkeit, das Tripel aus Verb, Argument und Rolle in einem rollensemantisch annotierten Trainingskorpus anzutreffen. Für die naive Implementation dieses Modells stellen mangelnde Trainingsdaten ein schwerwiegendes Problem dar, so daß wir mehrere Methoden der Datenglättung anwenden: Zum einen Good-Turing Smoothing, das die ermittelte Wahrscheinlichkeitsverteilung re-estimiert und ungesehenen Ereignissen eine geringe Wahrscheinlichkeit zuteilt, und zum anderen ein wortklassenbasiertes Verfahren, das von Wörtern hin zu Worklassen generalisiert und daher während der Estimierung erlaubt, Datenpunkte zusammenzufassen. Die Plausibilitätsvorhersagen des endgültigen Modells korrelieren für eine Reihe verschiedener Testdatensätze signifikant mit menschlichen Plausibilitätsbewertungen. Ein Vergleich mit zwei computerlinguistischen Ansätzen, die jeweils eine verwandte Aufgabe erfüllen, nämlich die Zuweisung von thematischen Rollen und die Berechnung von Selektionspräferenzen, zeigt, daß unser Modell Plausibilitätsurteile verläßlicher vorhersagt.

Unser Satzverstehensmodell, das Syntax/Semantik-Integrationsmodell, ist eine Kombination aus diesem Plausibilitätsmodell und einem inkrementellen, probabilistischen Satzverarbeitungsmodell auf der Basis eines syntaktischen Parsers mit breiter Abdekkung. Das Syntax/Semantik-Integrationsmodell interpoliert syntaktische Wahrscheinlichkeitsabschätzungen für Analysen einer Äußerung mit den semantischen Plausibilitätsabschätzungen für die Verb-Argumentpaare in jeder Analyse. Das Ergebnis ist eine global präferierte Analyse. Das Syntax/Semantik-Integrationsmodell sagt Verarbeitungsschwierigkeiten voraus, wenn entweder die syntaktisch und semantisch präferierte Analyse konfligieren oder wenn sich die semantische Interpretation der global präferierten Analyse in einem Verarbeitungsschritt nicht-monoton ändert. Das Syntax/Semantik-Integrationsmodell ist damit constraintbasierten Ansätzen verwandt, da es wie diese Präferenzen aus verschiedenen Informationsquellen benutzt, um eine global präferierte Analyse zu bestimmen, und da es Verarbeitungsschwierigkeiten vorhersagt, wenn sich die Präferenzen aus den verschiedenen Quellen widersprechen. Es unterscheidet sich von diesen allerdings durch seine breite Abdeckung und darin, daß keine Constraints und Gewichte von Hand ausgewählt und definiert werden müssen. Die abschließende Evaluation anhand von Befunden über menschliche Verarbeitungsschwierigkeiten, wie sie experimentell in acht Studien für vier Ambiguitätsphänomene festgestellt wurden, zeigt, daß das Syntax/Semantik-Integrationsmodell die experimentellen Daten korrekt voraussagt.

Acknowledgements

I would like to thank my supervisors Matthew W. Crocker and Frank Keller, who both separately and jointly have been excellent advisors. I very much enjoyed working in Matt's group, and I have profited a lot from his advice and guidance. Frank has always been equally accessible, despite the physical distance, and has always had valuable comments on my work and the directions it should take. Throughout the three years, I have especially appreciated the smoothness and consistency of Matt's and Frank's joint supervision.

Thanks also go to my friends and colleagues in Computational Linguistics and Psycholinguistics at Saarland University and in Informatics at the University of Edinburgh for many profitable as well as for many distracting conversations. Very special thanks are due to my family, of course, and most of all to Sebastian.

I gratefully acknowledge a DFG doctoral scholarship in the International Graduate College "Language Technology and Cognitive Systems".

Contents

List of Figures

List of Tables

1. Introduction

Human sentence processing is generally very fast, robust and accurate. In some cases, however, the human sentence processor displays difficulty or is even unable to assign an interpretation. In these cases, readers take longer to process an utterance or encounter conscious difficulty with understanding it.

This thesis is concerned with the modelling of human sentence processing. The task of a sentence processing model is to process easy and difficult sentences, to identify when and why processing difficulty is encountered, and to explain how it is overcome. While existing models have mostly concentrated on syntactic processes, our focus of attention is on the integration of a model of human plausibility intuitions into a wide-coverage, experience based model of human sentence processing.

Phenomena that cause the processing system difficulty afford a valuable insight into the inner workings of the sentence processor. Consider the famous *garden path* sentence in (1.1). Such sentences cause most first-time readers processing difficulty or even lead to processing breakdown.

(1.1) The horse raced past the barn fell.

This sentence is difficult to understand because until the last word, most readers assume that the sentence is an active clause about a horse racing past a barn. However, at *fell*, it becomes clear that the human sentence processor has been led up the garden path: While *raced* seemed to be a main verb, it was in fact part of a reduced relative clause that refers to *the horse*. The complete sentence can be paraphrased as *The horse that was raced past the barn fell*. Given the ambiguity of *raced*, the processor has initially subscribed to an interpretation that is not consistent with the complete input sentence.

This example highlights an important property of human language processing: Processing proceeds *incrementally*. While processing sentence (1.1) and many other locally ambiguous sentences, processing difficulty arises from the fact that the human processor constructs a syntactic analysis and with it a semantic interpretation of its input immediately upon encountering each new word, without waiting for further information that disambiguates the intended interpretation in case of ambiguity. **Incrementality of processing** is therefore an important desideratum for psycholinguistic models: The processor appears driven by the desire to assign its input a semantic interpretation – to understand it – as quickly as possible. This strategy forms the core of several theories of sentence processing (Pritchett, 1992, Crocker, 1996).

A first class of models that was proposed to explain the processor's difficulty with sentences like (1.1) are *principle-based* models. We take this term to denote all models which assume that the sentence processor bases all its structural decisions on a limited set of processing principles. These can be defined syntactically as in Frazier's influential Garden Path model (e.g., Frazier, 1987) or on the basis of immediate semantic interpretability (Pritchett, 1992, Crocker, 1996). For example, in Frazier's account, the main clause interpretation of sentence (1.1) is initially preferred by the principle of Minimal Attachment, because it requires the postulation of fewer syntactic nodes than the reduced relative analysis. Garden path sentences incur processing difficulty when the initial attachment made by the processor proves to be wrong, and the assumed interpretation of the input has to be revised by a dedicated repair strategy.

Attention in recent years has shifted away from principle-based approaches to *probabilistic, experience-based* models that do not stipulate general principles, but rely only on structural frequencies to account for the processor's choices. This shift in emphasis was motivated by experimental results which show that human sentence processing is sensitive to frequency information on different levels of processing, including lexical word class membership frequencies (e.g., Trueswell, 1996, Crocker and Corley, 2002), verb subcategorisation frequencies (e.g., Trueswell, Tanenhaus, and Kello, 1993, Garnsey et al., 1997), and structural frequencies (e.g., Cuetos, Mitchell, and Corley, 1996). These observations can to some degree be integrated into principle-based models, but are more naturally accounted for by experience-based models and their assumption that processing is fundamentally guided by preferences accumulated in language experience (see, e.g., Jurafsky, 1996). A probabilistic formulation of experience-based models accurately captures the existence and strength of structural preferences. We therefore posit a **probabilistic approach** as another desideratum for psycholinguistic models.

Probabilistic models typically estimate lexical and structural preferences from large corpora of naturally occurring utterances to model human language experience. This makes it easy to construct models which robustly and accurately process unseen input. This **wide coverage** of unseen input, which is a fundamental characteristic of the human sentence processor, is another desideratum for sentence processing models.

The probabilistic account for the processing of sentence (1.1) rests on the processor's preference for more frequent alternatives. First, main clauses are overall more frequent than reduced relative constructions, and second, *raced* is used in the simple past tense more often than as a past participle (as in the reduced relative interpretation). For these two reasons, probabilistic models predict the sentence processor to prefer the main clause interpretation.

Processing in probabilistic models usually is *parallel*, that is, they construct all or at least a number of possible analyses for the input. Parallel models do not require the stipulation of an explicit reanalysis strategy. A structural interpretation that becomes impossible, like the main clause interpretation of sentence (1.1) when reaching *fell*,

is simply replaced with the most probable remaining alternative, and difficulty is predicted to be caused by the replacement process.

Probabilistic models are implemented in a number of different architectures, from connectionist networks and approaches based on statistical parsers to constraint-integration models, which formulate the processor's decision about which structural analysis to prefer at each point as the integration of evidence for or against the analyses.

For a large number of difficulty phenomena, it has been demonstrated that the semantic plausibility of the alternative syntactic analyses has an effect on processing (see also the review of experimental results for four phenomena in Chapter 5). While the majority of psycholinguistic theories and models of sentence processing assumes that the human sentence processor makes use of plausibility information, fewest approaches specify this aspect more precisely. This thesis is especially concerned with modelling human plausibility intuitions and the influence of plausibility on sentence processing. Other semantic effects exist, for example an influence of discourse context on ambiguity processing (e.g., Altmann and Steedman, 1988), which are not treated here.

To illustrate an effect of plausibility, consider a second example. Sentences (1.2) and (1.3) from McRae et al. (1998) contain the same local ambiguity as sentence (1.1).

(1.2) The doctor cured by the treatment had developed it himself.

(1.3) The patient cured by the treatment had been diagnosed as terminal.

Sentence (1.2) causes readers more processing difficulty at and after the *by*-phrase than sentence (1.3). Since the structure of both sentences is identical up to *had*, which disambiguates the local ambiguity, the difference between the sentences must lie with the words in the subject NP. Indeed, *doctors* are likely *curers*, which points the human sentence processor towards a main clause interpretation in which *the doctor* is a *curer*. *Patients*, on the other hand, make less good *curers* in the main clause interpretation, but good *curees* in the alternative reduced relative interpretation. This plausibility bias towards the ultimately correct analysis facilitates reading the disambiguating main verb. This example underlines the influence of **plausibility** information in sentence processing. A fourth desideratum for models of human sentence processing is therefore to account for this factor, as well.

1.1. Plausibility in a Wide-Coverage Sentence Processing Model

We have identified four desiderata for a model of human sentence processing: Incrementality, wide coverage, a probabilistic architecture, and the incorporation of plausibility. This list is of course not exhaustive of all conceivable desiderata, but it helps to outline the types of models we are interested in. The first three desiderata have

received considerable attention in the modelling literature, but there are few accounts that specify the integration of plausibility, and none that conform to all four desiderata.

The lack of incremental, wide coverage, probabilistic models that also integrate plausibility is due at least to some extent to the fact that no general model of human plausibility intuitions exists. In this thesis, we propose such a model, and then integrate with it an incremental, wide coverage, probabilistic account of human sentence processing to create a sentence processing model that conforms to all four desiderata.

We propose a model of semantic plausibility that accounts for human intuitions about verb-argument relations. From the many available characterisations of plausibility, we choose the level of verb-argument relations both because it captures the basic who-does-what-to-whom information in a sentence, and because this is the typical level of plausibility manipulations in the experimental studies we will use to evaluate the model.

Recall the plausibility manipulation demonstrated by sentences (1.2) and (1.3). It relies on creating verb-argument pairs which are more plausible in one of the alternative relations made available by local ambiguity than in the other. To account for this kind of plausibility effect, a plausibility model has to identify the two possible relationships between *doctor* and *cured* that can be paraphrased as *the doctor cured* versus *the doctor was cured*, and then evaluate the plausibility of seeing *doctor* in either of these relations to *cure*.

The strategies employed in psycholinguistic modelling to account for plausibility all lack the potential for a wide-coverage model. A first strategy directly integrates human judgements for *the doctor cured* versus *the doctor was cured* into the model (e.g., McRae et al., 1998, Narayanan and Jurafsky, 2002). This approach does account for the influence of human plausibility intuitions in processing, but its reliance upon actual human judgements precludes the development of a general, wide-coverage model.

The second approach, used by connectionist models, is to approximate the plausibility of a verb-argument relation by learning the distributional behaviour of verbs and arguments (e.g., Elman, 1990, Rohde, 2002). This approach is limited because it requires large amounts of training data to account for the plausibility of verb-noun co-occurrences. Therefore, it is hard to apply to realistic vocabulary sizes: Even very large corpora of naturally occurring language data by Zipf's law contain many lexical items that are too infrequent to allow robust inferences about their distributional behaviour.

In computational linguistics, there are two tasks that are relevant to predicting plausibility and that have been addressed with wide-coverage models. The first task is semantic role labelling. It consists of assigning verb-argument pairs in syntactic context the appropriate thematic roles that semantically characterise their relation. Defining the two possible relations between *doctor* and *cured* in the garden path sentence (1.2) by assigning the appropriate thematic roles – the doctor being the *curer* or the *curee* – is the first step towards estimating how plausible the verb-argument pair in each relation is. The task of semantic role labelling can thus be seen as a subtask of predicting

human plausibility judgements. Semantic role labelling has been a popular research area recently, since role annotated corpora have become available for training (see, e.g., Carreras and Márquez, 2004, 2005). Therefore, the task is well understood and models perform on free text with increasing accuracy. However, most models heavily rely on syntactic properties of the whole input sentence, that is, they are neither incremental, nor do they pay much attention to the identity of the argument, because syntactic features are likely to be less sparse than information about specific arguments, and therefore have more generalisation power.

Selectional preference models take up where role labellers leave off and address the task of estimating the plausibility of a given relation for a verb-argument pair (Resnik, 1997, Li and Abe, 1998, Clark and Weir, 2002). They model the fit between a verb and its argument given a specified relation by comparing the similarity of the current argument and typical arguments for the verb in the relation, as seen in a large corpus. The models overcome the problem of sparse training data and unseen test instances by pooling observations using semantic classes. However, such models are somewhat too coarse for our purposes because they use the syntactic relationship between verb and argument to characterise their relation, rather than the thematic role. This characterisation is not sufficiently fine-grained to distinguish between the two possible relations between *doctor* and *cured* in *The doctor cured...*: Both the agentive and the passive reading are realised with *doctor* as a syntactic subject or external argument of *cured*.

In sum, while there are related approaches in computational linguistics that address similar tasks to those of a plausibility model and achieve wide coverage, none are directly suitable as a plausibility model. We propose a semantic plausibility model that borrows from the computational linguistics approaches by using corpus resources, relying on thematic roles to characterise the relationship between a verb and its argument, and by pooling data using semantic classes. It differs from the computational linguistics approaches, however, in that it simultaneously identifies all possible relations between verb and argument and estimates their plausibility. In Chapter 4, we show that our semantic plausibility model outperforms both related approaches from computational linguistics.

We use a corpus-based model on the assumption that the frequency of verb-argument relations in a corpus can be used to predict the plausibility of the verb-argument relations in unseen utterances. The approach gives the semantic plausibility model wide coverage of unseen verb-argument pairs (within the limits of the available training data) and ensures its compatibility with the probabilistic experience-based model with which it is combined.

We integrate the semantic plausibility model with an incremental, wide-coverage, probabilistic model of syntactic processing to construct a new model of human sentence processing that fulfils all modelling desiderata. The Syntax/Semantics (SynSem) Integration model transparently combines the syntactic preference predictions by a

statistical parser-based model and the semantic preferences based on plausibility evaluations of the proposed analyses. It predicts difficulty either if syntactic and semantic preferences conflict, or if the assumed interpretation of the input changes. The prediction of conflict cost reveals the SynSem-Integration model's close relationship to constraint-based accounts. Chapter 5 demonstrates that the SynSem-Integration model correctly accounts for the experimental results of processing difficulty found in eight reading-time studies.

1.2. Organisation of the Thesis

Chapter 2 discusses in more detail the three classes of probabilistic experience-based models that have been proposed in the literature and demonstrates that none of them meets all modelling desiderata. We therefore go on to propose the SynSem-Integration model and, as a precondition, the semantic plausibility model.

Chapters 3 and 4 focus on the semantic plausibility model. Chapter 3 discusses different approaches to estimating and smoothing the model and describes the selection of the best model instances. These instances are evaluated in Chapter 4, where we show that the semantic plausibility model predicts human plausibility judgements from a range of different studies.

Having established a plausibility model which is a reliable predictor of human plausibility judgements, we go on to present the SynSem-Integration model in more detail in Chapter 5, where we also discuss the implementation of the cost functions and parameter selection. Finally, in Chapter 6, we introduce four phenomena from the psycholinguistic literature that cause the human sentence processor difficulty and show that our model predicts patterns of human processing difficulty as observed for these phenomena in eight experimental studies. Chapter 7 concludes and gives an overview of future work.

2. Computational Models of Sentence Processing

In Chapter 1, we have argued for an incremental model of human sentence processing that accounts for the influence of human semantic plausibility intuitions and exhibits broad coverage of both psycholinguistic phenomena and the large range of easily-processed naturally occurring utterances. We have identified a probabilistic, experience-based architecture as a plausible basis for such a model. This chapter first briefly discusses the basic assumption made by this architecture, namely that processing preferences can be induced from structural frequencies in text, and then describes existing, implemented computational models of sentence processing from the three most prominent classes of probabilistic models: *Connectionist* models (described in Section 2.2), *probabilistic grammar-based* models (Section 2.4) and *constraint-integration* models (Section 2.3). We review each type of model with respect to our requirements and identify both shortcomings of the existing models, but also aspects which contribute to our own proposal. We introduce this proposal, the Syntax/Semantics (SynSem) Integration model, in Section 2.5. It accounts for syntactic preferences and semantic effects in human sentence processing in a transparent way, while preserving wide coverage of both experimental phenomena and naturally occurring language. The review of existing models shows that those models that integrate a notion of semantics usually rely on costly human judgements. In contrast, the SynSem-Integration model includes an experience-based model of human plausibility intuitions.

2.1. Assumptions of Probabilistic Models

Probabilistic models implement an experience-based approach to human sentence processing which assumes that the sentence processor prefers those structural and lexical interpretations of the input that it has encountered frequently during previous experience. Probabilistic models are motivated by the observation that all levels of human sentence processing are sensitive to frequency information (see, e.g., Jurafsky, 2003, for a comprehensive overview). Frequency effects have for example been found for lexical word class membership frequencies (e.g., Trueswell, 1996, Crocker and Corley, 2002), verb subcategorisation frequencies (e.g., Trueswell et al., 1993, Garnsey et al., 1997), or structural frequencies (e.g., Cuetos et al., 1996). Probabilistic models account for these preferences by estimating lexical and structural preferences from large

corpora of naturally occurring utterances. This approach makes the crucial assumption that frequencies observed in corpora will reliably reflect the preferences of the human processing system.

The predictive power of corpus frequencies for processing preferences is intensively discussed in the literature. In general, correlations between corpus frequencies and processing preferences are found, but usually only when a number of factors are carefully controlled. For example, while Cuetos et al. (1996) found a preference in English and Spanish to attach relative clauses in the way most frequently encountered in corpora, Mitchell and Brysbaert (1998) found that this is not the case for Dutch. Desmet, Brysbaert, and de Baecke (2002) and Desmet, de Baecke, Drieghe, Brysbaert, and Vonk (2005) in turn showed that readers did prefer the more frequent attachment, once animacy and concreteness of the first attachment site were taken into account.

A similar picture emerges for the correlation between verb subcategorisation preferences extracted from corpora and those found in production (completion studies) and comprehension (reading-time experiments). Merlo (1994) and Gibson, Schütze, and Salomon (1996) provide evidence against a positive correlation. However, it appears that there are several important factors that need to be controlled in the determination of subcategorisation preferences: Rather than being defined per verb lemma, a verb's subcategorisation preferences change by sense (Roland and Jurafsky, 2002, Hare, McRae, and Elman, 2003). Hare, McRae, and Elman (2004) demonstrate that the amount to which corpus-extracted verb-sense specific subcategorisation preferences correspond to the preferences assumed by experimenters in a number of studies predicts whether an effect of verb bias was found experimentally or not.

Hare et al. (2003) also find that preceding context can bias comprehenders to prefer one of the senses of ambiguous verbs, and Keller and Scheepers (2006) show for German that the subcategorisation preferences even of unambiguous verbs vary depending on preceding context. Finally, genre- and discourse type-dependent usage may also determine a verb sense's subcategorisation preferences (Roland and Jurafsky, 1998). This allows the hypothesis that verb subcategorisation preferences extracted from a large, balanced corpus that contains evidence of many different types of language data should yield more reliable subcategorisation preferences. Indeed, Lapata, Keller, and Schulte im Walde (2001) show that verb subcategorisation preferences extracted from the BNC (Burnard, 1995), a balanced corpus, are significantly correlated with the human norming results from four experimental studies.

In sum, it appears justified to assume that frequencies from a sufficiently large corpus correlate well with human preferences both in production (e.g., completion tasks) and comprehension. However, it is clear that this correlation holds only when additional variables such as verb sense or animacy of a possible attachment site are carefully controlled. These variables are also important indications of the grain size of experienced events to which the human sentence processor takes recourse, and the influence of factors like animacy point to more than purely structural events.

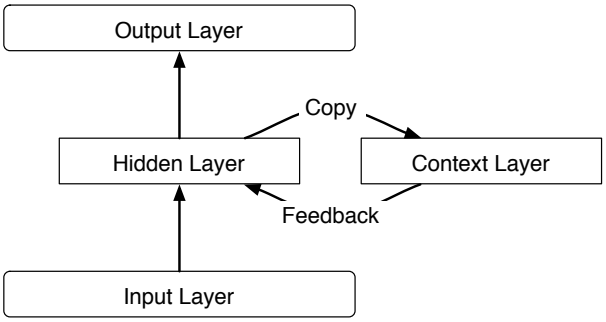


Figure 2.1.: Schematic diagram of an SRN: Input and output layer with intervening hidden layer and context layer, which stores the hidden layer's activation and feeds it back at the next time step.

2.2. Connectionist Models

Connectionist models (also known as neural networks) are inspired by the neural architecture of the brain. These models have been used to investigate how the structure and meaning of language can be learnt from exposure to language data in a cognitively plausible way. They are especially qualified for this task because they exhibit similar properties to the human language system in several respects: They are very good at detecting associations and structure in their training data, are robust to noisy input and show graceful degradation (as opposed to abrupt failure) in the case of damage. Since they require long training times and become very complex with increasing size, scaling up to cover a realistic amount of language phenomena and to process free text has been a great challenge that has only begun to be addressed recently.

All connectionist models consist of layers of nodes linked by weighted connections. Input information is usually specified to a separate input layer, and the models' output is read off an output layer. Any layers between the input and output layers are called *hidden layers*. Hidden layers allow the network to develop its own encoding of the input information which typically consists of activation patterns across all nodes in the layer. These representations allow generalisation through the creation of similar representations for similar concepts and robust processing in case of noise or damage.

Activation propagates through the network along the weighted connections. The activation that a node receives through its input connections is summed and passed on through all outgoing connections. Scaling ensures that small amounts of activity are suppressed and that large amounts are capped. The weights of the connections are incrementally adapted during many rounds of training to shape the network's output

reaction to an input more like a specified target output.

For incremental language processing, a sequence of inputs over time has to be encoded. This is facilitated by a type of model called the Simple Recurrent Network (SRN, Elman, 1990), which is the basis for most models of sentence processing. SRNs encode previous context through a feedback loop, which stores the state of the hidden layer at the last time step and feeds it back into the hidden layer together with the new input, as shown in Figure 2.1. In this way, the model uses both new information and the interpretation of the input at the last time step to arrive at an interpretation for the current time step. This provides the model with extended memory of previously processed input and allows it to learn structural information encoded in the order of input events, which is a typical feature of language.

There are two typical training regimes for connectionist models that entail different test tasks. Elman (1990) introduced the *next-word-prediction task* which is often used both for training and for the evaluation of trained models. The model is presented with input incrementally and has to predict the next word or next category. The model thus has to learn to associate input word strings and possible lexical continuations. When used in training, the task provides the model with an implicit training signal by comparing the actual next word to the prediction. Note, however, that Steedman (1999) characterises this task as equivalent to part-of-speech tagging rather than syntactic parsing, because the SRN does not arrive at a structural analysis of the input.

Alternatively, models can be trained in a completely supervised fashion by presenting the input signal together with an explicit target output. The target output is usually a semantic encoding of the language input (see, e.g., McClelland, St. John, and Taraban, 1989, Mayberry, 2003). The network's task is to associate the correct input and output patterns so that it can also correctly build semantic representations for novel input combinations. During testing, the accuracy with which the model produces the correct output is evaluated.

This task aims overtly at learning to understand language input, which is arguably more similar to a human language learner's task than next-word prediction. However, it also requires an explicit teacher signal which specifies the correct interpretation of the input. Such an extremely reliable signal is usually not available in human language learning, where there is scope for misunderstanding the relationship between an utterance and its intended meaning.

2.2.1. Early Models

One of the first connectionist models of sentence processing was developed by St. John and McClelland (McClelland et al., 1989, St. John and McClelland, 1990) to demonstrate that the syntax and semantics of language can be learnt simply through exposure to event representations and language input.

Their model was trained on a corpus of 630,000 sentence/event pairs. The sentences

were active or passive declaratives and contained a number of lexical items that were vague (e.g., *someone*) or ambiguous (e.g., *ball*). During training, the model received a sequence of sentence constituents as input and an encoding of the corresponding events as thematic role relations between entities and actions as target output. In its hidden layer, it incrementally built a representation for the complete input sentence, using a feedback loop similar to that of an SRN (but feeding back the activation of the output rather than the hidden layer). Once a representation for the sentence, termed the *sentence gestalt*, was constructed, the model could be queried about its understanding of the sentence in terms of thematic role relations between events and protagonists.

The trained model was able to answer probes about roles and fillers after processing (partial) input. It performed semantic tasks like correctly assigning thematic roles, finding the correct concept instantiation for vague descriptions, or inferring a thematic role and filler that was not mentioned in the input, for example an instrument role. The model also learnt syntactic generalisations, which enabled it to construct the same semantic representation for alternative syntactic realisations of the input, such as active and passive voice or the double-object diathesis alternation.

The model thus demonstrated that information about syntactic structure and semantic relations can be learnt from combinations of linguistic input and situation information without any innate structural knowledge about language, only by extracting regularities from a large number of input sentences.

Elman (1990) showed that a true SRN model trained on short sentences using the next-word prediction task can induce a representation for the lexical material according to its syntactic class and semantic properties, even if no explicit thematic role information is given. Clustering the hidden layer representations for the input words after training showed that verbs were represented separately from nouns, and subdivided by argument frames. Nouns were clustered, for example, according to animacy and being human. This hierarchy had fallen out from the usage of words in the training corpus, for example the fact that the verbs were always used with a specific argument frame or that only animates could take the argument position of certain verbs. With this clustering of similar words, the model's representations for the items in its lexicon are structured in a similar way as those of the human mental lexicon (see Section 2.5.1).

Finally, Elman (1991) presented results for training an SRN model on more complex sentence structure. The training data for this model encoded verb-argument agreement, argument structure preferences (such as transitive and intransitive verbs) and relative clauses which raise the complexity of processing agreement and allow recursion.

The network was trained and evaluated on the next-word prediction task, where accuracy was determined by comparing the activation pattern for the predicted next words to the statistical distribution of next words in the training data. The test set was novel, but not guaranteed to contain only unseen sentences. The model correctly accounted for verb agreement and verb subcategorisation preferences, even in complex sentences involving relative clauses. Inspection of the model's state space shows that its

representation of the syntactic structure of relative clauses and main clauses generalises over similarities between them. The model thereby showed itself capable of processing input with internal structure and of representing this structure directly in its hidden layer.

2.2.2. Approaching Realistic Coverage

The early models introduced above proved that recurrent networks are able to infer syntactic structure of the input and even learn some lexical semantic information such as selectional preferences from the distribution of words in the training data. However, they were small proof-of-concept studies with restricted coverage of syntactic constructions and vocabulary. More recently, the focus of research has been to build models that demonstrate a much wider coverage of constructions and that can process naturally occurring text. These models aim to account for human adult performance as well as for the language learning process.

Rohde (2002) for example describes an SRN model of sentence comprehension and production that covers a greater range of syntactic constructions. The model either generates sentences from a given semantic representation or, in comprehension mode, predicts the next input word. It covers a large spectrum of syntactic constructions, such as sentential complements and subordinate clauses, relative clauses of different kinds, prepositional phrases and coordination. In addition, the training data contains a relatively large lexicon of nouns, adjectives and verbs in different tenses, with some instances of lexical ambiguity. While its training data is not yet naturally occurring text, the statistics of the training corpus correspond broadly to those of the Penn Treebank corpus of English (Marcus, Santorini, and Marcinkiewicz, 1994).

The model is of special interest to us as the evaluation on the comprehension task includes the prediction of reading times for a range of ambiguity phenomena. Reading time predictions are based on the combination of two measures of comprehension difficulty encountered by the model. Difficulty is represented on the one hand by the amount to which the next input word is predicted by the model and on the other hand by the amount of change taking place in the semantic representation of the input sentence. The measure thus reflects both the syntactic and semantic expectedness of the current input word, which accounts for lexical and structural frequencies involved in the different possible analyses, and the amount of semantic processing necessary to integrate the new input, which consists of a more or less drastic change of the semantic representation of the complete input.

Comparison to reading time is done for four well-studied ambiguity and memory-load phenomena. Because the model cannot directly process experimental items from the literature, its predictions of reading time effects are computed on a controlled set of input sentences that manipulate the appropriate factors (e.g., thematic fit of an argument) in the context of the training data. Consequently, the model also cannot

exactly predict the results of any one reading study, but is evaluated on predicting the general reading-time profile found by a range of studies for the same phenomenon.

Across all phenomena, the model proves that it is sensitive to the manipulation of the experimental factors. The model generally predicts similar patterns of difficulty to those found experimentally for local ambiguity phenomena. However, it fails to correctly predict comprehenders' preferences for different types of relative clause, which appears to be due to specifics of training and to the distribution over structures in the training language. Overall, the model is clearly able to match human experimental results even in a quite indirect evaluation. Rohde's model thus constitutes a model of adult human sentence comprehension (and production) on a still restricted, but relatively realistic scale.

Mayberry (2003) takes a further step towards a wide coverage model by training and testing on actual corpus data. He uses the Redwood Treebank (Oepen, Flickinger, Toutanova, and Manning, 2002), a corpus of about 5,000 sentences of spoken language that were transcribed and annotated with parse trees and semantic representations in the Minimal Recursion Semantics (MRS) format (Copestake, Lascarides, and Flickinger, 2001). Mayberry's model learns to incrementally build the MRS representation of the input string. Evaluation thus focuses on understanding the input, which makes correct acquisition of syntactic structures and a mapping to the corresponding MRS representations necessary as a pre-condition.

Using MRS as a semantic representation introduces the problem of representing graph structure in the network. This is solved by the co-operation of a series of components: The first, an SRN model, reads in the input and retains a representation of it. Then, a second hidden layer called a Frame Map processes this representation to generate the nodes of the corresponding MRS structure. The Frame Map is self-organised, i.e., different areas specialise to encode different types of nodes like determiners, verbs or nouns. The arcs that link the different nodes into a graph are represented in the node representations as pointers from the Frame Map region that encodes the current node to other regions. Finally, the MRS representations are decoded into the output layer.

The trained model is able to reliably parse a completely unseen treebank test set. The main source of errors is that pointers tend to be slightly inaccurate and not indicate the correct target but a neighbouring node. Since similar nodes are generally represented in similar regions of the hidden layer, their confusability is indeed high. The model reacts robustly to the introduction of noise, both as pauses and dysfluencies in the input from the raw transcripts of the treebank sentences, and to noise directly introduced into the model's weights. The network also accounts for ambiguity in the input utterances, which is quite high in the treebank, where three quarters of sentences have more than one acceptable parse. The network represents alternative analyses by activating the corresponding MRS representations to different degrees, with the preferred analysis activated most strongly. In sum, the trained model is capable of correctly processing noisy, ambiguous naturally occurring language data, while also accounting for the

learning process that leads to fully competent language processing through the self-organised acquisition of semantic representations in the Frame Map. However, unlike Rohde's model, it has not been applied to the prediction of experimental reading time data.

2.2.3. Summary

Connectionist models of human sentence processing have demonstrated how it may be possible for children to acquire the structure and meaning of language simply from listening to utterances in a situation context, without any prior knowledge (other than the assumptions encoded in the model architecture). Furthermore, Rohde (2002) and Mayberry (2003) have constructed models that realistically account for adult language processing. The models mirror human processing effects in the psycholinguistic literature and account for comprehension of unseen, noisy naturally occurring utterances, respectively.

The most serious problem with connectionist models is that scaling them up to processing realistic language utterances makes them grow extremely complex and difficult to train. Therefore, further extending the existing relatively large-scale models remains a challenge. A related problem is that the emerging representations in large models are complex and opaque, which makes it harder to assess why and how the model arrives at its generalisations. Finally, the need to derive a representation for an input word exclusively from its distributional profile amplifies the sparse data problem: A data point presented only a few times does not influence the model's connection weights sufficiently to be well represented. This makes full-scale, realistic input models especially vulnerable to the sparse data problem inherent in every realistic task. In consequence, no model has yet been used to make predictions of processing difficulty for individual studies. Mayberry's model is the only one so far which is able to process free text, but given the relatively small size of its training corpus, it presumably would encounter a serious problem with unknown words when processing actual experimental items.

The representation of semantic information is also subject to the scaling problem. Connectionist models can acquire information about the distributional properties of lexical items that can be taken to represent their semantics, but for this they require a large amount of training examples. Therefore, full-scale semantic representations for a large input vocabulary can only be learnt through an intensive training regime.

2.3. Constraint-Integration Models

Constraint-integration models share the assumption that the preferred analysis at each processing step is determined by the immediate and simultaneous interaction

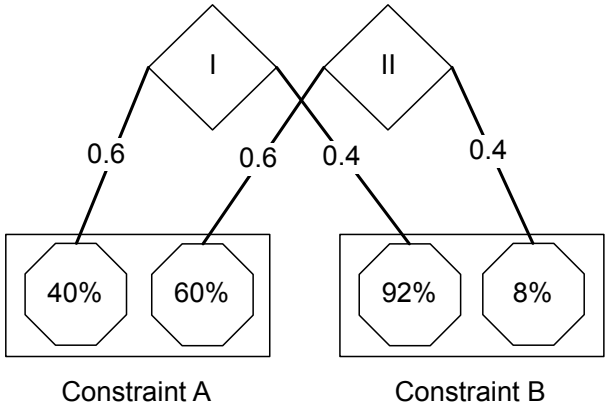


Figure 2.2.: Schematic diagram of the Competition-Integration model (adapted after McRae et al. (1998)). Constraints and example levels of support for alternative analyses.

of a number of constraints of different strength. Typical constraints are based on lexical biases like verb subcategorisation preferences, or on structural preferences like a general bias towards more frequent structures over infrequent ones. Each constraint is integrated into the analysis as soon as it becomes available. Constraint-integration approaches therefore account for the influence of semantic plausibility by simply positing it as an additional constraint.

Such models further assume that processing difficulty is directly related to the time it takes the processor to integrate conflicting evidence for which of the possible analyses should be preferred. If many strong constraints point to one analysis, the processor can decide easily and quickly, but it will take longer to reach a decision if all alternatives are about equally supported. Processing difficulty is therefore linked to the time it takes for a model to settle on a preferred analysis. Constraint-integration models differ from probability distribution parser models in that they predict a preferred analysis at each processing step. In contrast to ranking parser models, they make graded, stimulus-specific difficulty predictions. Constraint-integration models borrow the metaphors of weighted connections and spreading activation from connectionist networks.

There is a number of theoretical proposals for constraint-integration models (e.g., Bates and McWhinney, 1989, MacDonald, Pearlmutter, and Seidenberg, 1994, Boland, 1997) and an even larger number of publications which support the idea of constraint-based processing without proposing models in full detail. One of the few implemented accounts, the Competition-Integration model (Spivey-Knowlton, 1996, Spivey and

Tanenhaus, 1998) uses the competition approach to distinguish between possible analyses of the input, which are assumed as given. The model computes how much support each analysis receives from the range of weighted constraints, and iterates until a critical activation threshold is met by one of the analyses.

Figure 2.2 shows a schematic diagram of the model, based on the description in McRae et al. (1998). Two alternative analyses of the input, I and II, are supported by two constraints via weighted connections. Constraint A supports interpretations I and II almost equally, while Constraint B strongly supports interpretation I and only weakly supports interpretation II. During each iteration, a feedback mechanism serves to re-calculate each constraint's support for the analyses depending on the connection weights and the respective activity of each analysis. Since a constraint's support for the alternatives is normalised to sum to 100, strengthening its support for the more active analysis at the same time weakens its support for the less active analysis. The model thus moves towards preferring one analysis in the normal case. To ensure termination of the iterative cycle and to put an upper bound on the number of iterations, the activation threshold for settling is lowered at each iteration. Thus, the model will eventually settle on one alternative even if both analyses are supported equally strongly. In the Competition-Integration model, the number of iterations until settling is assumed to be linearly related to processing difficulty as reflected in longer reading times.

Tanenhaus, Spivey-Knowlton, and Hanna (2000) present several simulations of syntactic ambiguities that are influenced by thematic fit and context effects. The simulations demonstrate that the Competition-Integration model accounts for different types of data patterns that have often been interpreted as evidence against this type of model. However, the model encounters a methodological problem: The set of constraints to be used for modelling and the point in processing at which they are assumed to be available to the human processor have to be determined manually. This compromises the wide coverage of each individual model instance. Further, constraint strengths are usually inferred based on evidence from corpus or norming studies, which introduces the danger of finding inconsistent preferences (see, e.g., Roland and Jurafsky, 1998). Constraint weights must either be set by hand, or fit to off-line completion studies (see McRae et al., 1998). Spivey-Knowlton and Sedivy (1995) suggest using regression analyses or a connectionist network to learn the weight settings.

The integration of plausibility information in the shape of human judgements elicited in rating studies is at the same time an advantage of the Constraint-Integration model and another restriction of wide coverage, because the ratings have to be elicited anew for each new set of experimental data.

A related model proposed by Tabor, Juliano, and Tanenhaus (1997) aims to avoid the necessity of selecting constraints altogether. Their model draws on ideas from dynamical systems theory and describes the parsing of an input sequence as the

trajectory of an object through a metric space, past attractors that represent distinct processing states and correspond to decisions in a grammar-based parser. Reading times are predicted by the speed with which the object is drawn to one of the possible analyses, which depends on the force exerted by the attractors and their proximity. Tabor et al. use the clusters of similar syntactic analyses that emerge during training of a connectionist network as competing processing states. Thus, no constraints have to be specified by hand, and the strategy also allows Tabor et al. to give a tentative account of how linguistic classifications may emerge that are treated as given in non-connectionist models. However, the model inherits the poor scalability of neural networks to wide coverage (recall Section 2.2). For example, Tabor and Tanenhaus (1999) demonstrate as a proof of concept that the model is able to also learn thematic fit constraints, but they use a very restricted set of training data. In order to induce a more general account of thematic fit in the model, an impractically large amount of training data presumably would be needed.

2.3.1. Summary

The Competition-Integration model, as an implemented representative of a large number of constraint-integration models, provides a natural way of integrating semantic information into a processing model. However, used across different phenomena, it constitutes a consistent modelling architecture rather than one specific model of sentence processing, in that it requires a new set of constraints, biases and weights to be chosen for each ambiguity that is to be modelled. Consequently, individual model instances are unlikely to generalise to new constructions without further changes. The use of costly human judgements to model the influence of plausibility further restricts wide coverage of unseen input data. A second problem is that the constraint weights often are estimated from a diverse, potentially conflicting set of sources, for example various corpus studies as well as rating and completion studies (e.g., McRae et al., 1998). In addition, the Competition-Integration model does not account for the creation of syntactic analyses. Instead, the way it decides between given interpretations is considered an abstract characterisation of information integration during sentence processing. This characteristic however compromises its wide coverage, as the syntactic alternatives presented by the input have to be known beforehand for every input string.

An alternative proposal based on dynamic systems theory by Tabor et al. elegantly avoids the selecting of constraints and estimating of parameters by training a connectionist model. However, in this way it inherits problems from connectionist models, for example poor scalability to realistic input.

In sum, the simple integration of various sources of information into a single constraint-integration model is balanced by methodological and practical disadvantages such as lack of coverage and the necessity to hand-select constraints or manually set parameters.

2.4. Probabilistic Grammar-Based Models

While connectionist networks exclusively rely on the language structure they can infer from their training data, probabilistic grammar-based models (*parser models*) are presented with structural information in the form of annotation to the input and only need to learn the distribution over the given structure over the input data. Thus, parser models share with connectionist models the idea of learning from language experience, but they do so in reference to a given grammar that overtly specifies the structure of the input data. Therefore, parser models do not give an account of the acquisition of structural knowledge about language, but they do account for the acquisition of structural preferences. Since probabilistic grammars are a standard tool of computational linguistics, models can profit from developments in that field, which allow them accurate, wide coverage of unseen language data. For example, smoothing mechanisms can be employed to alleviate the problem of sparse training data.

Typically, parser models use a Probabilistic Context-Free Grammar (PCFG) to compute the probability of each possible structural analysis of an input sentence. A PCFG consists of a set of context-free rules. These define which daughter nodes in a phrase structure tree a mother node may have. Each rule is annotated with a probability which represents the likelihood of expanding the mother node into the daughter nodes. These rule probabilities are usually extracted from large corpora with syntactic annotation.

The annotation of grammar rules with probability information allows the ranking of generated tree structures by their probability according to the grammar: The probability of a syntactic structure (parse tree) is defined as the product of the probabilities of all rules applied in generating it. Parser models usually base their predictions on either the most probable parse tree generated by a grammar or on the probability distribution over all generated parse trees. For a more in-depth introduction to parsing with probabilistic grammars, see Section 5.1.1.

2.4.1. The Ranking Approach

One approach taken by probabilistic grammar-based models to account for human sentence processing is to predict processing difficulty and parsing preferences on the basis of a ranking of the best syntactic analyses of the input. In the *ranking* approach, the most probable syntactic analysis at each incremental processing step is predicted to be the one preferred by humans. Processing difficulty is linked to the processing effort made when a previously preferred analysis suddenly becomes dispreferred as more input is processed. Ranking models assume that a number of different analyses is entertained in parallel. Since the number of possible structures rises with the size of the grammar, human memory limitations are usually modelled using a *search beam* which contains only the most likely analyses.

The first, highly influential instantiation of a ranking model was introduced by

Jurafsky (1996). His model proposed a unified account of lexical and syntactic disambiguation on the basis of lexical and structural grammar rule probabilities. The model accounts elegantly for frequency effects on different levels of processing: A preference for the more frequent lexical category of a word is modelled through probabilistic lexicon entries which list all possible categories and the probability with which the word is realised as each. Structural preferences are captured by grammar rule probabilities. Jurafsky combines word-driven *bottom-up* information and rule-driven *top-down* information in a Bayesian reasoning system. This strategy in principle allows for the inclusion of a great number of constraints, for example semantic plausibility, but only lexical and syntactic information is considered in the examples. By making independence assumptions, the combination procedure can be simplified to multiplying the probabilities of top-down and bottom-up evidence, just as in a PCFG parse tree. Later work uses the Bayes net reasoning mechanism for combination and does not further require the independence assumption (Narayanan and Jurafsky, 1998).

The model thus takes both structural and lexical category preferences into account in computing the tree probability for each syntactic analysis. At each step in processing, the model exhaustively computes all syntactic analyses of the current input that are monotonic extensions of analyses in the search beam, but only analyses within a certain probability range are kept. Processing breakdown for difficult garden path sentences is linked to a situation where the correct analysis of the input is not contained in the search beam because it was too unlikely at some previous point in processing.

Recall the sentence *The horse raced past the barn fell* that induces processing failure in most comprehenders. In Jurafsky's model, the ultimately correct reduced relative analysis corresponding to *The horse that was raced past the barn* would be assigned only a small probability at *raced* because reduced relative clauses are relatively infrequent and because *raced* is biased towards the intransitive, active interpretation. The correct reduced relative analysis therefore drops out of the beam of accessible parses at *raced* and cannot be retrieved any more when *fell* is encountered. The model, like most readers, therefore cannot assign the sentence a correct syntactic analysis any more.

Thus, Jurafsky's model accounts both for human parsing preferences and for processing breakdown by one mechanism, which involves constructing probabilistic sentence analyses, predicting the most likely one to be preferred and discarding the least likely ones. One obvious restriction of this model is that it does not account for processing difficulty with phenomena that do not lead to complete processing breakdown, or for processing difficulty in unambiguous sentences (for example center-embedded relative clauses). Also, the grammar covers only a fragment of English, while the rule probabilities are established from a range of corpora. This introduces the danger of finding inconsistent preferences (see, e.g., Roland and Jurafsky, 1998).

Another ranking model, the ICMM (incremental cascaded Markov model) account described in Crocker and Brants (2000), avoids the problem of restricted coverage and makes predictions also for ambiguities that do not cause processing breakdown. The

model is an incremental probabilistic parser which extracts both its grammar and the rule probabilities from the Penn Treebank corpus data. This ensures wide coverage of language data and robust processing of unseen data. Input is processed by a sequence of Markov models, a type of probabilistic sequence model. Beginning at the word level, the first model assigns sequences of lexical categories to the input words. A second model determines probable sequences of syntactic categories given the lexical categories, and so on. Processing is fully parallel, such that all possible parse trees could in principle be considered, but a search beam pruning mechanism is used to improve runtime while hardly compromising accuracy (Brants and Crocker, 2000). This beam is however not used in the prediction of processing difficulty, as in Jurafsky (1996) above. Instead, difficulty is predicted if the preferred analysis changes from one processing step to the next, a situation generally termed termed a *flip*. This prediction function also accounts for difficulty with ambiguity phenomena that do not cause processing breakdown, since it allows for a new preferred analysis to be entertained after a flip has occurred and only punishes the change in preferred interpretations.

The ICMM achieves good accuracy and coverage on completely unseen sentences from the Penn Treebank. This behaviour accounts for the ease with which humans comprehend the vast majority of utterances, while the model also makes correct predictions of human behaviour for a variety of psycholinguistically interesting phenomena (Crocker and Brants, 2000). It correctly predicts the preferred analyses for lexical category ambiguities, due to the probabilistic assignment of lexical categories during parsing. It also proves able to model structural and verb preferences through syntactic rule probabilities encoded in the grammar.

2.4.2. The Probability Distribution Approach

A second type of probabilistic parser-based model, the class of *probability distribution* models, has a slightly different focus from the other models reviewed here, as it predicts not ambiguity resolution preferences, but processing load on the reader, which is then linked to elevated reading times. Probability distribution models monitor the incremental changes in the probability distribution over all parses for the input to predict cognitive load, assuming fully parallel processing and making no predictions about preferred structures.

The model proposed in (Hale, 2001) computes the total probability of finding any parse at all for the current (partial) input string, the string's *prefix probability*. This probability is assumed to be inversely related to the cognitive load encountered in processing the input string: Processing a highly probable string causes little load, processing an improbable one causes much more. Hale (2001) defines the cognitive load spent on processing a particular word as the ratio of the prefix probability of the input before seeing the word over the prefix probability of the input including the word. The logarithm of this measure, termed the *surprisal*, thus captures the amount

of change in the set of possible parses in terms of the total probability of the set. High surprisal is assumed to be mirrored in longer reading times. If, for example, a word is encountered that causes a previously highly probable analysis to become impossible, that analysis will drop out of the set of possible parses and the prefix probability of the input string will be much lower than before, causing high surprisal for the word in question and making a similar prediction to the ranking models. However, high surprisal can also be caused by the abandoning of a large number of relatively improbable analyses, a case in which ranking models would not predict difficulty. Hale's model is thus sensitive to the amount of restriction on the set of alternative analyses that each word places and does not resort to beam search or explicit reanalysis.

Hale demonstrates his approach using a small, hand-written grammar with rule probabilities extracted from the Penn Treebank. Levy (2005) shows that surprisal as computed by a wide-coverage probabilistic parser using a grammar and rule probabilities extracted from a syntactically annotated corpus of German (Skut, Brants, Krenn, and Uszkoreit, 1997) correctly predicts processing difficulty observed in reading German verb-final clauses. In addition, Levy derives an equivalent information-theoretic formulation for the surprisal measure which is defined more generally over input string probabilities instead of prefix parse probabilities, which always depend on a given grammar. The particular grammar used for parsing is reduced to a mere source for deriving the string probability and is equivalent to any other grammar that defines the same probability distribution over strings. In this way, the predictions of the surprisal approach become independent from any particular grammar formalism.

A related model, the *Entropy Reduction Hypothesis* (Hale, 2003) relies on measuring the uncertainty about the interpretation of a given partial input at any moment in processing. This uncertainty is measured in terms of the entropy of the probability distribution over all possible analyses. A distribution with many equally probable parses shows a larger entropy than one which clearly favours a single parse. The intuition that a tie between several analyses should cause processing difficulty is similar to the approach of constraint-integration models (which however still aim to identify a preferred analysis from the set of possible alternatives). Since entropy reduction is not bound to ambiguity, Hale's model is able to account for processing difficulty in unambiguous sentences.

Again, processing load per word is computed by the ratio of the entropy at the last processing step over the entropy of the current processing step. The Entropy Reduction Hypothesis predicts that any uncertainty reduction is linked directly to reading times, with large uncertainty reductions causing longer reading times due to higher processing load.

Hale (2003) demonstrates that this model correctly predicts a difference in reading time for two related syntactic structures that was found by Sturt, Pickering, and Crocker (1999), and models the processing difficulty associated with different types of relative clauses as well as predicting processing difficulty with multiple central embeddings.

2.4.3. Integration of Plausibility

Since all parser models rely only on probabilistic grammars and structural frequency information to make predictions about human sentence processing, neither the ranking nor the probability distribution models reviewed so far incorporate a notion of plausibility. This shortcoming is addressed explicitly in work by Narayanan and Jurafsky (2002), which integrates syntactic and semantic factors through a combination of Bayesian belief nets (a formalism for reasoning about events based on partial probabilistic information). The Bayesian architecture incrementally integrates a probabilistic grammar-based parsing model with lexical preferences and thematic fit information in a mathematically clean and consistent way. The parser is cast as a set of belief nets representing the possible syntactic analyses, which have to be pre-specified. Each net computes the structural probability of the parse it represents, incrementally updating its estimate as more input is encountered. A second belief net integrates thematic fit and lexical (verb tense/voice and valence) probabilities to compute the support for each alternative analysis from lexical and semantic evidence. The predictions of the nets are combined into a single probability value for each structure. Processing difficulty is predicted using the flip measure, by predicting a constant amount of difficulty if a previously preferred syntactic analysis becomes no longer tenable with the next word of the input.

The Narayanan and Jurafsky model thus demonstrates how to cleanly integrate semantic information into a probabilistic parser model. However, the model still encounters a methodological difficulty, because, as for constraint-integration models, the sources of information to be integrated have to be chosen by hand for each new phenomenon that is to be modelled. Also, the approach inherits several weaknesses of the Jurafsky (1996) model. One is that the necessary conditional probability values usually are determined from a range of different resources, for example different corpora or, in the case of semantic plausibility, human judgements elicited in rating studies. The approach therefore runs the risk of incorporating conflicting preferences if corpus resources with different or even conflicting biases are used.

Finally, wide coverage of the model is compromised by two restrictions: First, again as in the Constraint-Integration model, costly human judgements are used to model the influence of plausibility. Another problem inherited from the original model is that a grammar with restricted coverage is used. Since the parser implementation assumes that Bayes nets corresponding to all relevant syntactic analyses are available before processing can begin, it would be hard to use this model for processing free text. While addressing the lack of plausibility information is parser models, the Narayanan and Jurafsky approach thus lacks wide coverage, another vital requirement for a model of human sentence processing.

2.4.4. Summary

In sum, probabilistic grammar-based models are well suited to model processing by a fully-developed language processing system. Those models that extract both grammars and rule probabilities from large corpora exhibit wide coverage of language material and robust processing of unseen data, just like the human language processor.

Rank and probability distribution model differ for example in their difficulty predictions for unambiguous phenomena which induce processing difficulty. Ranking models only predict processing difficulty for the processing of ambiguous input because they rely on a change in the preferred analysis of the input, which does not take place in unambiguous contexts. Probability distribution models also can account for difficulty in unambiguous constructions, for example in cases of multiple center-embeddings of relative clauses.

Another important difference between the probability distribution and ranking approaches to modelling is that the ranking approaches more easily account for how a semantic interpretation of the input is incrementally constructed during parsing. Since the ranking approaches predict one syntactic structure to be the preferred one, the semantic interpretation of the input can simply be assumed to be the interpretation licensed by the preferred structure. This interpretation is either maintained and extended with further input, or has to be abandoned and replaced if the underlying syntactic analysis becomes untenable. Probability distribution approaches cannot naturally explain the construction of a semantic interpretation of the input in this way, because they do not pay attention to individual parses. The argument made in Levy (2005) even assumes that the set of possible parses does not need to be enumerated as long as the corresponding probability distribution can be inferred.

A crucial deficit of most probabilistic grammar-based models, be they ranking or probability distribution models, is that they do not incorporate a notion of the semantic plausibility of the analyses they arrive at. While connectionist models can learn an approximation of a word's meaning through its distribution in the training data or through direct association with a semantic representation and constraint-integration models specify plausibility constraints, probabilistic grammar-based models focus their predictions only on syntactic analyses of the input. The single exception is the model described in Narayanan and Jurafsky (2002), which integrates thematic fit information with a probabilistic parser. However, we noted above that this model encounters methodological problems and does not achieve wide coverage of naturally occurring language data, which is generally a strength of probabilistic parser models.

2.5. The SynSem-Integration Model: A Wide-Coverage Model of Syntactic and Semantic Preference

In Chapter 1 we identified four desiderata for a model of human sentence processing: wide coverage, incrementality, an experience-based, probabilistic framework and the integration of semantic plausibility. The review of existing models of human sentence processing above has demonstrated that while all models are capable of incremental processing, and to varying degrees rely on corpus frequencies for parameters, only some explicitly account for the influence of plausibility, and there is no model to date that simultaneously satisfies all four desiderata. We have argued that connectionist models are well suited to modelling the language acquisition process through their ability to learn syntactic and semantic properties of language from large amounts of input data. However, connectionist models capable of wide-coverage processing of realistic utterances are extremely complex and therefore opaque with regard to their predictions. Also, learning reliable generalisations about the plausibility constraints on a realistic vocabulary requires very large amounts of training data as well as large amounts of training time.

Constraint-integration models distinctively include plausibility constraints for modelling, and usually determine the strengths of their probabilistic constraints from corpora. However, no wide-coverage implementation has been proposed. The majority of theoretical proposals require an individual set of constraints to be chosen by hand for each phenomenon, which makes it hard to achieve wide coverage of phenomena (or natural utterances) with any one model instance. An alternative implementation avoids these problems by relying on a connectionist network for constraint selection and strength setting, but it also inherits scalability problems from connectionist networks.

Finally, probabilistic grammar-based models can easily achieve wide coverage by estimating all their parameters from syntactically annotated corpora. They are also extremely transparent with regard to the reasons for their predictions, but to date lack any treatment of semantic plausibility. The one parser model which allows the integration of plausibility information (Narayanan and Jurafsky, 2002) does not have wide coverage because the set of admissible structures must be modelled individually. In addition, we have found that those constraint-integration and parser models that do integrate semantic plausibility rely on human judgements (Narayanan and Jurafsky, 2002, McRae et al., 1998); no model exists of human plausibility intuitions.

We therefore propose the SynSem-Integration model, which combines an incremental, probabilistic grammar-based syntactic model with a semantic model of human plausibility estimates trained on corpus data. As shown in Figure 2.3, this semantic model computes the plausibility of the verb-argument relations in each syntactic structure proposed by the syntactic model. The plausibility predictions are integrated with

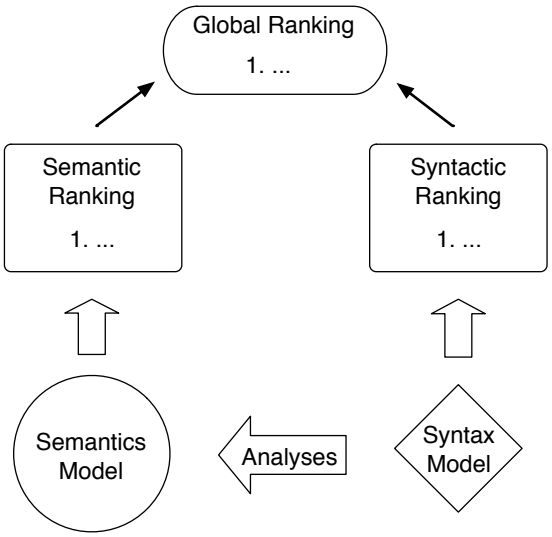


Figure 2.3.: The architecture of the SynSem-Integration model

the syntactic probability of the structures to determine the *globally preferred structure*. The syntactic and semantic models also each identify a preferred structure given the ranking of structures by their respective evaluation measure. Predictions of difficulty in human sentence processing are made by transparent cost functions defined over the preferences of both models in relation to the globally preferred structure.

The SynSem-Integration model thus overcomes the limitations of pure parser models by integrating an explicit source of semantic plausibility estimates. At the same time, the model does not require the stipulation of constraints, and the vast majority of parameters in the syntactic and semantic models can be learnt automatically from corpus data (Chapter 3 discusses the setting of the remaining parameters for the semantic model). Learning both the semantic and the syntactic model from corpora also gives the SynSem-Integration model broad coverage both of structures that are processed effortlessly as well as those that cause disruption. This allows the model to cover a range of different psycholinguistic phenomena without requiring modifications. Finally, the SynSem-Integration model operates strictly incrementally, integrating each word into the syntactic representation immediately and making plausibility estimates as soon as a new verb-argument pair is encountered. It therefore fulfils all four desiderata identified in Chapter 1.

The syntactic model is instantiated by an off-the-shelf incremental PCFG-based parser using a grammar and rule probabilities extracted from a large corpus. The

syntactic model is used to incrementally predict the possible syntactic analyses of the input and their probabilities. We follow the ranking approach to probabilistic grammar-based modelling and rank the explicitly enumerated parse alternatives by their syntactic probability.[1] The semantic model has the task of evaluating the plausibility of each of the syntactic parse alternatives and ranking them by their plausibility scores. Section 2.5.1 discusses how we propose to construct such a model and account for human judgements of semantic plausibility based on language data in a large corpus. Section 2.5.2 describes when and how cost predictions are made on the basis of syntactic and semantic preferences.

2.5.1. Modelling Semantic Plausibility

Within the SynSem-Integration model, the role of the semantic model is to evaluate the semantic plausibility of the syntactic analyses of the input proposed by the syntactic model. Since a full evaluation of all semantic and pragmatic meaning aspects is clearly impracticable, we restrict the scope of the semantic model to the level of verb-argument relations (which we operationalise below) in the input. This level furnishes the basic who-does-what-to-whom information necessary to roughly evaluate and compare the plausibility of the described event. Further, the verb-argument level of analysis corresponds to the level of typical plausibility manipulations in experimental psycholinguistics, where plausibility is usually operationalised as the fit of an argument to a verb in a specified relation. Finally, recall that the task of the semantic model is to compare the plausibility of different analyses of the same input string. This means that lexical material is the same in all analyses and that relevant differences in plausibility between the analyses are caused by different attachment decisions that may lead to the stipulation of different verb-argument pairs.

Thematic Roles Instantiate Verb-Argument Relations

Several possible instantiations for verb-argument relations could be considered. One is the *grammatical function* that characterises the syntactic relationship between a predicate and its argument. This approach has been used by models of verb selectional preference, which solve a related task to our semantic model (e.g., Resnik, 1996, Clark and Weir, 2002, Li and Abe, 1998), to approximate verb-argument relations. This instantiation is maximally close to the syntactic analysis that yields the verb-argument pairs in the first place. However, while grammatical functions are a source of information about the verb-argument relation, they do not yield enough semantic generalisation to be useful for the evaluation of plausibility. The existence of regular verb diathesis (see Levin, 1993) demonstrates that the syntactic relationship between two words does not

[1] For details about the implementation of the syntactic model, see Section 5.1.2.

determine their semantic relationship. Instead, the same semantic relationship between a verb and its argument can often be expressed by a range of syntactic relationships. Grammatical functions are thus often indicative of a semantic relation, but they do not define the relation unambiguously.

We choose *thematic roles* instead to describe the relationship between a verb and its argument. Thematic roles characterise the semantic nature of the syntactic relationship between the assigning verb and the receiving argument or adjunct. Thus, they are both defined close to the observed syntactic relationships and introduce a semantic interpretation of the relationship between a verb and its argument.

Thematic roles have long been relied upon as a pivotal link between syntactic and semantic processing in psycholinguistics. Carlson and Tanenhaus (1988) for example gave an influential characterisation of thematic roles as a level of preliminary semantic analysis that allows the processing system to make fast semantic commitments that can be retracted at relatively low cost if they later turn out to be incorrect. A number of theories and models of sentence processing like Pritchett (1992) or Crocker (1996) even explain the human processor's strategies from a desire to interpret the input by assigning thematic roles as early as possible, attributing processing difficulty of different severity to the re-assignment of thematic roles in different situations.

Using thematic roles, the semantic plausibility model proceeds as follows to evaluate the plausibility of different sets of verb-argument relations corresponding to different syntactic analyses: First, the syntactic relationships in each set are semantically characterised by thematic roles. Then, the plausibility of each verb-argument-role triple is evaluated by estimating the goodness-of-fit of the argument as a bearer of a specific thematic role assigned by the verb. The more plausible syntactic analysis is the one that gives rise to the more plausible verb-argument-role triples.

A Probabilistic Model

To support wide coverage in the SynSem-Integration model, the semantic model should cover as many verb-argument relations as possible after a single training session, and should not require retraining for individual sets of test data. Therefore, the semantic model cannot rely on human judgements as do many existing models that integrate semantic information. Eliciting enough human judgements to allow the model acceptable coverage is far too costly. Rather, we propose a probabilistic model of human plausibility intuitions that is estimated from corpus data. The probabilistic approach allows fast and cheap estimation, while yielding a model that is experience-based and has far broader coverage of verb-argument relations than what can be reasonably achieved with human judgement elicitations. Two large corpora annotated with thematic roles exist: FrameNet (Ruppenhofer, Ellsworth, Petruck, and Johnson, 2005) and the Proposition Bank (PropBank, Palmer, Gildea, and Kingsbury, 2005). We compare both corpora as training data in Chapter 3.

The doctor cured the patient	Verb lemma: *cure* Argument lemma: *doctor*	Verb sense: *healing* Grammatical function: subj

Figure 2.4.: Factors in the semantic model

The semantic model will be used to identify all possible thematic roles that can link a given verb-argument pair in sentence context, estimate the plausibility of all verb-argument, and assign the most plausible thematic role as the humanly preferred one. This thematic role and its plausibility estimate are the basis for the comparison between the overall plausibilities of different syntactic analyses of the input material. We identify a number of factors that are relevant for the tasks of identifying possible thematic roles and estimating their plausibility. As an example, consider the sentence *The doctor cured the patient*.

- **Verb and argument head lemmas** The identity of the verb and the argument head, *cure* and *doctor*, are obviously important to determining their relationship and its plausibility.
- **Verb sense** Polysemous verbs normally assign a completely different set of thematic roles in each of their senses. For example, the *preserving* sense of *cure* assigns roles that can be paraphrased as *preserver* and *preserved*, while the *healing* sense assigns a *healer* and a *healed*. Therefore, it is the verb sense rather than the verb lemma that finally determines the thematic roles applicable to syntactic relations.
- **Grammatical function** The grammatical function that links verb and argument does not determine the semantic relation between verb and argument, as argued above, but is an important factor in inferring the intended thematic role: *The doctor* in the syntactic subject position of *cured* is intended to fill a different role than if it is realised as a syntactic object.

As an example, the complete set of factors for the relation *doctor–cure* in the sentence *The doctor cured the patient* is listed in Figure 2.4.

We propose to equate the plausibility of a verb-argument-role triple with the probability of seeing the thematic role with the verb-argument pair. This approach ensures that the model's plausibility estimates reflect the observed frequency distribution in the training corpus. Modelling plausibility of events through language data implies some crucial assumptions, which we discuss below.

We formulate a generative model of the probability of assigning a thematic role to a verb-argument pair. Generative models attempt to estimate the joint probability

distribution that underlies (or *generates*) the co-occurrence of the input factors (here, the four influential factors we have identified) and the output factors (in our case the thematic role). This enables them to predict missing input or output values on the basis of the joint distribution, a property we use both for the prediction of thematic roles and the treatment of missing input values (see discussions below and in Section 3.1.4).

The plausibility of seeing a thematic role with a verb-argument pair is thus computed as the joint probability of seeing together the argument head a, the verb v in its sense s, the grammatical function gf of a and the role r in Equation 2.1.

$$Plausibility(v, r, a) = P(v, s, a, gf, r) \tag{2.1}$$

To predict thematic roles for verb-argument pairs, we use the model to enumerate all possible thematic roles for the verb-argument pair and their plausibilities and choose the most plausible one, as in Equation 2.2.

$$assigned\ role(verb, argument) = \underset{role}{argmax}\ Plausibility(verb, role, argument) \tag{2.2}$$

We determine the plausibility of a syntactic analyses by assuming independence between the different verb-argument pairs it gives rise to and multiplying the plausibility estimates for all pairs. This assumption is necessary due to data sparseness, but it is overly strong, as there is a dependency between thematic roles assigned to different arguments of the same verb. We take this fact into account in the implementation of the model by positing the constraint that each role can be assigned only once, and then optimising the probability of the set of assigned roles given this constraint (see Section 5.2.1 for details).

The plausibility model specified in Equation 2.1 has a number of properties worth discussing. First, it allows us to deal easily with missing variable values, because it is an instance of a generative model. Treating unknown values appropriately is important, since the values of the five model variables are never all known when a plausibility prediction has to be made. For example, the appropriate sense of a polysemous verb is never specified in the input to the semantic model. Since the verb sense determines the set of thematic roles that the verb can assign, the verb sense has to be disambiguated to make plausibility predictions. The generative formulation of the model allows us to elegantly incorporate this disambiguation task into the prediction process: We allow the model to generate role and plausibility predictions for all applicable verb senses and for each thematic role choose the sense value that leads to the highest plausibility prediction. The verb sense used to predict the most plausible role is the one the model assumes to be appropriate for the verb-argument pair.

Second, the model operates incrementally over verb-argument pairs, which is a precondition for its use in the SynSem-Integration model. As soon as an analysis from the

syntactic model contains at least one verb-argument pair, the plausibility model assigns each pair a thematic role and plausibility estimate. Through the independence assumption between pairs, the plausibility of each partial syntactic structure is computed incrementally. The model as defined here has a tendency to assign lower probability values to sentences with more verb-argument pairs. This runs counter to the intuition formalised in several theories of sentence processing that the human processor is eager to make as many role assignments as possible (see, e.g., Pritchett, 1992). This model bias is addressed in the implementation by a normalisation procedure (see Section 5.2.2).

Modelling Plausibility Using Corpora

In our discussion of modelling human plausibility intuitions, we have so far identified the level of verb-argument-relation triples as the appropriate one in the context of modelling experimental results on human sentence processing. We have also determined thematic roles to be the best way of defining the relations between verb and argument, and have argued for the use of a probabilistic model. To induce such a model, corpora annotated with thematic role information can be used.

However, the fundamental assumption of this approach is that plausibility information can be modelled using frequency patterns in linguistic utterances. This assumption is certainly justified if the plausibility of a verb-argument-role relation is a linguistic property deriving from the admissibility of using the verb and argument in the given role. In this case, information about acceptable combinations can be derived from corpus data just as information about acceptable syntactic structures can be derived for a model of syntax. At least some aspects of the plausibility of verb-argument-role triples are certainly linguistic, for example the set of thematic roles that a verb may assign, and its preference for how to realise these roles syntactically.

However, it is unclear to which degree the plausibility of a specific argument, such as *doctor*, as a role filler for a verb, such as the *healer* role of *cure*, is determined linguistically and to which degree world knowledge is necessary to predict the plausibility of the described event. It appears likely that the larger share falls to world knowledge, because even frequency patterns observed in language use such as the existence of typical role fillers are to a large extent caused by speakers' frequent references to plausible real-world events and therefore reflect world knowledge. Therefore, we have to assume that some amount of world knowledge in addition to linguistic knowledge is necessary to predict the plausibility of verb-argument-role triples.

In this case, using corpora to train a probabilistic model of the plausibility of verb-argument-role relations is useful only to the degree to which corpus frequencies of verb-argument-role triples reflect the plausibility or frequency of events in the world. We cannot assume the parallelism between the frequency of events and the frequency of utterances about events to be perfect, because humans usually make utterances with the goal of communicating information to a hearer. Infrequent events may be perceived

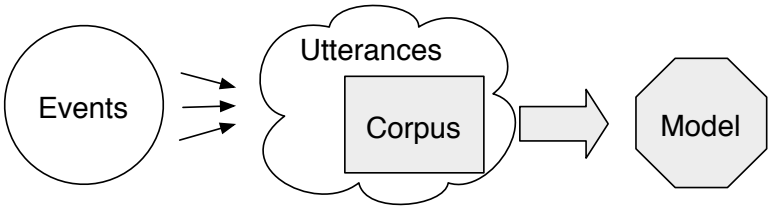


Figure 2.5.: Modelling the plausibility of events via the frequency of utterances about the events, as represented in a corpus.

as more informative or interesting and therefore more worthy of being communicated, which may cause them to be discussed disproportionally more often than they are experienced. By the same logic, frequent events may be perceived as less newsworthy and therefore be mentioned less often than they occur. Also, while we can choose a corpus resource that is as balanced as possible with regard to the provenance and genre of utterances, we can never be sure that the corpus frequencies are a reliable sample from the distribution of all utterances.

However, in addition to the intuition that information such as the existence of typical role fillers reflects facts about the plausibility of real-world events, there is some experimental evidence that corpus data is indeed a useful training source for a plausibility model: A rating study investigating the plausibility of verb-argument-relation triples (see Section 4.1.2) shows that triples that were seen in a corpus are rated as significantly more plausible than unseen triples made up of the same verbs and arguments, but an admissible unseen relation. This suggests that a plausibility model can make useful predictions if it assigns a higher probability to verb-argument-relation triples that are seen or even typical in the training data than to triples that are atypical or not attested at all.

In sum, it appears that a plausibility model trained on corpus data should be able to capture both linguistic and extra-linguistic information necessary to predict the plausibility of a verb-argument-role triple. Figure 2.5 illustrates our modelling approach: In order to model the plausibility of events in the world, we rely on both linguistic and extra-linguistic information encoded in utterances made about the events. However, we do not base our plausibility model on the complete set of utterances ever made about any event, but rather use a corpus, which samples a subset of the utterances. The model uses linguistic knowledge such as the set of roles a verb may assign like a filter that rules out verb-argument-role triples that cannot be expressed in the language, such as *cure-doctor-ingestor*, while the knowledge about plausible real-world events induced from utterances about the world allows the model to predict the respective plausibility

of linguistically admissible verb-argument-role triples such as *cure-doctor-healer* and *cure-patient-healer*.

Estimating the Semantic Model

We estimate the joint-probability model introduced above from the training data using Maximum Likelihood (ML) estimation, which defines the probability of an event as its relative frequency in the training corpus (see, e.g., Manning and Schütze, 1999). However, this approach immediately encounters a problem: The model contains five variables (verb, sense, argument head, role and grammatical function), most of which have a large number of different values. It is clear that the variable co-occurrences observed in any training corpus can only cover a small amount of all possible combinations of these variables. While many of the unseen cases will be "grammatical zeroes" reflecting implausible combinations, the semantic model will also be unable to make predictions for a large number of plausible unseen cases.

The semantic model therefore requires a smoothing strategy that allows it to infer plausibility estimates even for unseen input combinations. Experience-based models of syntax alleviate the data sparseness problem by abstracting away from lexical items to abstract categories like part of speech or phrase type. ML estimates are then made on the basis of abstract categories instead of lexical items. This *class-based smoothing* approach, simplified, allows syntactic models to estimate the probability of seeing a noun phrase in the subject position of a verb phrase, instead of estimating directly how likely it is that *doctor* occurs as a subject of *cure*. In this way, syntactic models pool similar observations. This allows them to base their predictions on a broader data basis, making them more robust and reliable.

For a model of semantic plausibility, class based smoothing can be employed in a similar way. The generalisations we are interested in are necessarily semantic, however. Grouping together words with similar meaning (and similar syntactic behaviour, in the case of verbs) enables the semantic model to pool co-occurrence information and use frequent observations to make inferences about the behaviour of infrequent class members. In computational linguistics, for example, models of selectional preferences that estimate the goodness-of-fit between a verb and its argument in a specified relation use noun classes for smoothing (see Section 4.3).

From a cognitive perspective also, using semantic classes for making generalisations about verb-argument plausibility is a natural approach, since semantic categories are a much-researched basic tool for human reasoning about the world (see, e.g., Medin and Aguilar, 1999, for a short, high-level overview). Human semantic categories group together words that are similar by criteria relevant to the reasoning task at hand. The exact grouping mechanisms are still controversial: Are human semantic categories defined by similarity, or is similarity between concepts a sign of a shared category? It is clear, however, that classes of similar words play a large role in human reasoning about

the world and therefore presumably also in making plausibility judgements. Class-based smoothing can therefore be seen as an approximation of this human strategy, although we restrict ourselves to one constant set of word classes and do not attempt to re-define semantic classes for each prediction task.

There is evidence for the existence of semantic classes also at another level of human cognition, namely as an organisational principle of the human mental lexicon. Evidence from *priming* studies, which measure the processing speed of a target word after a prime word has been presented, suggest that the representations of words and their meaning are grouped in semantic fields, such that the presentation of a prime from the same semantic field as the target facilitates processing. This evidence is supported by clinical studies which show that semantic fields can be selectively damaged or spared in patients whose language faculty is partially impaired. Within these semantic fields, words are assumed to be linked through conceptual similarity (including superordination and antonymy), frequent co-occurrence or shared word class. The stronger these links are, the more the words are evocative of one another in production (association naming) and perception (priming) (see, e.g., Aitchison, 2003, for an overview). A set of conceptually similar words of the same word class therefore can be assumed to be connected by strong associational links in the mental lexicon and therefore be evocative of one another to comprehenders.

Counteracting data sparseness with semantic classes is thus justified on the basis of both computational and cognitive criteria: From a computational point of view, grouping words together into classes allows for more robust probability estimates than looking at each word individually. In cognition, semantic categories are pervasive both in the organisation of the human lexicon and in reasoning about the world, which is a task similar to the one set for the semantic model. Chapter 3 is dedicated to further discussing smoothing methods and selecting the best smoothing regime for the semantic model.

2.5.2. Predicting Difficulty

In isolated sentences with syntactic ambiguities, processing difficulty may be observed in two regions: During the processing of an ambiguous region, where several syntactic analyses of the input are possible, and at the point of disambiguation towards one of the alternative analyses. We attribute the difficulty observed in these two situations to two separate effects: *Conflict* and *revision*. These explanations for processing difficulty have been identified previously by constraint-integration and ranking parser models, respectively. Constraint-integration models attribute all occurrences of processing difficulty to a conflict of constraints, while ranking parser models only predict difficulty due to a revision of the preferred structure. We argue here, however, that the factors are complementary to one another in the context of our sentence processing model and that only a combination of both leads to the correct prediction of difficulty.

- **Conflict** arises during the processing of an ambiguous region if there is conflicting evidence concerning which syntactic analysis to prefer. In these situations, either the syntactic or the semantic model does not agree with the globally preferred structure. Since conflict does not necessarily lead to a change in preferred structure, the revision measure does not account for this source of processing difficulty.

- **Revision** occurs at the point of disambiguation if the analysis that was preferred during the ambiguous region is rendered impossible by the disambiguation. In contrast to parser-based models, we define revision as a semantic process, which allows us to abstract away from the exact syntax of the preferred analyses.

Difficulty due to revision can be captured by a conflict-based measure if it is assumed that the evidence for the previously preferred, but disconfirmed structure continues to be available. This evidence then conflicts with the strengthened evidence for a previously dispreferred analysis. However, at a point of disambiguation, this is not always the case in our model. Recall that the SynSem-Integration model operates on the set of possible syntactic analyses of the input. If syntactic disambiguation rules out the preferred analyses of the previous time step, its semantic interpretation is often no longer available to compete with the interpretation of the confirmed alternative analysis. This ensures that every semantically preferred analysis is also syntactically compatible with the input. Therefore, the conflict measure alone cannot account for difficulty due to revision, because at the point of disambiguation, only one possible analysis of the input may remain.

In sum, we attribute processing difficulty to two separate factors, neither of which suffices to predict difficulty on its own. The occurrence of one of these factors does not necessarily exclude the occurrence of the other, however: It is possible in the SynSem-Integration model that conflict and revision co-occur, whenever the situation that leads to the revision of the globally preferred structure does not completely rule out the previously preferred structure and leaves its semantic interpretation to compete with that of the new preferred structure. In such cases, we treat conflict cost and revision cost as additive.

The SynSem-Integration model predicts difficulty based on these two factors in the following way: First, it computes the globally preferred syntactic analysis, which is predicted to be the one that humans assume. This analysis is found by interpolating the syntactic and semantic evaluation for each candidate syntactic structure predicted by the syntactic model, and by ranking the structures according to the resulting goodness score. Cost is predicted according to two cost functions: *Conflict cost* is incurred when the preferred structure predicted by the syntactic or semantic model conflicts with the globally preferred structure. *Revision cost* is predicted when the interpretation of the globally preferred structure changes non-monotonically. The SynSem-Integration

model's final cost prediction for the processing of an input region is the sum of all conflict and revision cost predicted for the test stimuli in this region, normalised over the number of stimuli. In Chapter 5, we discuss different implementations of the two cost functions and select the best-performing ones as well as the best-performing interpolation factor on a development set.

In predicting cost, we have a choice between three levels of granularity for difficulty predictions: *Qualitative* predictions are binary flags for the existence of difficulty (for example, as in ranking parser models), *near-quantitative* predictions specify the relative size of processing difficulty (as in constraint-integration models), and truly *quantitative* predictions directly link a model's output to reading times in milliseconds.

In practice, quantitative predictions are hard to make since a number of factors like word length, word frequency and predictability also influence reading times (Just and Carpenter, 1980, MacDonald and Shillcock, 2003). Qualitative predictions, on the other hand, carry only a limited amount of information. The SynSem-Integration model's predictions are therefore defined to be near-quantitative. This is achieved by summing predictions for individual stimuli and normalising by the number of all stimuli. In consequence, the model can make near-quantitative predictions even if the individual difficulty predictions are made on a qualitative level, because the overall prediction depends also on the number of stimuli for which difficulty is predicted: Conditions with few qualitative difficulty predictions are predicted to be easier to process than conditions with many qualitative difficulty predictions.

2.5.3. Cognitive Claims

The strength of implemented models of human sentence processing is that they specify exactly how each aspect of sentence processing is modelled and therefore make testable predictions. However, the process of creating a fully specified model is bound to lead to some design choices that serve simply to create a running implementation, but do not have the same status as the architectural claims that underlie the model and that are to be tested. In this section, we discuss which properties of the SynSem-Integration model make claims about the architecture of the human sentence processing system, and which are due to implementational design decisions.

Architecture The implemented SynSem-Integration 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 output of the syntactic model. This modular architecture is an implementational choice, not a claim about the architecture of the human processing system. What the SynSem-Integration model does claim is that the input is analysed with regard to its syntactic structure and semantic plausibility, and that the degree of consistency between the results of these analyses influences processing ease. We do not address the question whether this analysis takes

place in a modular or integrated fashion in the human brain. Note, however, that even given no explicitly modular architecture it is still plausible that semantic analysis takes place on the result of syntactic analysis, under the assumption that syntax serves to communicate semantics.

Mechanisms underlying syntactic and semantic processing The syntactic and semantic models are derived from annotated corpus data and operate on the assumption that phenomena frequently encountered in corpora are preferred by the human sentence processor in its incremental construction of a syntactic and semantic interpretation of the input. The underlying assumption of all probabilistic experience-based models, and therefore also of the SynSem-Integration model, is that human syntactic processing takes a similar approach. This is motivated by the overwhelming evidence for frequency effects in sentence processing, for which an experience-based architecture can elegantly account.

However, we do not claim that humans evaluate the semantic plausibility of verbs and their arguments only on the basis of their linguistic experience. Rather, we use the frequency of utterances about an event as an approximation of the typicality of real world experiences, assuming that there is a strong link between the frequency and therefore plausibility of a real-world event and the frequency with which it is mentioned in utterances about the world. The implementation of the semantic model therefore should be understood as a modelling device, not a claim about human reasoning about the world.

Difficulty prediction mechanisms The SynSem-Integration model makes predictions by determining a globally preferred parse and then evaluating the semantic and syntactic models' output with regard to this parse. This mode of operation also contains some assumptions about the mechanisms involved in human sentence processing. The first is that humans incrementally form a clearly defined preferred analysis of the utterances they encounter. This assumption is supported by the more basic assumption that sentence processing aims at constructing a complete semantic analysis of the input that immediately integrates each new word as it is encountered. Predicting a specific preferred analysis of the input at each processing step explains how an incremental analysis of the semantics of the input can be constructed by comprehenders: They simply rely on the interpretation of the preferred structure.

Second, the SynSem-Integration model's two cost functions make few explicit claims about the processing mechanisms of the human mind, but they are compatible with assumptions made by other models. In conflict situations, constraint-integration models attribute longer processing times to competition for activation or processing resources between two roughly equally supported analyses, which delays processing until one analysis is decided on (Spivey and Tanenhaus, 1998, Tabor et al., 1997). The competition

view can easily be applied to the SynSem-Integration model, as well: Difficulty can be attributed to the existence of a conflicting interpretation that uses up processing resources and therefore slows down processing. Note that this interpretation does not entail a change in the preferred structure.

The second source of cost, revision of the preferred analysis of the input, is compatible with the predictions of all models reviewed above. Constraint-integration models quantify this cost by predicting competition between the still highly-activated previously preferred analysis and a now better supported alternative. In this situation, we argue, however, that a disproven analysis immediately drops out of competition, and we therefore attribute difficulty more abstractly to a change in the preferred analysis like ranking parser models do. The difficulty prediction measure for connectionist networks proposed by Rohde (2002) similarly takes changes in his model's semantic interpretation into account, and probability distribution parser models base all their predictions on the amount of change in the probability distribution over all parses, which is bound to be large if a highly probable analysis suddenly drops to a very low probability. In sum, the prediction that abandoning a highly preferred analysis should incur cost is the least contentious claim about the SynSem-Integration model's cost prediction mechanism.

2.6. Summary

In this chapter, we have reviewed probabilistic models of human sentence processing. We have demonstrated that none of the existing models fulfils the four desiderata for a model of human sentence processing that we have identified in Chapter 1: Wide coverage, an experience-based probabilistic framework, incrementality and the integration of semantic plausibility.

We therefore proposed the SynSem-Integration model, which fulfils all the desiderata by integrating a wide-coverage model of human plausibility intuitions about verb-argument-thematic role triples with a ranking parser model. Two cost functions defined over the output of the models predict processing difficulty due to two different processing situations: One is the conflict cost function, which applies if the syntactic and semantic models prefer a different analysis during an ambiguous region. The other, revision cost, applies if the semantic interpretation of the input changes between processing steps, for example at the point of disambiguation.

As a precondition for the SynSem-Integration model, we also proposed a probabilistic, experience-based model of the semantic plausibility of verb-argument-thematic role triples. The semantic model aims to capture the frequency and thereby plausibility of events in the real world by estimating the frequency of utterances about these events, collected in a corpus.

The SynSem-Integration model shares characteristics with two classes of compu-

tational models: It incorporates a parser model and uses the idea of difficulty being caused by revision of the preferred analysis, but on the other hand, the conflict cost function borrows from the idea of processing difficulty due to conflicting constraints in constraint-integration models. The resulting model is most similar to constraint-based architectures, in that it combines the preferences of the syntactic and semantic models at each processing step to arrive at a globally preferred analysis and to predict processing difficulty.

The SynSem-Integration model however avoids the drawbacks of both parser models and constraint-integration models. By incorporating semantics, but retaining wide coverage, it extends parser models without compromising their advantages. By specifying models that can be to a large part automatically estimated from corpus data, the SynSem-Integration model avoids problems with selecting and setting constraints for individual constructions and is able to account for a number of different phenomena with a single model instance.

We finally have argued that the architecture of the model and the claims it makes about the human sentence processor are compatible with experimental findings about human processing.

3. Estimating the Semantic Model

In Chapter 2, we have motivated the SynSem-Integration model of human sentence processing that incorporates a wide-coverage, probabilistic model of human intuitions about verb-argument plausibility. This chapter focuses on this plausibility model, which we term the *semantic* model to contrast it with the syntactic component of the SynSem-Integration model. In this chapter, we discusses in detail how the semantic model presented in Section 2.5.1 is best estimated. This question requires special attention, since any attempt at deriving the model from training data without applying smoothing techniques immediately faces a severe sparse data problem. We discuss the origins of this problem in Section 3.1 and motivate the use of two orthogonal smoothing methods to alleviate it. One method uses semantic generalisations based on semantic word classes to yield more data points in estimation, while the other applies to the estimated probability distributions and assigns some probability mass to unseen combinations.

From the set of possible model instances that use the smoothing methods on their own or in combination, we select the best-performing model based on its ability to predict human plausibility judgements in Sections 3.2 to 3.5. The validation of the selected model in Section 3.6 shows that we need to combine both smoothing methods to allow the semantic model to make accurate predictions of human data with good coverage of unseen verb-argument pairs.

The final model is evaluated more thoroughly with regard to its ability to predict human data across data sets with different characteristics and compared to existing related approaches in Chapter 4. Chapter 5 describes its integration into an implementation of the SynSem-Integration model, including two extensions that allow it to process free text.

3.1. Sparse Data and Smoothing

This section discusses the semantic model's need for smoothing and the methods used to alleviate the sparse data problem. We first introduce an equivalent alternative formulation of the semantic model in Section 3.1.1, which we will use throughout the chapter. We then discuss the conceptual sources of the sparse data problem the semantic plausibility model is faced with in Section 3.1.2, and then describe the ways in which different smoothing methods contribute to alleviate the sparse data problem, and

the way they profit from the decomposed model formulation in Section 3.1.3. Finally, we discuss restrictions on our strategy of generating missing values in the context of smoothing (Section 3.1.4).

3.1.1. Alternative Formulations of the Semantic Model

Recall our proposal to estimate the plausibility of a verb-argument-role triple as the joint probability of the argument head a, the role r, the verb v in its sense s and the grammatical function gf that links v and a. For the sake of concise presentation, we collapse the variables for verb lemma and verb sense into v_s. Equation 3.1.1 repeats Equation 2.1 for convenience.

$$Plausibility_{v_s,r,a} = P(v_s, r, a, gf) \tag{3.1}$$

In this and the following chapters, we use a decomposed formulation of this model, which we derive using the chain rule. Since we do not make independence assumptions, the formulation in Equation 3.2 is equivalent to the joint formulation.

$$\begin{aligned} Plausibility_{v_s,r,a} &= P(v_s, r, a, gf) \\ &= P(v_s) \cdot P(gf|v_s) \cdot P(r|v_s, gf) \cdot P(a|v_s, gf, r) \end{aligned} \tag{3.2}$$

We make this decomposition for two reasons: First, the decomposed model formulation allows a more intuitive understanding of what kind of information about verbs, arguments and their relations the semantic model uses to make plausibility predictions. The model subterms can be interpreted as yielding linguistically relevant information about the verb-argument pair. For example, the $P(gf|v_s)$ term captures information about the verb's syntactic subcategorisation preferences when used in sense s: It reflects the probability of seeing an argument realised in each possible grammatical function. The $P(r|v_s, gf)$ term shows how the verb prefers to realise its thematic role fillers syntactically. Finally, the $P(a|v_s, gf, r)$ term is similar to the term estimated by selectional preference models, which perform a task related to that of the semantic model, as discussed in Chapter 1. The only difference is that our term is more specific: It pays attention to verb sense, and uses the thematic role linking verb and argument in addition to the grammatical function.

The second, practical reason for decomposing the joint probability model is that this formulation has advantages in combination with the application of smoothing methods, since variables are introduced in the order of their expected sparseness in the training data, from least sparse to sparsest. We discuss in Section 3.1.3 how smoothing approaches can profit from this formulation.

3.1.2. Sparse Data

The most straightforward way of estimating a probability distribution like the ones involved in the joint and decomposed semantic model formulations from training data is by Maximum Likelihood Estimation (MLE). This method estimates the distribution so that the likelihood of seeing the training data is maximised. This is done by equating the probability of each event in the training corpus to its relative frequency. To estimate the joint probability model, for example, we therefore compute the relative frequency of seeing each specific combination of an argument head, a verb in a specific sense, a grammatical function and a role in the corpus. In this way, MLE ensures that the estimated model is as faithful to the training data as possible, and no further assumptions about the shape of the estimated distribution are necessary. However, no probability mass is assigned to unseen events. This is problematic, because, in practice, the semantic model estimated with MLE suffers considerably from a lack of training data: At most 7% of verb-argument-role triples in both the development and test set used later in this chapter have been seen together during training, and therefore are covered by the unsmoothed semantic model.

There are three reasons for this large sparse data problem: One is quite simply that all corpora are limited. This means that we cannot be sure to find any given combination of a verb and argument in a corpus, and the more variables we introduce into a model (such as the thematic role or grammatical function that link them), the less likely it becomes that we will find the combination. Second, words in corpora are distributed in a Zipfian manner, so that very few very frequent words make up a large portion of the corpus, while the majority of covered words is very infrequent. Therefore, even if a word is present in a corpus, it is very likely to be represented only a few times, with few arguments. The third reason for the sparse data problem is specific to our combination of test and training data: We train the semantic model on annotated corpora of financial news text or general written language data, and test it on sets of psycholinguistic stimuli that have been constructed by experimenters to show very specific plausibility characteristics (see Section 3.2.2). We have to expect a difference in vocabulary between the two data sets, and possibly also in the type of world knowledge that they cover. Furthermore, probabilistic models generally have difficulty when training and test data come from different domains. To give one recent example for a related task, this problem was encountered for systems participating in the CoNLL shared task of semantic role assignment (Carreras and Márquez, 2005).

To counteract the sparse data problem, smoothing methods have been developed. For models of syntax, the sparse data problem is alleviated by the use of abstract categories (like part of speech or phrasal category), as discussed in Section 2.5.1. Categories allow the estimation to abstract away from the specific lexical material and pool individual lexicalised occurrences into evidence for the abstract category. In Section 3.4, we will evaluate following the same approach for semantic estimation, by classing verbs and

arguments together into more abstract groupings that represent some semantic concept. However, the use of semantic classes does not per se eliminate the problem of no or too little evidence of word co-occurrences in the training data. The model's inability to make predictions for unseen combinations of input values can be addressed if necessary by standard smoothing algorithms, which re-estimate the probability distribution determined by MLE and assign small amounts of probability to still unseen events. This allows predictions to be made even for unseen events.

While class-based and re-estimation smoothing methods also address the problem of differences between training and test data, they do not completely eliminate it: When we test the semantic model on a data set that is more similar to the training data in Section 4.1.2, we do indeed find better coverage and even more reliable predictions for that data set than for test sets that are more dissimilar to the training sets. However, evaluation results throughout Chapter 4 show that while the semantic model does better on data from the training domain, it is still able to make reliable predictions for test items from different domains.

3.1.3. Smoothing Approaches

The considerations in Section 3.1.2 have demonstrated the need for smoothing methods and have touched upon two different types of smoothing: Class-based estimation and re-estimation of the probability distribution derived from data. Class-based smoothing applies directly to the MLE process and is therefore orthogonal to the strategy of re-estimating a term's already-induced probability distribution in order to assign a small, uniform probability to unseen co-occurrences. This section discusses how the two different smoothing methods affect the proposed semantic model to demonstrate what kind of performance increases we can expect from them.

We begin by considering smoothing by re-estimation in case the target combination of verb, role, argument and grammatical function is unseen in the training data. Using only MLE, the semantic model cannot make a plausibility prediction in this case. If a smoothing method that re-estimates the MLE results and assigns a small probability to unseen events is applied, the model can assign at least a smoothed plausibility estimate. This estimate is always the same in the case of the joint probability formulation of the semantic model. However, if we use the decomposed model formulation from Equation 3.2, we can do better.

$$\begin{aligned} Plausibility_{v_s,r,a} &= P(v_s, r, a, gf) \\ &= P(v_s) \cdot P(gf|v_s) \cdot P(r|v_s, gf) \cdot P(a|v_s, gf, r) \end{aligned} \tag{3.3}$$

Equation 3.3 repeats Equation 3.2 for convenience. Recall that r denotes the role, gf the grammatical function linking verb and argument, v_s the verb in its sense s and a

the argument head. We chose the formulation in Equation 3.3 for two reasons: One is that the subterms can be interpreted as capturing linguistically relevant information about the verb-argument pair. The second reason for decomposing the joint probability model in this way, which is more relevant here, is that variables are introduced in the order of their expected sparseness: The verb in some sense is expected to be the least sparse variable, followed by the grammatical function seen together with the verb. Since there is no one-to-one correspondence between roles and grammatical functions, triples of verb, grammatical function and role are presumably sparser than verb-grammatical function tuples. Finally, the argument head is introduced, which is expected to be sparsest, because many different argument heads can fill a thematic role in a specific syntactic realisation, but only few of the potential fillers can be expected to be present in the training corpus.

The decomposed model formulation in combination with smoothing by re-estimation improves the semantic model's predictions over those of a smoothed joint formulation in the following way: First of all, the first three model subterms ($P(v_s)$, $P(gf|v_s)$ and $P(r|v_s, gf)$) now contain fewer variables than the joint probability formulation and therefore should be less sparse in the the semantically annotated training data and less require smoothing. This means that in many cases of data sparseness, only the $P(a|v_s, gf, r)$ term has to be substituted by a smoothed estimate. In these cases, the overall probability prediction is still significantly influenced by the known verb preferences for the sense, the grammatical function and the role encoded in the first three model terms. This linguistic information about the verb is a helpful heuristic for the plausibility of the verb-argument pair if the argument is unknown. For example, the plausibility prediction will be higher for a role that the verb preferentially assigns than for one it assigns infrequently. In consequence, the model's plausibility predictions are much more specific to the input values than a uniform smoothed value output by the smoothed joint model formulation. If other model terms apart from $P(a|v_s, gf, r)$ are also sparse, less and less specific information is available, but only if the verb is unseen, the smoothed estimate is as unspecific as for the joint formulation. We evaluate the use of the *Good-Turing* smoothing method with and without *linear interpolation* as an instance of re-estimation based smoothing in Section 3.3.

We can address the lack of argument-specific information due to unseen argument heads independently by applying class-based smoothing to the $P(a|v_s, gf, r)$ term. Intuitively, this approach supplements (possibly sparse) information about a word by using information about semantically similar words. For example, when making a prediction for the plausibility of <cure$_1$, doctor, healer, agt>, we can also consider information about the co-occurrence of <cure$_1$, physician, healer, agt> if we know that a doctor and a physician are similar. This similarity information is supplied by semantic classes, which group together similar nouns and thereby abstract away from the lexical items to a higher-level semantic concept (for example *person with advanced medical training*), as discussed in Section 3.1.2. Using semantic classes affects the estimation

process directly: Instead of estimating how likely a noun is given a verb, grammatical function and role, we estimate the likelihood of seeing any member of the noun's semantic class in this position. The $P(a|v_s, gf, r)$ term thus becomes $P(class_a|v_s, gf, r)$, which allows us to pool the co-occurrence counts for all nouns from a semantic class. This increases both coverage of lexical items, because a noun class may be seen in the training corpus while a class member is not, and the accuracy of predictions, because probability estimates become more reliable the less sparse they are.

Such semantic generalisation on the basis of classes of similar concepts can be made for verbs as well as for nouns. The crucial parameter for the class-based smoothing method is of course the choice of semantic class, both in terms of class size and in terms of the semantic generalisation the class makes. We describe an in-depth evaluation of different types of verb classes and two choices of noun class granularity in Section 3.4.

In sum, using a re-estimation smoothing approach allows us to fully cover the test data, because it allows predictions to be made for any unseen event. Combining it with a decomposed version of the semantic model allows us to make the best possible use of plausibility information in the training data for these predictions. The class-based smoothing approach is orthogonal to re-estimation smoothing: It affects probability estimation directly by making semantic generalisations. It both increases coverage of lexical items, thus often avoiding the need for re-estimation smoothing, and allows more accurate predictions for both seen and unseen variable value combinations. Its performance however immediately depends on the semantic classes used.

3.1.4. Restricted Generativity

One important requirement for the semantic model is the ability to deal with incomplete input in the form of unspecified values. For example, in all evaluation tasks below, only the verb lemma is given and the verb sense is left unspecified. If an input value is unspecified, the model, thanks to its formulation, is able to generate the value that allows the most probable prediction by exhaustively substituting all seen values for the unknown value. This model feature allow us to make predictions in real-world settings where not all input values are known, and integrates in a natural way typical pre-processing tasks like word-sense disambiguation or the specification of possible roles for a verb-argument pair.

In some contexts, it is however preferable to restrict the generation of possible values. One such case is the treatment of unspecified input values, specifically of unknown grammatical functions, and the other is given by the interaction of the generative model formulation and smoothing.

Unspecified Values

The current implementation of the semantic model allows to specify input values for the verb lemma, the verb sense, the argument head and the grammatical function. The verb and argument lemmas have to be specified, while the thematic role that links verb and argument always remains unspecified – the model exhaustively generates all roles which are consistent with the specified values, ranked by plausibility.

If any other input value is unspecified, our general strategy is to generate the value that allows the most plausible role predictions. This is for example always applied to missing verb sense information, which determines the appropriate set of thematic roles to be used. However, for the treatment of missing grammatical functions another strategy is more promising. If no grammatical function is specified, we generate predictions using all seen grammatical functions and sum them, which amounts to dropping the grammatical function feature from the model. The motivation for this strategy is that the grammatical function that links verb and argument is a useful indication for which thematic roles are plausible, but if it is not specified, there is little advantage in generating the most frequent syntactic realisation if we can instead drop the grammatical function information altogether and derive our plausibility predictions from a more robust data set that pools syntactic realisations.

We apply the same strategy if a grammatical function is specified, but unseen with the verb. Since most roles can be syntactically realised in different ways, we assume that the exact value of the grammatical function feature is less crucial to the prediction of roles and plausibility ratings than for example the verb or argument lemma. Therefore, it is preferable to drop the grammatical functions feature and make verb- and argument-specific predictions than to output a smoothed estimate for an unseen input value combination.

Consistency of Predictions

The generative nature of the semantic model and its ability to deal with incomplete input information interacts with the application of smoothing methods. The unsmoothed plausibility model only instantiates missing values with values that have been seen in the training data together with the specified parts of the input, because unseen combinations have zero probability. Smoothing in principle allows us to generate predictions for any combination of values by supplying a smoothed probability estimate for unseen combinations. This would on the one hand enable the model to make predictions for value combinations that are unseen in our limited training data, but on the other hand it can lead to the prediction of plausibility estimates for inconsistent value combinations, for example for verb senses that are inconsistent with the specified verb lemma or roles that are incompatible with the verb sense. Additionally, it would also dramatically increase the search space for the optimal prediction. To keep search

Verb	Argument	Role	Rating
cure	doctor	agent	6.8
cure	doctor	patient	3.8
cure	patient	agent	1.4
cure	patient	patient	6.1

Table 3.1.: Test item: Verb-argument-role triples with ratings on a 7-point scale from McRae et al. (1998). 1: Implausible, 7: Plausible

manageable and in particular to ensure consistent predictions, we conservatively allow predictions to be made only for seen value combinations. Therefore, any smoothing that allows predictions for unseen events is applied only to the sparsest $P(v_s, r, gf, a)$ model term. We apply the re-estimation smoothing method also to the other model terms, but there it serves not to allow predictions for unseen value combinations, but to smooth the noise-prone estimates for seen events with low frequency.

3.2. Training and Evaluating Model Instances

Our goal in this chapter is to select the optimal estimation and smoothing approach for the semantic model. Section 3.2.1 explains how we compare different instances of the semantic model using different smoothing approaches by their performance in task-based evaluation. The objective function we wish to optimise is the reliable prediction of human plausibility judgements. Section 3.2.2 describes the training, development and test data.

3.2.1. The Judgement Prediction Task

The task we use for evaluation is the prediction of human intuitions about the plausibility of predications. Intuitions can be measured in terms of plausibility judgements for verb-argument-role triples. An example item from McRae et al. (1998) is presented in Table 3.1 (an item is a complete set of stimuli showing all manipulations, in this case argument and role identity). The ratings reflect the judges' intuitions that *doctors* typically *cure* instead of being *cured*, while the reverse is true for *patients*. The judgements also reflect the fact that it is possible for *doctors* to be *cured* by assigning a rating towards the middle of the scale for *cure-doctor-patient*.

We evaluate the quality of the predictions made by different instances of the semantic model by correlating the predicted plausibility values (probabilities ranging between 0 and 1) and the human judgements (average ratings ranging between 1 and 7). The

judgement data is not normally distributed, so we use Spearman's ρ (a non-parametric rank-order test). The ρ value ranges between 0 and 1 and indicates the strength of association between the two variables. A positive value that is significantly different from 0 indicates that the semantic model's predictions are significant predictors of human intuitions.

We compare model performance to human agreement on the judgements. The inter-agreement between human judges, computed as the average correlation between a single judge's ratings and the average ratings of all other judges, is usually around 0.7 (see Section 4.1.2 and Keller and Lapata, 2003). Thus, humans do not agree perfectly on the plausibility judgements.

We also report coverage of the tested data points as the percentage of data points for which a prediction was made. Since the semantic model only generates predictions for combinations of verb, argument, role and grammatical function that have been seen in the training data (see Section 3.1.4 above), it is possible that it does not make predictions for all verb-argument-role triples in the test data, and that coverage is imperfect.

The judgement prediction task is very hard to solve if the verb is unseen during training. Backing off to syntactic information or a frequency baseline is problematic for both available training corpora: In both resources, thematic roles are specific to verb senses, which makes it impossible to assign theoretically meaningful roles to unknown verbs. We therefore exclude items with unseen verbs from the development and test data.

3.2.2. Training, Development and Test Data

This section describes the training data used to estimate the semantic model instances, both during model selection in this chapter and during evaluation of the final model in Chapter 4. We also introduce the development set used for model selection and the test data on which we evaluate the final model in Section 3.6, to ensure that the semantic model makes appropriate predictions also on an unseen test set.

Training Data

To train the semantic model, we require language data with thematic role annotation, as motivated in Section 2.5.1. To date, there are two main efforts to semantically annotate corpora: PropBank (PB, Palmer et al., 2005) and FrameNet (FN, Ruppenhofer et al., 2005). The two approaches have substantially different goals and characteristics: PropBank annotates a corpus of running text, while FrameNet compiles references of verb behaviour for a lexicographic approach to verb semantics. Figure 3.1 gives an example of PB and FN style annotation. We train models on both corpora and compare their performance in this chapter and Chapter 4. Details on data preparation and the extraction of relevant features for training can be found in Appendix A.1.

PropBank The PropBank annotation project aims at creating a large corpus of English with both syntactic and semantic role annotation. The PropBank adds a layer of semantic annotation to the Wall Street Journal section of the Penn Treebank (Marcus et al., 1994). It contains c. 120,000 propositions and covers c. 3,000 verbs. Arguments and adjuncts are annotated for every verbal proposition in the corpus. A common set of argument labels *Arg0* to *Arg5* is used for each frame set of each verb, but argument labels are interpreted as verb sense specific in order to avoid difficulty with defining a closed set of thematic roles, and are not semantically defined. Some consistency in mapping has been achieved, so that agents are generally labelled *Arg0* and patients/themes *Arg1*, as in Figure 3.1. For adjuncts such as location and manner, *ArgM* roles are used that generalise across verbs.

There are no explicit verb groupings.[1] Each verb can have a number of senses defined by syntactic usage. These are indicated by different *frame sets*, i.e. sets of syntactic verb frames and the thematic roles tied to them that belong to the same syntactic verb usage. A new frame set is created for each role profile a verb exhibits. For example, *decline* has two frame sets: One for the *reject* sense, with a role for the agent and theme of the rejection, and one for the *fall* sense, which lacks the possibility of expressing theme, but allows a starting point, end point or extent of the falling event to be specified. Frame sets often, but not always, correspond to semantic sense distinctions. Thus, the PropBank sense distinctions and role labels are defined on a level that is very close to syntax. This means that they can often be correctly inferred by the syntactic configuration of verb and argument, but also that few semantic generalisations across verbs or roles are available.

FrameNet The FrameNet annotation project is primarily concerned with the lexical semantics of verbs (and other role-taking word classes), characterised by their different senses and the thematic roles they assign to their arguments. Consequently, FrameNet groups verbs with similar meanings together into frames (i.e., descriptions of prototypical situations). A frame then introduces a set of frame-specific roles for typical participants in these situations. In Figure 3.1, these are a *Healer* and a *Patient* in the *Cure* frame. Frames can also introduce non-core roles like *Location* or *Duration* that are the same across all frames and that generally apply to adjuncts. All verbs within a frame must be able to realise all of the frame's roles. In this criterion, a weak link to Levin's verb classification on the basis of patterns of argument realisation (Levin, 1993) is evident. Across frames, the same role name often, but not always indicates similar role semantics. Both the definition of frames as semantic verb classes and the semantic characterisation of frame-specific roles introduce semantic generalisations

[1] The VerbNet groupings have been used to some extent, for example such that the functions of roles numbered higher than 1 generally correspond between members of the same VerbNet class. We evaluate the use of VerbNet classes for class-based smoothing in Section 3.4.2.

PropBank	heal.01	[The doctor $_{Arg0}$]	cured	[the patient $_{Arg1}$]
FrameNet	Cure	[The doctor $_{Healer}$]	cured	[the patient $_{Patient}$]

Figure 3.1.: Example thematic role annotation: PropBank (above) and FrameNet (below).

into FrameNet annotation that are not present in PropBank.

The FrameNet resource (release 1.2) contains c. 57,000 verbal propositions and c. 2,000 verbs. FrameNet is thereby about half the size of PropBank. There are additional data sets for nouns and adjectives that can take arguments. Corpus annotation proceeds by frame, identifying frame members and then annotating example sentences extracted from the British National corpus (BNC, Burnard, 1995). The annotation aims to present each verb with all roles and in all syntactic diatheses, which in general yields good coverage even of non-core roles. Apart from the smaller size of the FrameNet corpus, there are two more caveats: Since the annotated corpus is constructed to serve lexicographic uses, only some senses of a verb may be present, and word frequencies in the FrameNet corpus may not be representative of English.

Development and Test Data

Our development and test sets for the evaluation of the smoothing methods come from the judgement data reported in McRae et al. (1998). Four example data points from this set were given in Table 3.1: One verb is paired with two arguments and two roles each. This data set was chosen for two reasons: First, each of the two arguments for each verb is highly plausible in one of the rated roles and implausible in the other. This means that the set is unbiased with regard to an overall preference for one role, ensuring that the semantic model can only correctly predict the judgements if it uses semantic plausibility information. Second, the human ratings the model has to predict are clearly distinct, which makes the model's task as straightforward as possible at this first model selection stage.[2]

A third reason for choosing this data set is that, at 160 data points, it is the largest argument role data set from the literature that is available to us. This allows us to split it into a 60 point development and a 100 point test set. The split was done randomly, with the constraint that all data points containing the same verb had to be in the same data set, to ensure that the data sets are truly independent.

The plausibility judgements for this data set were gathered by asking raters to assign a value on a scale from 1 (not plausible) to 7 (very plausible) to questions like *How*

[2]We test data sets with somewhat less distinct plausibility judgements in Chapter 4.

common is it for a doctor to cure someone? and *How common is it for a doctor to be cured?*, that prompted the agent and patient role for each argument.

All test pairs were hand-annotated with FrameNet and PropBank roles following the specifications in the FrameNet on-line database and the PropBank frames files, and using the FrameNet and PropBank corpora for reference in cases of doubt. If the verb sense appropriate to a verb-argument pair was not attested in FrameNet or PropBank, we could naturally not assign roles. Instead, the verb-argument pair received roles *None1* and *None2* for the two readings. For example, the verb *lift* is only attested in the FrameNet training data with senses *body movement* and *theft*. Therefore, we could not assign appropriate roles to the verb-argument pair *lift-infant* with the intended sense of *move object*, and assigned *None* roles instead. The roles were annotated by a single annotator only, but results on Inter-Annotator Agreement both in the PropBank project (Palmer et al., 2005) and in the Salsa project (FrameNet annotation of a German corpus Burchardt, Erk, Kowalski, and Pado, 2006) are high ($> 85\%$ for Salsa annotation, $\kappa = 0.91$ for PropBank). This indicates that annotators generally agree very well on the role labels they assign. It therefore appears justifiable to use only one set of annotations.

In Section 3.2.1, we have argued that no meaningful predictions can be made for items with unseen verbs. We therefore test only items with verbs seen in the training data (regardless of whether the seen sense is correct for the test item). On the McRae development set, 48 out of 60 data points remain for the FrameNet training data, and 56 out of 60 for the PropBank data. On the test set, the picture is similar with 92 data points for the PropBank model and 64 for the FrameNet model (both out of 100). The verb-argument-role triples from both the development and the test set are generally unseen in training: A completely unsmoothed model trained on either training corpus covers one data point of the development set and two of the test set.

3.3. Good-Turing Smoothing

In this section, we evaluate the use of re-estimation smoothing with the decomposed model (see Equation 3.2) which allows us to make verb-based predictions in case of sparse data. We use the Good-Turing (GT) method (Good, 1953, Manning and Schütze, 1999, for an introduction) to smooth the distribution of co-occurrence counts that we estimate from the training data and to assign a small probability to unseen events.

GT smoothing relies on re-estimating the frequency of seen and unseen events based on knowledge about more frequent events. Events are collected in classes according to the frequency with which they have been observed, and each class is assigned a proportion of the total number of occurrences observed for the next more frequent class. Thus, the class of unseen events is assigned some observations that have been "borrowed" from the frequency estimates for the more frequent classes.

Technically speaking, the GT method re-estimates the number of observations N

made of all events with observed frequency r as

$$r^* = (r+1)\frac{N_{r+1}}{N_r} \tag{3.4}$$

where the re-estimate r^* is computed as the ratio of the number of events with $r+1$ observations over the number of events with r observations. The $r+1$ factor adjusts for the fact that there are much fewer highly frequent events than infrequent events.

GT re-estimation is usually only applied to the lowest values for r, since the classes of low-frequency events are largest and the counts therefore are most accurate. Also, for the largest r, no observations exist for $r+1$, of course. We apply GT smoothing for $r \leq 2$.

The re-estimated observation frequencies are used to compute probabilities by dividing by the total number of observed events, as with the MLE method. Since the total number of events in the training data is unaffected by the re-estimation, those observed frequencies to which no smoothing has been applied are discounted to ensure that the resulting probabilities form a distribution which sums to 1. The most important difference between the resulting distribution and the distribution reached by the pure MLE method is that the former assigns probability mass to unseen events, as desired. It also differs in the re-estimated probabilities for smoothed infrequent events and in the proportionally somewhat lower probabilities to highly-frequent events, caused by the re-estimation and discounting.

3.3.1. Evaluation

Method

We estimated the probabilities for the semantic model from both the PropBank and the FrameNet training corpora and subsequently applied GT smoothing. The predictions of the smoothed model for the development set were correlated to the human judgements for the same verb-argument-role triples.

Results and Discussion

Table 3.2 lists the results of the Good-Turing smoothing (GT) model. For comparison, we also give coverage numbers for the completely unsmoothed model. For this model, a correlation cannot be computed because there are too few data points covered.

In comparison to the unsmoothed result, GT smoothing clearly allows satisfactory coverage of the test items. For PropBank, predictions can now be made for every development set data point, while for FrameNet, coverage is high, but not perfect.

[3] Levels of significance are specified in all tables in this chapter for two-tailed tests and follow the usual conventions: ns : not significant, $*:\ p < 0.05$, $**:\ p < 0.01$, $***:\ p < 0.001$

Train	Smoothing	Coverage	ρ
PB	None	2%	–
	GT	100%	0.188, ns
FN	None	2%	–
	GT	96%	0.150, ns

Table 3.2.: Good-Turing (GT) smoothing. Coverage and correlation strength (Spearman's ρ) for PB and FN data on the development set.[3]

Recall that our semantic model is restricted to making predictions only for roles seen with the target verb and that we do not exclude items where the verb is only seen in inapplicable senses. For such items, the model is often unable to predict the correct role.

Regarding the correlation of predictions to human judgements, it is obvious that Good-Turing smoothing makes imperfect predictions of plausibility. This is not unexpected, since the verbs in the development set appear with one good agent and one good patient each, and virtually all verb-argument-role triples are unseen (as is clear from the unsmoothed coverage). Recall that the smoothed model in this case makes predictions based only on the verb's co-occurrence with roles and grammatical functions. These predictions are necessarily the same for both unseen role fillers, so that we cannot expect this model to account for plausibility effects that depend on the identity of the role filler.

3.3.2. Adding Linear Interpolation

We also experimented with adding a second smoothing method, Linear Interpolation (LI), which is typically used for smoothing n-gram models (Manning and Schütze, 1999). This approach allows the use of less fine-grained probability information to form a more accurate estimate of (possibly sparse) probability distributions. It re-estimates the probability of the n-gram in question as a weighted combination of the n-gram, the n-1–gram and the n-2–gram. To ensure that a probability distribution is returned, the weights (usually denoted with λ) need to sum to 1. For our model, the LI re-estimate of, for example, the model term $P(a|v_s, gf, r)$ is computed as a combination of interpolation terms (which are estimated using MLE):

$$P_{LI}(a|v_s, gf, r) = \lambda_1 P_{MLE}(a|v_s, gf, r) + \lambda_2 P_{MLE}(a|v_s, r) + \lambda_3 P_{MLE}(a|v_s) \quad (3.5)$$

This approach thus allows us to use more general knowledge about the co-occurrence

of the argument just with the verb and the role in order to estimate our sparsest model term. If the argument is altogether unseen, this is of course not helpful, but if it was seen with the verb in some combination (for example realised in a different grammatical function or with a different role), the interpolation method, unlike the pure GT method, allow us to take this information into account. Abstracting away from the grammatical function probably will however not make much difference in a model that allows the grammatical function to be unspecified in the input and in such cases drops it through marginalisation. Therefore, the largest gain we can expect is in predictions for roles that are unseen with an observed verb-argument combination.

We constructed an LI model instance in which we re-estimated the model terms using the linear interpolation technique. Each of the semantic model's four conditional probability terms (see Equation 3.2) requires three λ values, as shown in Equation 3.5. The optimal λ values were estimated separately for each conditional probability term on the training data. We used five-fold cross-validation and set the λ terms to maximise the likelihood of the held-out fold. The final λ values are the average of the results across the five folds.

Evaluation shows that using LI on its own fails completely on our test data, because our sparse data problem is so serious that verbs and arguments are virtually never seen together, neither in the target role relation or any other. This means that for most data points, the $P(a|v_s, gf, r)$ term remains zero after interpolation, which precludes probability predictions.

We evaluated the LI method again after first applying GT smoothing to all model and interpolation terms both for λ estimation and evaluation. The combined model did not outperform the pure GT model (PB: $\rho = 0.156, ns$, FN: $\rho = -0.002, ns$). Inspection of the λ values for the $P(a|v_s, gf, r)$ term for the combined model shows why: Even after the application of GT smoothing, the training data is so sparse that the estimation process de-emphasises the sparsest (and most specific) λ term in order to maximise the likelihood of the test fold. In the extreme case (for FrameNet), $P(a|v_s, gf, r)$ is weighted at 0.075, $P(a|v_s, r)$ is weighted at 0.001, and $P(a|v_s)$, the least specific term, is used almost exclusively, at 0.924. These λ values mean, however, that the all-important argument-specific information is not used efficiently on the judgement prediction task, even when it is available. It therefore appears that maximising the data likelihood during λ estimation does not approximate our final task well enough. A better solution might be to use the correlation task directly as a λ estimation criterion, but this is much more complex, requiring us to estimate all λ terms simultaneously.

In sum, we conclude the LI method does not appear to be particularly suited for our task and our data sets, and its application indeed does not add anything beyond using the GT method. We will therefore restrict ourselves to using GT smoothing below.

3.4. Class-Based Smoothing

Good-Turing smoothing allows us to make predictions for almost all data points. However, the predictions are verb-specific at best if the argument is unseen, which is clearly reflected in the evaluation results. Therefore, we evaluate the use of class-based smoothing to improve the amount of argument-specific predictions the semantic model can make.

Class-based smoothing affects the estimation process by generalising from word tokens to word classes. The method is therefore especially appropriate for alleviating the sparseness of lexical items. By substituting word classes for both nouns and verbs, we maximise the number of verb-argument co-occurrences we can consider for estimation. We apply class-based smoothing to the $P(a|v_s, gf, r)$ term of the decomposed model to counteract the sparseness of the argument head variable and improve the amount and quality of argument-specific plausibility predictions. When the verb and argument head lemmas are substituted with their semantic classes, the term we estimate becomes $P(class_a|class_v, gf, r)$. In case a verb or noun is a member of several classes, we choose the class that allows the highest plausibility prediction.

There is a difficulty with estimating the class-based term using the MLE method if words can be members of more than one verb or noun class, which we want to explicitly allow to account for polysemy. If a word is a member of several classes, the training observations it occurs in are counted several times to establish co-occurrence counts for different class combinations. This means that the total of co-occurrence counts differs from N, the total of observations. To illustrate this problem, assume that the data point <cure$_1$, doctor, healer, agt> has been seen five times in the training data. If we assume that *doctor* belongs to four semantic classes and *cure*$_1$ belongs to two, there are eight combinations of classes that add the five observations of co-occurring class members to their total co-occurrence count. The total number of observed word co-occurrences is still only five, however. This makes it impossible to apply MLE and be returned a well-formed plausibility distribution over class co-occurrences.

There are two strategies to avoid this problem and ensure that $P(class_a|class_v, gf, r)$ remains a probability distribution: Either, observations can be split among classes that lay claim to them, effectively adding partial counts to class co-occurrence totals. Alternatively, counts that are used multiple times can be added multiple times to N to ensure that the MLE method returns a well-formed probability distribution where all probability values sum to 1.

We cannot use the count-splitting method here, because it is not compatible with GT smoothing and therefore makes it impossible to cleanly combine class-based and GT smoothing (see Section 3.5 for our combination strategy). GT smoothing is defined for integer co-occurrence counts (recall that it is based on grouping together events with the same observed frequency). Apart from binning observations together, which requires the setting of arbitrary bins, there is no way to adapt GT smoothing to the continuum

of real-valued co-occurrence counts that are the result of the splitting method.

Therefore, we adjust the total number of counts. We add counts to the total as many times as there are class combinations that use the counts. In the above example case, we would adjust the total number of data points by adding another seven times five counts to the actually observed total of five data points. This artificially inflates the amount of training data, but relative frequencies are maintained just as with the count-splitting strategy, and co-occurrence counts remain integers.

The success of the class-based smoothing approach hinges on the nature of the semantic word classes used. These classes need to reliably group together words with similar meanings, while covering as many words as possible from the training corpus and grouping them with other known words to maximise the coverage gain for smoothing. The need for reliability points towards using hand-created lexicographic classes from resources like WordNet or VerbNet. However, the need for good coverage of the training data and for grouping known verbs suggests inducing word classes from the training corpora.

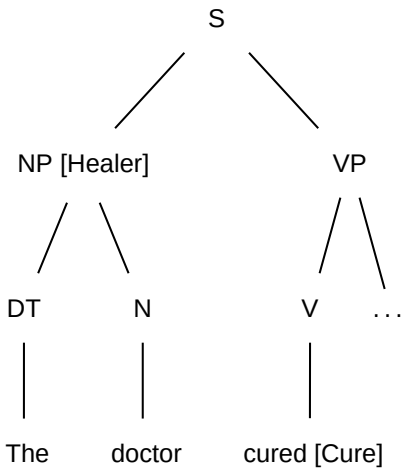
We experiment with both types of semantic classes for verbs, but restrict ourselves to lexicographic classes for nouns both for reasons of time efficiency and because we assume nouns to be sparser than verbs in a corpus, which makes it harder to infer meaningful classes. We first discuss the induction of verb classes in Section 3.4.1 and then compare induced versus lexicographic verb classes in Section 3.4.3, combining both with lexicographic noun classes of different granularities.

3.4.1. Induced verb classes

Inducing verb classes from the training data allows us to group verbs together according to semantic dimensions that are relevant for our task. In order to form semantic verb classes, we cluster verbs according to linguistic context information. This is feasible already using purely syntactic information: Levin (1993) has demonstrated that verbal subcategorisation and diathesis patterns allow the formation of verb classes that are broadly semantic as well as syntactic. Korhonen, Krymolowski, and Marx (2003) and Schulte im Walde and Brew (2002) demonstrate that similar verb classes can be induced automatically by clustering verbs according to subcategorisation information acquired from large corpora.

In our case, however, we also wish to exploit the semantic role annotation in our training data. This will allow us to induce classes of verbs that realise similar roles in similar ways, a type of information that is not always equivalent to realising similar argument structure. Only such verb classes allow meaningful class-based smoothing for the task at hand. We did not employ complete subcategorisation frames exactly because we are especially interested in the behaviour of verbs with regard to specific roles and argument heads.

We extract three types of features for each instance of a verb-argument pair: Semantic,



Feature	Value
Verb sense	Cure
Role	Healer
Path	NP-up-S-down-VP
Path-Role	Healer / NP-up-S-down-VP
Arg Head	doctor

Figure 3.2.: Clustering features: Syntactic parse tree with [FN semantic annotation] and corresponding feature set.

syntactic and lexical. Figure 3.2 shows the complete set of features for the verb-argument pair *doctor-cured*, assuming FrameNet semantic annotation. There are two semantic features: Verb sense and argument role. These are inferred directly from the semantic annotation. The sense feature gives information about human-drawn sense distinctions between verb usages, and the role feature tells which roles a verb realises. There are also two broadly syntactic features. The first is a description of the path through the parse tree from argument to verb. The sample parse tree in Figure 3.2 shows that a syntactic subject position is characterised as moving from an NP node up to the S node and down again to the verb's parent VP. The path feature thus gives similar information as a grammatical function feature, but also contains some information about the phrasal categories headed by verb and argument. We also use a combined feature of role and syntactic path, which specifies which role was realised as which syntactic argument, making linking information explicit. Finally, the lexical feature, the argument head lemma, yields some information about typical role fillers, but is expected to be sparse. These features are closely related to the features used to estimate the semantic model (namely verb lemma, verb sense, grammatical function, role and argument head). This ensures that we group verbs according to features that are relevant for the prediction of plausibility. We explore the relative importance of each of these features in Section 3.4.1.

To induce verb classes, we employed two soft clustering algorithms in the imple-

mentation of Marx (2004)[4], the Information Bottleneck method (Tishby, Pereira, and Bialek, 1999, previously used for verb clustering by Korhonen et al. (2003)) and the closely related Information Distortion algorithm (Gedeon, Parker, and Dimitrov, 2003). Unlike hard clustering approaches, these algorithms allow a verb to be a member of several clusters, thus making it possible to account for different senses of one verb. A description of the clustering algorithms and can be found in Appendix A.2, in Section A.2.1.

We varied three parameters for the induction of verb classes: The clustering algorithm, within that, the amount of smoothing during clustering, and the number of clusters. We compared the resulting sets of verb classes by task-based evaluation, measuring the performance of the semantic model on the judgement prediction task (Section 3.2.1) when each set of classes was used for class-based smoothing. Only sets of verb classes that allow the model to reliably predict human judgements on the development set were chosen. See Section A.2.2 in Appendix A.2 for a detailed account of the selection of the best-performing sets of verb classes.

Table 3.3 contains the most successful sets of verb classes identified in Appendix A.2 and for each set specifies the values of the two most informative of the three varied parameters, the clustering algorithm used and the number of clusters for each set. Each set is further identified for reference in later sections by the training set used for induction and a number. We present several verb classes for each training set to demonstrate that the smoothing performance of the verb class sets generalises across several different parametrisations. The results suggest the conclusion that the Information Distortion algorithm does somewhat better for FrameNet data, while the Information Bottleneck method works better for PropBank data.

As discussed in Appendix A.2, reliable verb classes could be induced for the PropBank training set only after the input data had been reduced to covering only NP arguments.[5] This measure strongly restricts the set of roles seen with each verb, so that the *Arg1* and *Arg0* roles, which allow most semantic generalisation, are seen almost exclusively. After the adaptation, it was possible to identify stable class sets for PropBank, as well. Those sets tend to be smaller than the FrameNet sets, however. Section 3.4.1 helps to explain this difference by giving some insight into the importance of different clustering features for the different sets of training data.

Finally, the coverage and correlation ρ numbers in Table 3.3 (for performance on the development set) confirm that the selected sets of verb classes indeed allow the semantic model to make reliable predictions of human data. While coverage is relatively low, it is still much increased from the unsmoothed model, which which covers only 2% of the development data (cf. Table 3.2). The correlation of predictions to observations is

[4]Many thanks to Zvika Marx for generously providing his software, and for interesting discussions about the outcome of the experiments.

[5]Model estimation is still done on the full argument data.

Corpus	Verb Classes	Algorithm	# of Clusters	Coverage	ρ
PB	PB 1	ID	3	16%	0.700, *
	PB 2	IB	11	25%	0.563, *
	PB 3	IB	7	11%	0.943, *
FN	FN 1	ID	11	15%	0.847, *
	FN 2	ID	13	21%	0.790, *
	FN 3	ID	13	21%	0.790, *

Table 3.3.: Induced verb class sets for both training corpora. Class set identification for reference below, algorithm, number of classes and smoothing performance (coverage and Spearman's ρ) on the development set.

generally strong at values around $\rho = 0.8$ or even above.

We will combine these induced sets of verb classes with lexicographic noun classes in Section 3.4.3 and compare their performance to that of hand-created lexicographic verb classes.

Feature Evaluation

During the class set selection process, differences in performance between the FrameNet and PropBank corpora were observed. We now investigate the contribution of each feature to the clustering performance to gain an insight into the reason for the very different results.

Recall the set of features for each verb-argument occurrence from the example in Figure 3.2 in Section 3.4.1: There are two semantic features, namely the argument role and the verb sense, one lexical feature, namely the argument head lemma, and two syntactic features, namely the syntactic path between verb and argument and the combination of path and role, which encodes linking information, are syntactic features.

Method We test each feature's contribution to the overall clustering result by leaving it out during clustering and comparing the smoothing performance of the resulting classes to the performance for classes induced with the standard feature set. We report example results for the parametrisations of verb class sets PB 2 for PropBank and FN 3 for FrameNet.

Results and Discussion Table 3.4 first gives the performance for using sets of verb classes induced with the full feature set and then the performance when each feature is

	FN			PB		
Features	Coverage	ρ		Coverage	ρ	
All	21%	0.790,	**	25%	0.563,	*
No role	19%	0.745,	*	23%	0.683,	*
No sense	21%	0.657,	*	38%	0.183,	ns
No arg head	21%	0.790,	**	23%	0.298,	ns
No path	21%	0.414,	ns	21%	0.420,	ns
No path-role	19%	0.711,	*	11%	0.722,	**

Table 3.4.: Testing the importance of features: Development set performance for the judgement prediction task using FN 3 and PB 2 verb class sets for smoothing. Coverage and correlation strength (Spearman's ρ) on the development set.

left out at a time. Both models show significant correlations to the human judgements data when all features are used.

FrameNet Looking at the feature classes in turn, we find that indeed, as expected, semantic information is important to forming useful semantic verb classes: The absence of the verb sense and role information causes a drop in correlation ρ. This is not surprising given that the FrameNet frame distinctions are an explicit semantic grouping of verbs and the roles also carry abstract semantic information because they are mostly frame-specific, but similar role names still often have similar meaning across frames. Note that the induced FrameNet classes do not mimic the distinctions between all nearly 300 FrameNet frames present in the training data, but make further generalisations over these frames into only 13 classes.

Despite the availability of informative semantic features, the group of syntactic features still has the strongest influence on cluster formation. No significant correlation is reached without the path feature, and the absence of linking information also causes a noticeable drop in correlation ρ. This underscores the claim that the syntactic and linking information are instrumental to the forming of semantic classes.

Finally, we see that the argument head feature has no impact on the smoothing results at all. It is not surprising that the argument head feature is not used, since its values are unlikely to generalise well.

PropBank The PropBank results differ noticeably from the FrameNet pattern. The one similarity is that the absence of the syntactic path feature causes problems: PropBank clustering also relies strongly on syntactic generalisations. The differences are

interesting: The role and path-role features contribute some generalisations (see the drop in coverage), but the classes induced without them perform better, so the generalisations seem to add noise, as well. Interestingly, this is the case although we restricted the input data to contain mostly $Arg0$ and $Arg1$ roles, which do allow some semantic generalisation across verbs. Possibly, the small number of semantic roles introduces a tendency to over-generalise, however, which would explain the results for leaving out the role-related features.

Very interestingly, also, the features that allow fewest generalisations have a large impact: Leaving out the PropBank sense feature and the argument head feature severely affects the clustering result. This happens despite the fact that PropBank senses do not allow generalisations across verbs and that the argument head feature is bound to be sparse for the overwhelming majority of verbs and argument heads. The reliance of the clusterer on these sparse features shows that the rest of the feature set (for example the path feature or the role feature) do not allow a strong grouping of verbs that would override the arbitrary signals of the non-generalisable features in the optimisation procedure of the clustering algorithm. This strongly suggests that the semantic information contained in the PropBank annotation does not yield strong hints towards a semantic grouping of verbs, and that the syntactic cues on their own, although important, do not allow a strong grouping, either. This may again be in part the result of restricting the input data to only NP arguments, which are bound to have similar path information. Recall, however, that no useful clusters could be induced at all using all arguments.

Summary In sum, we conclude that for both sets of training data, syntactic information is the backbone of grouping verbs together into meaningful classes, even if explicit semantic role and sense annotation are available. Another important insight is that the semantic content of the FrameNet annotation clearly facilitates the formation of clusters much more than the intentionally more shallow PropBank annotation. The induced FrameNet verb classes rely largely on syntactic and semantic information, while the shape of the PropBank verb classes proves to be influenced also by features that allow no or very little semantic generalisation, like the PropBank verb sense or the argument head. This observation suggests a larger susceptibility of the PropBank classes to overfitting the development set used for class selection, while the FrameNet classes can be hoped to generalise well across test sets due to their more robust reliance on features that allow generalisation.

3.4.2. Lexicographic Classes

Hand-created semantic noun and verb classes should be especially accurate and reliable. However, depending on their granularity and the similarity criteria, they may contain too many or too few alternatives, or many alternatives that are unseen in the

training data and therefore do not help the estimation process. We therefore compare lexicographic verb classes to our induced classes, and evaluate two instantiations of lexicographic noun classes at different levels of granularity.

Noun Classes

We extracted noun classes from WordNet 2.0, a hierarchical lexicographic data base (Fellbaum, 1998). In WordNet, nouns are grouped together in synonym sets (*synsets*) representing specific concepts. The synsets are linked into a graph structure by hypernymy relations. There are 25 unique beginner sets that each form the root of a WordNet subgraph. These have been grouped together into a top-level ontology of eleven supersets by making some further generalisations, such as grouping *animal, person* and *plant* together under *organism*.

We tested two sets of noun classes with different grain size. One is the top-level ontology, which we modified slightly by undoing the last level of generalisation for the two most frequent classes, *entity* and *physical object*, in an attempt to avoid overgeneralisation by the resolution of both of a verb's argument to these classes. This set of noun classes is extremely general and may cause overgeneration of semantic alternatives.

The second set of classes is the lowest level of generalisation, the noun synsets themselves, which contain only synonyms of the target word. These noun classes are expected to make extremely reliable generalisations, but probably will not increase coverage greatly. We did not attempt to further optimise the level of generalisation made by the noun classes to keep the evaluation practicable. While our semantic model's performance compares favourably to that of a set of selectional preference models that go to great lengths to select the correct level in the WordNet hierarchy (see Section 4.3), it is conceivable that model improvement may be gained by further exploiting the WordNet noun hierarchy.

Verb Classes

We use verb classes from two resources: WordNet and VerbNet. Verbs in WordNet are arranged in a top-level ontology of 15 semantic fields represented by unique beginners that head subtrees of verb sets. Example classes are stative verbs, as well as verbs of motion, perception and communication. Verbs are organised in synonym sets, which tend to be small as true verb synonymy is rarer than noun synonymy. As a test of synonymy, only verbs that have similar selectional restrictions are entered into a synset.

The synsets are organised in a hierarchy according to *troponymy*, a form of entailment that also shares characteristics with meronymy. It captures the observation that many meaning distinctions between verbs relate to some kind of change in manner. A test of troponymy is that "V_1 is to V_2 in a specific way", e.g., *to amble is to walk in a specific way*.

The verb hierarchy is relatively shallow (generally not more than four layers deep), and synsets are small, so we used the top-level classification as verb classes.

We also used verb classes from VerbNet (Kipper, Dang, and Palmer, 2000, Version 2.0,), which is a verb classification based on the Levin classes. While the Levin classes organise verbs by the types of diathesis transformations they can undergo, VerbNet makes the semantics of each verb class explicit by characterising the shared meaning components of the member verbs. VerbNet also links the arguments of each verb in a class to a thematic role and specifies selectional restrictions both for classes and, if necessary, specifically for each verb. Each sense of a listed verb corresponds to membership in the appropriate class. VerbNet adds further layers to the Levin classification by making more fine-grained syntactic and semantic sub-divisions. We used the most general verb classification, which corresponds to the most general level of Levin classes.

3.4.3. Evaluation

The coverage and accuracy improvements yielded by class-based smoothing hinge on the verb and noun classes used for generalisation. To evaluate this smoothing approach, we therefore create model instances using a number of different verb and noun classes. The results give an insight both into which kinds of classes perform best and into what coverage and what quality of predictions class-based smoothing allows the semantic to reach.

Method

To evaluate the class-based smoothing approach, we combine lexicographic and induced verb classes with both levels of lexicographic noun classes. We create different instances of the semantic model by using the different combinations of verb and noun classes during the estimation of the semantic model's probability terms. We then correlate the model's predictions for the development set items to the corresponding human plausibility judgements. For interesting comparisons, we test for significant differences between the correlation ρ values in this section and below using the method described in Raghunathan (2003). This method allows missing values in either underlying data set and therefore allows us to compare correlations on the FrameNet and PropBank data sets, which each contain only a subset of the original McRae development set due to the exclusion of items with unseen verbs. Note, however, that in general, significance is hard to reach even for numerically large differences in ρ values if the size of the underlying data sets is very different. This is an instance of the usual phenomenon that small sample sizes do not allow to make inferences with great certainty in significance testing.

	Synsets				Toplevel Classes			
Train	Verb Classes	Cov.	ρ		Verb Classes	Cov.	ρ	
PB	WN	96%	0.092,	ns	WN	96%	0.043,	ns
	VN	96%	0.125,	ns	VN	96%	0.043,	ns
	PB 1	25%	0.469,	ns	PB 1	96%	-0.034,	ns
	PB 2	45%	0.423,	*	PB 2	96%	-0.020,	ns
	PB 3	18%	0.774,	**	PB 3	96%	-0.044,	ns
FN	WN	29%	0.042,	ns	WN	96%	-0.002,	ns
	VN	2%	–		VN	96%	-0.017,	ns
	FN 1	19%	0.678,	ns	FN	96%	-0.042,	ns
	FN 2	37%	0.572,	*	FN	96%	-0.025,	ns
	FN 3	37%	0.572,	*	FN	96%	-0.024,	ns

Table 3.5.: Lexicographic versus induced verb classes and WN synsets versus top-level noun classes. Coverage and correlation strength (Spearman's ρ) for PB and FN training data on the development set.

Results and Discussion

Table 3.5 gives an overview of the results in terms of coverage and correlation strength. For comparison with unsmoothed performance, recall that only 2% of the development set are covered without smoothing for either training set. Most combinations of lexicographic classes reach very good coverage results that are near the perfect GT smoothing coverage (Section 3.3). However, high coverage numbers generally come with low ρ values. The lexicographic models appear to overgenerate alternative words to the point where predictions become arbitrary.

- **Lexicographic Noun Classes** The results in Table 3.5, especially those for combining the induced verb classes with both levels of noun classes, support our hypotheses about the performance of the noun classes. Relatively low coverage figures show that the noun synsets do not generate as many alternatives as the top-level classes, but they apparently lead to more accurate predictions than the top-level classes: The combinations of induced verb classes and noun synsets even lead to significant correlations of predictions and human judgements. The top-level noun classes in contrast allow almost full coverage for all verb classes, but the small, and even negative correlation coefficients show that their generalisations over-generate alternative nouns to the point of making the model's predictions arbitrary.

- **Lexicographic Verb Classes** The contribution of verb classes to the overall performance is most easily seen in the combinations with noun synsets, because the noun top-level classes lead to uniformly high coverage, but unreliable predictions in combination with all verb classes. On FrameNet, the lexicographic verb classes perform disappointingly: Coverage is below 30% for the WordNet verb classes, and the correlation ρ is close to zero. The result of using the VerbNet classes even corresponds to using no smoothing at all.

  For PropBank, the lexicographic verb classes generate enough alternatives for almost full coverage even when used with the noun synsets. Apparently, the vocabulary and class size of these resources is much better suited to generate alternative verbs for the PropBank training data. However, the low correlation coefficients imply that the lexicographic verb classes do not propose the right verbs to make accurate plausibility predictions. Also, VerbNet and WordNet classes do not differ in performance, although VerbNet classifications were used in the creation of PropBank verb senses and role definitions (see Section 3.2.2).

- **Induced Verb Classes** The combinations of induced verb classes and noun synsets numerically, though not significantly, outperform the combination of lexicographic verb classes and noun synsets. These combinations are the only ones to achieve significant correlations to the human data. They do not perform significantly differently from the combinations of lexicographic verb classes and noun synsets mainly because they cover only a relatively low number of data points, which makes it harder to establish with certainty that performance differences are not due to chance.

  The reason for the nonetheless good quality of the predictions made by the induced classes is that they were selected for grouping together the seen observations so that coverage and prediction of human data are optimised. In a sense, these classes make the most of the semantic and distributional information in the training data. While they reach much lower coverage values than most of the lexicographic classes, they allow the model to solve its task. The WordNet and VerbNet classes appear to capture a level of generalisation that is less useful for our task and data. This supports concerns about the grain size of lexicographic classes and their ability to generate semantic alternatives that are both present and relevant in the current data set.

The significant correlations to human data for the induced verb classes prove that the class-based smoothing approach, unlike the GT smoothing method, allows our semantic model to make predictions that capture the argument-specific plausibility judgements. These predictions are made solely on the basis of smoothed estimates, as only one data point in the development set has been seen for either training set.

Verb Classes for Final Evaluation

Given the observations above, we select our induced verb classes for final class-based smoothing. With regard to noun classes, we have clearly identified the highly accurate, but small WordNet synsets as more useful for class-based smoothing than WordNet top classes, which yield good coverage, but do not provide informative probability predictions. We will therefore combine our induced verb classes and the WordNet noun synsets as semantic classes for the class-based smoothing approach below.

As noted above, smoothing with different sets of induced verb classes creates different instances of the semantic model. We continue to evaluate all models based on different induced class sets to demonstrate that the model performs robustly across different sets of induced sets of verb classes.

3.5. Combining the Smoothing Methods

We have seen in Section 3.3 that GT smoothing allows the model to make predictions for almost all data points, but that these predictions at best take the verb's preferences for roles and grammatical functions into account. In Section 3.4, we have demonstrated that class-based smoothing leads to accurate predictions that are specific to both the verb and the argument. Class-based smoothing also serves to increase coverage, but not quite to the desired level. We therefore test now whether a combination of class-based smoothing and GT smoothing will do better than either of the smoothing methods on its own.

Equation 3.6 illustrates our strategy to combine class-based and GT smoothing: GT smoothing is always applied to the first three model terms. Since we do not allow predictions for events that are unseen in these three terms (see Section 3.1.4), GT smoothing mainly serves to smooth the counts for events that only appear once in the training data, because these are prone to noise.

The final, sparsest model term $P(a|v_s, gf, r)$ is estimated in a series of backoff steps, given in Equation 3.7.

$$Plausibility_{v,r,a} = P_{GT}(v_s) \cdot P_{GT}(gf|v_s) \cdot P_{GT}(r|v_s, gf) \cdot P_{BO}(a|v_s, gf, r) \tag{3.6}$$

where

$$P_{BO}(a|v_s, gf, r) = \begin{cases} P_{CB}(class_a|class_v, gf, r) & \text{if } f_{CB}(class_a, class_v, gf, r) > 0 \\ P_{CB}(class_a|class_v, r) & \text{if } f_{CB}(class_a, class_v, gf, r) = 0 \\ & \text{and } f_{CB}(class_a, class_v, r) > 0 \\ P_{GT}(class_a|class_v, r) & \text{else} \end{cases} \tag{3.7}$$

Train	Smoothing	Coverage	ρ	
PB	PB 1 + Syn + GT	100%	0.276,	*
	PB 2 + Syn + GT	100%	0.305,	*
	PB 3 + Syn + GT	100%	0.251,	ns
FN	FN 1 + Syn + GT	96%	0.254,	ns
	FN 2 + Syn + GT	96%	0.260,	ns
	FN 3 + Syn + GT	96%	0.260,	ns

Table 3.6.: Combining Class-Based and GT smoothing. Coverage and correlation strength (Spearman's ρ) for PB and FN data on the development set. Induced verb classes, WN synsets (Syn) as noun classes.

First, we try to estimate $P(a|v_s, gf, r)$ using class-based smoothing. If a combination of classes, grammatical function and role is unseen, we apply class-based smoothing again, but drop the grammatical function term.[6] Recall that in Section 3.1.4 we have argued that the grammatical function is less crucial to our plausibility predictions than the other terms, and that we can therefore drop the grammatical function information from the model to be able to make more robust predictions based on all syntactic realisations of a role if necessary. If class-based smoothing fails entirely, we back off to a GT estimate of seeing an unknown combination of classes. Note that Equation 3.7 is simplified for ease of exposition. In order to ensure that a probability distribution is returned by the backoff sequence, the backoff terms have to be weighted appropriately. See Section A.3 in the Appendix for details of the weighting regime.

3.5.1. Evaluation

Method

Using the backoff approach introduced above, we realise model instances (with variations due to the different sets if induced verb classes used) by estimating the probability terms of the semantic model using a combination of class-based smoothing and GT smoothing. We then correlate the resulting models' predictions for the development set to the corresponding human plausibility judgements.

Results and Discussion

Table 3.6 contains the results of evaluating the combined smoothing methods on the development set. Coverage of the combined approach is of course the same as for the GT method, at the desired level of full or almost full coverage of the development set. Correlation ρs of the combined model are numerically, though not significantly, higher than for the GT method alone, where they are below 0.200 for both the FrameNet data and PropBank data (see Table 3.2). Additionally, there are significant correlations for two of the three PropBank models. These models perform slightly better than the FrameNet models, and they have the further advantage of a larger development set (56 data points as opposed to 48), which makes it easier for the PropBank ρ values to reach significance. The combination of smoothing methods therefore clearly performs better than pure GT smoothing.

The combined model's ρ values are numerically much lower than those seen for pure class-based smoothing. However, in comparing the ρ values of the combined approach and pure class-based smoothing, the large difference in coverage between the two models has to be taken into account. Indeed, the combined results are not significantly different from the values seen for just class-based smoothing (with the exception of the PB 3 model, $p < 0.05$, one-tailed). At the same time, the combined models have much higher coverage than the models using only class-based smoothing.

In sum, combining pure GT smoothing and class-based smoothing is our best attempt yet at reaching both good coverage and reliable predictions. In Section 3.6 below, we evaluate our semantic model using the combination of GT smoothing and class-based smoothing on the test set to ensure that the semantic model is able to make reliable predictions of completely unseen human data.

3.6. Validation of the Final Model

We now present results on the test set using the estimation strategy of combining GT and class-based smoothing that we selected as optimal in Section 3.5 above. We have seen during model selection that class-based smoothing alone furnishes reliable predictions, but with generally low coverage. Adding GT smoothing, we were able to also achieve satisfactory coverage on the development set. The PropBank model reached significant correlations to the human data on the development set.

[6]This is only relevant if a known grammatical function is specified. Recall that if the grammatical function is unseen with the verb or unspecified, we drop the grammatical function term from the model completely.

Train	Verb Classes	Coverage	ρ	
PB	PB 1	98%	0.098,	ns
	PB 2	98%	0.097,	ns
	PB 3	98%	0.105,	ns
FN	FN 1	88%	0.278,	*
	FN 2	88%	0.364,	**
	FN 3	88%	0.415,	**

Table 3.7.: Validating the Final Model. Coverage and correlation strength (Spearman's ρ) for PB and FN data on the test set. Induced verb classes, WN synsets as noun classes.

Results and Discussion

Table 3.7 presents the results of correlating the final model's predictions for the test set with human plausibility judgements.

On the larger test set, the performance of the models is reversed: The FrameNet models all significantly predict the human judgements and numerically clearly outperform the PropBank models, which are far from reaching significant correlations. [7]

The reason for the performance difference observed for the PropBank models between the development and test sets appears to be largely sparse data: The verbs in the test set are much sparser in PropBank than the verbs in the development set. In the development set, twelve of the 14 verbs are present in the PropBank corpus more than 30 times, but out of the 23 verbs in the test set, only 13 were this frequent. This results in less influence of reliable predictions from class-based smoothing: On the test set, only 21% of data points can be predicted using just class-based smoothing (as opposed to 45% on the development set). FrameNet does not show these coverage differences between development and test data, and the FrameNet models consequently perform much better than the PropBank models. In addition, it may be that the PropBank models overfit the development set and fail to generalise well to the test set, as predicted from the results of the evaluation of clustering features in Section 3.4.1. We will continue to test models for both training sets in Chapter 4 to ensure a fair comparison over several test sets.

Coverage of the test data is good, but not perfect: Recall that our semantic model can only assign roles seen with the verb in the training data. This means that for sparsely attested verbs or verbs only seen in inapplicable senses, no prediction can be made.

[7] The FrameNet models' correlation ρs do not significantly differ from the PropBank models' due to the large difference in test set size (92 data points for PropBank, and 64 for FrameNet).

3.7. Summary and Discussion

In this chapter we have evaluated two smoothing approaches intended to alleviate the sparse data problem encountered by the semantic model. Good-Turing smoothing re-estimates the model distributions and assigns a small probability to unseen events. In combination with decomposing the joint probability model, this smoothing approach ensures full coverage of the test data, but makes at best verb-specific predictions.

We also evaluated class-based smoothing, which is geared especially at overcoming the sparse data problem for argument heads, which are the sparsest variable in the semantic model. Class-based smoothing affects the model estimation process by substituting verb and noun classes for the argument head and verb lemmas. This generalisation allows better coverage and higher accuracy of predictions than an unsmoothed model.

We evaluated lexicographic and induced verb classes, and found that the induced verb classes perform best, because they capture generalisations that are relevant to the training data and the task. Especially for the FrameNet training data, the lexicographic verb classes appear to make generalisations on the wrong level of abstraction. We also evaluated two sets of lexicographic noun verb classes. One, the top-level ontology of WordNet, proved to over-generate alternative nouns to the point at which the semantic model's predictions cease to be meaningful. The other, the set of synonyms for each word, yield few generalisations with high accuracy and therefore performed better. Class-based smoothing with a combination of induced verb classes and WordNet synonym sets enables the semantic model to make predictions that are significantly correlated to human data. However, the model does not achieve full coverage.

Optimal model performance is reached by combining class-based smoothing and Good-Turing smoothing. Validation on an unseen test set shows that the resulting model's predictions are significantly correlated to human plausibility judgements and that the semantic model reaches full coverage.

On the test set, as throughout the chapter, models with induced verb classes trained on FrameNet outperformed models trained on the PropBank corpus. An analysis of the verb class induction process suggests that FrameNet semantic annotation leads to the formation of more informative classes than the more shallow PropBank annotation.

4. Evaluation of the Semantic Model

In this chapter, we present a thorough evaluation of the semantic model formulation selected in Chapter 3. This is the model that combines GT smoothing and class-based smoothing (using induced verb classes and WordNet noun synsets). We continue to estimate versions of the semantic model both from the PropBank and the FrameNet training corpus and to present results for three model instances per training corpus to demonstrate the model's robustness across different verb classes for class-based smoothing.

As in Chapter 3, the evaluation task is predicting human semantic plausibility judgements. The first set of experiments, in Section 4.1, tests the semantic model's predictions of human semantic judgements across several new data sets. We also explore model performance on seen and unseen data points to verify the appropriateness of the smoothing methods and to demonstrate that the model is able to differentiate within the sets of generally plausible and generally implausible data points.

The second set of experiments compares the semantic model to models that solve two related tasks from computational linguistics, namely semantic role labelling (Section 4.2) and the induction of selectional preferences (Section 4.3). Our evaluation shows that the model reliably predicts human judgement data over a variety of test sets and performs better at this task than the related approaches.

4.1. Predictions on Different Test Sets

The first part of our evaluation will demonstrate that the semantic model performs well across three previously unseen sets of judgement data with different characteristics: One, similar to the test set in the previous chapter, consists of items from the literature that were chosen by experimenters to exhibit extreme role preferences (Section 4.1.1). The items of the second test set were extracted from the training corpora, making them more similar in vocabulary to the training data of the model (Section 4.1.2). Also, for this set, role biases are less pronounced. Finally, in Section 4.1.3, we look at a third data set with judgements for adjunct roles, in contrast to all previous data sets, which contained verbs and their arguments. This data set again contains ratings on the full range of the scale. We also explicitly address the semantic model's performance on seen and unseen verb-argument combinations in Section 4.1.2.

Verb	Argument	Role	Rating
throw	ball	agent	1.4
throw	ball	theme	6.2

Table 4.1.: Example item from Trueswell et al. (1994): Pair of verb and inanimate argument with FN roles. Ratings for the agent and theme role on a 7-point scale. 1: Implausible, 7: Plausible

4.1.1. Trueswell Materials: Arguments

The first test set we consider is taken from Trueswell et al. (1994). Table 4.1 shows an example item: Verbs are paired with one argument and rated in two roles. As for the McRae test set used in Chapter 3, for each verb-argument pair, one of two rated roles is highly plausible, but the other is implausible. Therefore, the distribution of plausibility ratings is heavily biased towards points on the high and low ends of the scale. We use this test set to ascertain that the semantic model is able to predict human judgements for more than one data set from the literature.

Method

We test on 76 data points from Trueswell et al. (1994), which consist of verb-argument pairs where the argument is highly plausible as an object (in a patient or theme role), but implausible as a subject (in an agent role). This is achieved by using only inanimate arguments, a manipulation which has a strong effect because many verbs require plausible agents to be animate. The data were gathered in the same rating study as the McRae et al. data, so we can assume consistency of the plausibility ratings on a $1-7$ scale across the two studies. However, the data set crucially differs from the McRae et al. set in that it contains only one argument per verb and lacks ratings for plausible agents. Therefore, models with a bias towards preferring patient or theme roles have an advantage in predicting the judgements from this data set.

After eliminating items with verbs unseen in PropBank, 72 data points remain. 54 data points are covered by the FrameNet corpus. Out of these, three are seen the FrameNet training data, while the PropBank training data contains 12 out of the 72 test data points. We correlate the predictions of three models per training corpus with the human judgements using Spearman's ρ. As in Chapter 3, for interesting comparisons we test differences between ρ values for significance using the method described in Raghunathan (2003).

Training	Verb Classes	Coverage	ρ	
PB	PB 1	100%	0.306,	**
	PB 2	100%	0.350,	**
	PB 3	100%	0.334,	**
FN	FN 1	81.4%	0.397,	**
	FN 2	81.4%	0.427,	**
	FN 3	81.4%	0.522,	***

Table 4.2.: Trueswell materials: Coverage and correlation strength (Spearman's ρ) for FN and PB training corpora.[1]

Results and Discussion

For both training sets, the semantic model achieves good coverage and significant correlations to human data (see Table 4.1.1). Coverage for the FrameNet models is below 100% because of our model's restriction to predicting only seen roles (see Section 3.1.4).

On the Trueswell test set, both the FrameNet and PropBank models achieve significant correlations to the human data, which are not significantly different from each other in strength. The good result for the PropBank models, which is in contrast to their performance on the McRae test set in Chapter 3, could be due to the much larger percentage of test data points seen in the PropBank training data (17% of the Trueswell data points, in comparison to 2% of the McRae data points). Generally, the more data points are seen, the better our model tends to perform (see Section 4.1.4). However, the models' predictions for just the twelve seen data points are not significantly correlated to human judgements ($\rho < 0, ns$), so the good correlation results must be caused by another factor. The most likely reason for the significant correlation is a bias towards the *Arg1* role in the training data. This is the most frequent role in the PropBank corpus, as it applies to both patients and themes (e.g., the subjects of motion verbs). Therefore, the models often prefer it in the absence of argument-specific information. This is a successful strategy for predicting the Trueswell ratings, which reflect that all arguments are plausible patients, but implausible agents. Therefore, the good performance of the PropBank models has to be taken with a grain of salt.

The FrameNet models also achieve somewhat higher ρ values on the current data set than on the McRae test set, where the highest value was $\rho = 0.415$. Since the percentage of seen data points is the same for the Trueswell set as for the McRae test

[1] Levels of significance are specified in all tables in this chapter for two-tailed tests and follow the usual conventions: ns : not significant, $*:\ p < 0.05, **:\ p < 0.01, ***:\ p < 0.001$

set at 6%, this probably also reflects a small bias for patient-type roles in FrameNet. However, the large number of frame-specific names for patient- and theme-type roles in FrameNet makes a strong overall bias like the one observed in the PropBank training data impossible. The small bias we observe here most likely stems from a general syntactic bias to see more object than subject arguments with the test verbs, because subjects are not always realised, for example in passive constructions.

4.1.2. Padó Materials: Arguments

Our second experiment explores the semantic model's performance on items which were extracted from corpus data rather than constructed by experimenters to show certain plausibility properties. Stimuli that are strongly biased towards plausibility in just one role, as in the McRae and Trueswell materials, might make the model's task easier because it can achieve good results by strongly preferring one role and dispreferring all others. We therefore gathered our own test set that reflects a less extremely biased distribution of ratings. This data set further addresses the lack of verb coverage of FrameNet-trained models by ensuring complete coverage of the verbs by the FrameNet training set. Finally, the data set is more similar to the training data in vocabulary, thereby eliminating some of the sparse data problems caused by genre differences (recall Section 3.1.2).

Material Acquisition –Method

Materials To ensure that all the verbs in the new test set are covered in our training data, we used 18 verbs that appear in both FrameNet (release 1.1, largely a subset of the data in release 1.2 which we use for training) and PropBank. To vary the type of thematic roles assigned by the verbs to their subjects and objects, we chose six verbs each from three verb classes: Experiencer verbs like *hear* that assign an experiencer and a stimulus, Patient verbs like *hit* that assign an agent and a patient and Communication verbs like *tell* that assign a speaker and an addressee. The verbs were selected for the thematic roles they assign according to VerbNet.[2] This resource was chosen because it defines thematic roles, but is (at least to some degree) independent of FrameNet and PropBank.

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. Table 4.3 shows an example for the verb *hit*, which is presented with one filler each for the subject and object roles from each training corpus. In total, for each verb, there were usually six arguments from each corpus, and twelve arguments altogether. We constructed verb-role-argument triples by combining each

[2]Communication verbs assign an agent and a recipient in VerbNet.

FN Arguments				PB Arguments			
Verb	Argument	Role	Rating	Verb	Argument	Role	Rating
hit	man	**agent**	5.6	hit	player	**agent**	5.6
hit	man	victim	4.9	hit	player	victim	5.7
hit	baby	agent	6.2	hit	ball	impactor	5.7
hit	baby	**victim**	3.6	hit	ball	**impactee**	6.1

Table 4.3.: Example Padó stimuli: Verb-argument-role triples for *hit*. Arguments from FrameNet and PropBank, FrameNet roles, seen roles in **bold face**. Ratings on a 7-point scale, 1: Implausible, 7: Plausible.

verb-argument pair with both roles, obtaining 24 verb-role-argument triples per verb. In this way, we elicited ratings for both roles, regardless of the argument's original role in the corpus. In all, there are 414 verb-role-argument triples instead of the full $24 \times 18 = 432$, because some arguments were seen in both corpora or roles.

Procedure We collected ratings for the verb-argument-role triples on the World Wide Web (using the WebExp package, `www.webexp.info`). To avoid participants rating the same stimuli in both the subject and object interpretation and due to the large number of stimuli, we presented four separate lists of stimuli 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.

Participants 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 list), we excluded one more participant who had rated only one item. 99 participants remained, approximately half of whom were from Great Britain. 40 were from the United States, and another 12 from Canada, Ireland and Australia.

Normalisation and Annotation To minimise noise, we excluded ratings that were more than 2 points from the stimulus median. The average number of ratings per stimulus was 21.

Annotation of the verb-argument pairs with the appropriate roles as described in Section 3.2.2 was mostly unproblematic, apart from the annotation of one verb with FrameNet roles. For *raise* in the sense of *affect the value of an item on some scale*,

FrameNet specifies the roles of *Attribute* (scale along which the item ranges in value, for example *price*, *amount* or *height*) and *Item* (entity whose position on the scale is changed). This ontological distinction is often hard to make in practice, even for the FrameNet annotators. It turned out that in the FrameNet corpus, *raise* never occurs with *Item*. Instead, many arguments that should clearly be *Items* are labelled *Attribute*. For many of our verb-argument pairs, *Item* was however clearly the correct role for the object reading (e.g., for raise-dividend). We therefore annotated the correct, but unseen role at the cost of coverage.

Material Acquisition – Results and Discussion

Even though our verb-argument pairs were not selected to be extremely plausible in one of the two roles and implausible in the other, we cannot exclude the presence of biases. First, we test for a bias for the *Arg1* role when testing on the PropBank data set to exclude a confound of role bias in the training data with the predictions, as observed for the Trueswell data set in Section 4.1.1. A Wilcoxon's independent sample rank-sum test shows that there is no difference in the mean ratings for the 173 occurrences of *Arg1* and the mean ratings for the 241 occurrences of other roles (mostly *Arg0* and *Arg2*) ($W = 18808, p > 0.8$). We therefore do not expect the PropBank model's predictions of our ratings to be confounded with a model bias for the *Arg1* role.

Another possible bias in the data that we have to be aware of stems from the sampling process: It is likely that seen combinations of verb, argument and role are rated more to be more plausible than unseen ones. While unseen combinations do not have to be implausible, seen combinations were used by some author to describe a state of the world and should therefore be plausible to raters in most cases. Indeed, for both FrameNet and PropBank, verb-argument-role triples seen in the training set were rated higher than unseen triples (Wilcoxon's independent sample rank-sum test: FrameNet $W = 24602.5, p < 0.001$, PropBank $W = 24323.5, p < 0.001$, both two-tailed). This makes it more likely that a probabilistic model that assigns higher probabilities to seen triples will be able to model the seen data points correctly. However, note that far from all triples were seen in the training data. For FrameNet, only 112 out of the 414 data points were seen, leaving roughly 75% of the test data points unseen. For PropBank, the number of seen data points is not much higher at 135. Thus, the model's task is easier for this data set than for the preceding ones, but it still is not trivial.

Finally, a typical property of the literature data sets is the skewed distribution of human judgements, because the verb-argument pairs are selected to be highly plausible in one of the tested roles, and implausible in the other. We note that our data set contains many more intermediate ratings than the experimenter-chosen biased data sets. The ratings for our data set range from 1.0 to 6.9, with 32% of all ratings at and between 3.0 and 5.0, two points away from the extremes of the scale. In comparison, the Trueswell set (testing the PropBank model) shows a range of 1.0 to 6.6 with 17% of all ratings at

Training	Verb Classes	Coverage	ρ	
PB	PB 1	100%	0.270,	***
	PB 2	100%	0.250,	***
	PB 3	100%	0.286,	***
FN	FN 1	96.9%	0.514,	***
	FN 2	96.9%	0.521,	***
	FN 3	96.9%	0.515,	***

Table 4.4.: Padó materials: Coverage and correlation strength (Spearman's ρ) for FN and PB training corpora.

and between 3.0 and 5.0 and the McRae set (again testing the PropBank model) has a range from 1.0 to 7.0 with 13% of all ratings at and between 3.0 and 5.0. This means that for our own data set, our semantic model has to make more appropriate intermediate predictions in order to model the data correctly. Making consistently high predictions for seen and consistently low predictions for unseen pairs is not enough.

Evaluation – Method

The construction of the experimental stimuli ensured that all verbs are covered in both training corpora. We therefore test all 414 verb-argument-role triples. As mentioned above, for FrameNet, 112 out of the 414 data points were seen, and 135 for PropBank. Again, we evaluate by correlating the models' predictions to the human judgements using Spearman's ρ, as, even for this data sets, human ratings are not normally distributed.

For this data set, we can compute the inter-rater correlation as a plausible comparison mark for correlation ρ. This measure shows how well the ratings of the human judges agree. We compute it by correlating the ratings of each judge to the average ratings of the remaining judges and averaging the result over all judges. High inter-rater correlation signals high agreement about the plausibility of the rated stimuli. The inter-rater correlation reaches $\rho = 0.68$, which is reasonably high, but far from a perfect correlation of 1.

Evaluation – Results and Discussion

Table 4.4 contains the results of correlating the semantic models' predictions to the judgements from the Padó set. Despite the careful choice of the test items, coverage is still slightly below 100% for the FrameNet models, which is mostly due to the

idiosyncrasy of the training data with regard to *raise* described above. However, these cases are rare in comparison to other test sets, and we did not exclude items a priori.

The human data is clearly significantly predicted by both the PropBank and the FrameNet models. This time, the PropBank models' performance does not stem from a role bias, since the materials lack any object bias (see above). The PropBank models' results however still stay significantly below the FrameNet models' ($p < 0.001$ one-tailed comparing FN 1 and PB 3), in keeping with the prediction we have made in Section 3.4.1 that the FrameNet models should generalise better to unseen data sets. Indeed, the correlation ρ of the FrameNet models to the human data are the highest seen so far. Neither model however quite reaches the level of human rater inter-correlation of $\rho = 0.68$

Not surprisingly, the higher percentage of seen verb-argument-role triples and the greater similarity of the test items to the training data indeed make it easier for the models to correctly predict human judgements. For a more in-depth discussion of the models' coverage of seen and unseen triples see also Section 4.1.4 below. There, we find that the models' performance is not carried by accurate predictions for the seen data points alone, but that performance is still reliable for the unseen verb-argument-role triples. The present results thus also indicate that the models' performance does not depend on the existence of an extreme bias towards the ends of the plausibility scale in the human judgements, but that they are also able to predict data sets with a larger number of intermediate ratings.

4.1.3. Ferretti Materials: Adjuncts

We are also interested in the semantic models' performance on roles that are normally realised as adjuncts. These roles do not belong to the verb-specific role inventory, but can be assigned by all verbs, for example to a prepositional phrase specifying place or instrument.

Instrument and location roles are annotated with *ArgM* roles in PropBank, which are separate from the verb-specific roles in that they can apply to all verb senses. In FrameNet, adjuncts are generally assigned *non-core* roles that are also identical across frames, but frames may or may not allow the assignment of specific non-core roles. While FrameNet annotation does not focus on non-core roles, the project aims at providing at least one annotated instance of each possible role for the verb, so adjunct roles may be less frequent in the corpus than core roles, but will generally be attested. PropBank annotates running text, so coverage for *ArgM* roles can be expected to be more uneven across verbs.

We test the models' predictions on norming data for instruments and locations kindly provided by Ken McRae. These are the complete sets of ratings gathered for the study presented in Ferretti, McRae, and Hatherell (2001), including typical and less typical instruments and locations. Judgements on a 7-point scale are for one role only.

Test	Verb Classes	FN			PB		
		Coverage	ρ		Coverage	ρ	
Instruments	1	45.8%	0.258,	*	81.4%	-0.011,	ns
	2	45.8%	0.232,	*	81.4%	0.019,	ns
	3	45.8%	0.248,	*	81.4%	0.006,	ns
Locations	1	65.4%	0.183,	ns	63.3%	-0.003,	ns
	2	65.4%	0.259,	**	63.3%	-0.001,	ns
	3	65.4%	0.202,	*	63.3%	0.002,	ns

Table 4.5.: Ferretti materials: Coverage and correlation strength (Spearman's ρ) for FN (left) and PB (right) training corpora and Instrument and Location test sets.

Method

For FrameNet, we test only items with verbs in frames that specify a role for instruments and locations (many of the experiencer verbs do not allow these non-core roles). As before, a verb is tested even if it seen only in frames that do not describe the sense it is used in. Out of the 278 data points for instruments, the models are tested on the 162 stimuli with known verbs; for the locations, predictions for 156 out of 249 data points are tested. 14 data points are seen for instruments and eleven for locations.

PropBank covers 242 of the 278 instrument ratings and all 249 location data points. For each test set, two data points are seen in the training data.

Results and Discussion

While both coverage and correlation ρs are lower for the adjunct roles than for the argument roles for both models, the FrameNet models still achieve significant correlations with human preferences for both test sets. For FrameNet, coverage of the location data is higher than of the instrument data, which points to a lower frequency of role-labelled instruments than locations in the training corpus. If a verb is unseen with adjunct roles, this precludes predictions under our conservative strategy for ensuring prediction consistency. Sparseness in the training corpus is also the reason for the relatively low ρ values. Even where they are present, adverbial roles are much less frequent per verb than other roles, which makes class-based smoothing harder due to a lack of role fillers available for generalisation.

The PropBank models reach much higher coverage values for instruments than the FrameNet models, reversing the coverage trends for FrameNet. Obviously, the PropBank strategy of annotating running text results in more annotations for these roles. However, the correlation ρs are extremely low at around zero. The role fillers

observed in PropBank do not seem to allow useful generalisation to the test data, which was suggested already by the low number of seen location and instrument data points in comparison to the numbers for the much smaller FrameNet corpus. Due to the large difference in coverage for the instrument test set, however only the performance of the best-performing FrameNet model (FN 1) is significantly better than the performance of the best PropBank model (PB 2, $p < 0.05$, one-tailed), while for the location data even the lowest-performing FrameNet model (FN 1) significantly outperforms the best-performing PropBank model (PB 3, $p < 0.05$, one-tailed).

In sum, while the plausibility of adjunct roles proves harder to predict for both corpora, the FrameNet models still succeed in reaching a significant correlation to human judgements. Although fewer adjunct roles are attested with FrameNet verbs than with PropBank verbs, the role fillers for the FrameNet verbs appear to allow better semantic generalisation to the test data. The lower coverage of adjunct roles in FrameNet may point at genre differences between the source corpora or, more probably, at biases in the FrameNet corpus caused by the annotation strategy, which focuses on argument roles.

4.1.4. Seen and Unseen Verb-Argument Combinations

After successfully evaluating our models on three new data sets, we now explore their behaviour with regard to seen and unseen data points in greater detail. For two of the test sets discussed above, almost all data points were unseen, so we can already attest the semantic model instances good coverage of unseen data (especially when using the FrameNet corpus). Now, we take a closer look at the models' behaviour for seen and unseen data points within one test set. The Padó data set (see Section 4.1.2) lends itself especially well to this comparison as approximately half the verb-argument pairs are unseen for each training corpus.

Method

We split the Padó test set into the seen and unseen verb-argument pairs for each training corpus and evaluate both subsets separately. A seen verb-argument pair is defined as a verb-argument pair from the relevant training corpus with both the role that it originally appeared with as well as the other role for which we have elicited a judgement. For FrameNet, out of the 207 verb-argument pairs corresponding to 414 stimuli, 108 verb-argument pairs were seen and 99 unseen. For PropBank, 123 verb-argument pairs were seen and 84 unseen.

Out of the 108 pairs seen in FrameNet, 112 verb-argument-relation triples are covered without smoothing in FrameNet. This means that for four pairs, verb and argument were seen with both roles. For the PropBank corpus, 27 verb-argument pairs were seen with both roles. Since only these very low numbers of verb-argument-role triples were

Training	Test	Verb Classes	Coverage	ρ	
PB	Seen	PB 1	100%	0.377,	***
		PB 2	100%	0.312,	***
		PB 3	100%	0.400,	***
	Unseen	PB 1	100%	0.305,	***
		PB 2	100%	0.328,	***
		PB 3	100%	0.313,	***
FN	Seen	FN 1	96.8%	0.568,	***
		FN 2	96.8%	0.572,	***
		FN 3	96.8%	0.569,	***
	Unseen	FN 1	96.9%	0.374,	***
		FN 2	96.9%	0.390,	***
		FN 3	96.9%	0.383,	***

Table 4.6.: Seen and unseen data: Coverage and correlation strength (Spearman's ρ) for FN and PB training corpora and seen and unseen data from Padó test set.

also attested with the "unseen" role in the corpora, we expect to see a bias for assigning higher probability to the "seen" role. Recall also from the discussion of the data set in Section 4.1.2 that the ratings are biased to be higher for seen verb-argument-role triples. We therefore also expect a stronger correlation for data sets containing many seen verb-argument-role triples. The truly interesting question then is how well the models predict the unseen data points, which generally tend to be rated less plausible, but some of which still receive high plausibility ratings. As always, we correlate the models' predictions to the human judgements using Spearman's ρ.

Results and Discussion

Table 4.6 lists the results of correlating the semantic model predictions and the human data for the seen and unseen data points. The PropBank model covers 100% of the test set, while the FrameNet model stays slightly below that mark for the seen and unseen test sets, again due to the idiosyncrasies of the training data for one verb, as described in Section 4.1.2.

Clearly, the models reliably predict human data both for seen and unseen data points. Generally, as expected, the correlations for seen data are stronger than those for the complete data set, and as before, the correlation ρs are larger for the FrameNet models than for the PropBank models.

The separate analysis of the PropBank models' predictions for unseen data points

achieves better correlations to the human data than the overall result (see Table 4.4); the magnitude of the difference however appears to be largely an artefact of the rank correlation measure since it is vastly reduced if Pearson's r, a parametric correlation test, is employed. Note, however, that Pearson's r assumes a normal distribution of the data points, which is why it is not generally applied here.

Overall, the PropBank ρs for seen and unseen data are more homogeneous than the FrameNet ρs, which differ significantly between the seen and unseen data sets at $p < 0.05$, one-tailed. The PropBank models profit from seen verb-argument-role triples, but their performance for unseen triples is about the same as their overall performance for the complete data set. The large difference in the FrameNet models' performance for seen and unseen data presumably stems both from generally better performance for the seen data points and from some difficulty covering the unseen PropBank vocabulary, which causes lower performance on the unseen data set than on the complete data. Despite this difficulty, the FrameNet models still predict the unseen data on about the same level of ρ vales as the PropBank models predict the seen data.

It is clear from the significant correlations to both seen and unseen data points that both types of model do not just assign uniformly high or uniformly low predictions respectively, but that the smoothing methods we employ enable them to differentiate within the seen and unseen data points and predict graded human judgements.

4.1.5. Summary

Our experiments in Sections 4.1.1 to 4.1.3 have demonstrated that the FrameNet-based models, using the estimation and smoothing strategies determined in Chapter 3, reliably account for human judgements across a range of different data sets. Results for the PropBank models were either significantly less reliable or stemmed from a confound with test set biases. This fits well with the prediction based on the evaluation of clustering features in Section 3.4.1 that the PropBank verb classes could be prone to overfitting the development set.

The FrameNet models performed well for the Trueswell data, a literature data set with extreme plausibility properties and biased distributions of ratings, as well as for the Padó set of materials that were extracted from the training corpora and show a more balanced rating distribution. The Ferretti data set showed that the FrameNet models make reliable predictions also for adjunct roles, although their performance is higher for argument roles, which are more frequently attested in the training corpora.

The experiments also show that especially the FrameNet models generalise well to unseen data. The only precondition is that the verb has to be known, since thematic roles are assumed to be verb-specific in both sets of training data and therefore, no role-specific predictions can be made for unseen verbs. The models perform best for seen stimuli from the same genre as the training materials. This is not surprising, since the models are probabilistic and assign higher probability to more frequent, i.e.

seen, data points. The models are however able to differentiate within sets of seen and unseen data, demonstrating that class-based smoothing allows them to make graded predictions.

4.2. Comparison Against a Standard Role Labeller

We now turn to our second evaluation, the comparison of our semantic model with models that address two related tasks in computational linguistics. First, we consider a standard semantic role labeller. The assignment of thematic roles to verbs' arguments is a subtask of predicting the plausibility of verb-argument-role triples. In addition to defining the possible relations between a verb and its argument by assigning thematic roles, models trained on this task could in principle be able to also predict whether an argument is a plausible or implausible role filler by assigning the role with high confidence for plausible fillers and with less confidence for implausible fillers. In this section, we therefore compare our semantic model to a standard thematic role labeller, evaluating the models both on the task of assigning the preferred role to the verb-argument pairs in our test set and on the task of predicting verb-argument-role plausibility.

4.2.1. Semantic Role Labelling

The availability of sufficient amounts of annotated training data in the form of the FrameNet and PropBank corpora has allowed the application of supervised machine learning methods to the task of assigning the correct semantic roles to a verb's arguments.

Beginning with work by Gildea and Jurafsky (2002), who showed on an early, smaller release of FrameNet that automatic semantic role labelling was feasible, influential research by Surdeanu, Harabagiu, Williams, and Aarseth (2003) or Xue and Palmer (2004) explored useful features and established modelling procedures. There has been a large community interest in the task, as evidenced by its adoption as a shared task in the Senseval-III competition (FrameNet data, Litkowski, 2004) and at the CoNLL-2004 and 2005 conferences (PropBank data, Carreras and Márquez, 2004, 2005). Participants in these competitions have explored a range of machine learning models, information sources and pre- and post-processing procedures, further consolidating our knowledge about the task.

The task of assigning thematic roles to a verb's arguments is usually split into several steps. First, there may be a word sense disambiguation stage at which the correct verb sense is chosen. This is especially important for FrameNet annotation, of course. In the next step, a second model identifies the verb's arguments. This non-trivial task ideally requires a syntactic analysis of the input sentence and knowledge about the

Syntactic Features	Lexical Features
path argument – target verb	target verb token
voice	target verb word class
position	target verb/argument lemma
target verb subcategorisation	first/last word in argument phrase
argument phrase type	lemma+phrase type
argument governing category	target verb lemma+argument head
preposition head of PP	voice+position
target verb/argument POS	

Table 4.7.: Features used for the Standard Labeller.

verb's subcategorisation preferences to choose the correct nodes in the parse tree from the overwhelmingly large number of incorrect ones. Finally, the identified arguments are assigned a thematic role by a third model. Since the errors of probabilistic systems are additive along this pipeline, model performance can be drastically improved by specifying the correct verb sense and argument boundary information. However, a system operating on free text of course needs to perform well all stages.

In addition to a standard way of breaking up the task, a standard set of useful features has emerged. These features are mostly syntactic and lexical in nature, capturing regularities in the way a verb (in a certain sense) preferredly realises its arguments. There are global features describing the syntactic configuration of verb and argument, such as the path through a syntactic parse tree from the argument to the verb, the sentence voice (active or passive), the argument's position with regard to the verb or the verb's overall argument structure. More local features for each argument describe its phrase type and governing category or the preposition head of a PP. Finally, a set of lexical features describes the verb's and argument's parts of speech, as well as the lemmas of the argument head and the verb. This is the only type of feature that allows a standard role labeller to account for argument-specific role predictions.

4.2.2. The Standard Labeller

We base our standard role labelling system on the labeller described in Giuglea and Moschitti (2004) (see Giuglea and Moschitti (2006) for more recent results). At the time of the experiments, this was the only of the best-performing labelling systems made freely available by its authors upon request.[3] The labeller uses an SVM (support vector machine) learner, a group of learners that have proven well-suited to the role-labelling task (e.g., Pradhan, Hacioglu, Krugler, Ward, Martin, and Jurafsky, 2005). Another asset is the integration of a feature extraction tool which makes it easy to adapt the

original feature set to our requirements.

The labeller separates argument recognition and argument labelling as outlined above. Since we are aiming at building an un-tuned, standard labelling model, we restrict the feature space to the standard features that we described above; see Table 4.7 for a complete list of the features. Note especially that we do not use the additional features introduced by Giuglea and Moschitti which are based on the combination of information from PropBank and VerbNet for FrameNet classification.

The features listed in Table 4.7 are used to train a classifier for the final role labelling step only, since for our test set, the argument heads are provided directly and do not need to be automatically recognised. Note that we did not specify the verb sense for the labeller model. This is partly out of fairness towards our semantic plausibility model, which performs word-sense disambiguation itself, and partly because we test all verbs covered in the training data, even if they are not observed with the correct sense, so there is not always a correct sense available to be specified. If no verb sense is given, the labeller considers all possible roles for a verb-argument pair, because it cannot constrain the role set to the correct verb sense. This gives the labeller freedom to generalise in cases of sparse verb data, but on the other hand increases its risk of making inappropriate predictions.

The argument-labelling classifier is formed by a set of role-specific classifiers, one for each role, that decide whether an argument does or does not carry their role. The classifiers output a score with their decision, and the role with the highest positive score is selected as the role prediction for the argument.

The performance of thematic role labellers is usually given in terms of F score, which is a measure combining the labeller's accuracy, which is the percentage of correct role assignments in all assignments made, and its recall, which is the percentage of correct assignments in the number of input arguments, and differs from accuracy only if the labeller does not process all input arguments. F score is defined as the harmonic mean of precision and recall, as in Equation 4.1.

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4.1}$$

The standard labeller's F score on an unseen 10% of FrameNet data is $F = 80.5$ (using gold argument boundaries). This is a reasonable result given that the average F score achieved in the Senseval-3 competition by models employing more than the standard features was $F = 85$, also using gold boundaries. We also trained the labeller on the PropBank data, resulting in an F score of $F = 96.2$ on Section 23, the standard test set, and again using gold boundaries. Again, this is a good result compared to recently-published systems, which reach up to $F = 98.7$ on the role labelling task using (probably lower-quality) system boundaries (Che, Zhang, and Liu, 2006).

[3]Many thanks to Alessandro Moschitti and Ana-Maria Giuglea for the software and their friendly help.

4.2.3. Evaluation

Method

We evaluate the models on the McRae and Padó test sets. The Trueswell test set is not used because of its inherent bias towards the object role discussed in Section 4.1.1. The McRae and Padó sets are more informative of the models' ability to assess the plausibility of verb-argument-role triples.

To keep the results table as concise as possible, we compare the role labeller to the two best-performing instances of our semantic model (on the McRae test set) using the PB 3 PropBank model and the FN 3 FrameNet model.

We set two tasks: Role labelling and judgement prediction. In the role labelling task, the models have to predict the correct role for each verb-argument pair. We define the correct role label to be the one with the higher plausibility judgement. We assume the predicted role to be the most probable one in the semantic model and the one with the highest score for the standard labeller.

We formulate frequency baselines for role labelling based on our training data. For PropBank, always assigning the most frequent *Arg1* role results in $F = 45.7$ on the McRae test set and in $F = 38.6$ on the Padó set. For FrameNet, we assign each verb-argument pair the most frequent role given the verb, which places the baseline at $F = 34.4$ for the McRae set and at $F = 28.5$ for the Padó set.

For interesting comparisons, we test the difference between F scores for significance using a randomisation test. This type of test was advocated for F score significance testing by (Yeh, 2000), because it avoids certain independence assumptions. We use an implementation by Sebastian Padó, `www.coli.uni-sb.de/~pado/sigf`.

We also evaluate the models on the correlation task. Our semantic model's predictions for each of the two target roles are correlated to the human judgements as above. For the labeller, we normalise the role scores across all roles and then correlate the normalised scores for the target roles to the human ratings.

Recall that the labeller heavily relies on features extracted from parsed input. Also, it does not process its input incrementally. Therefore, we had to present the verb-argument pairs from the test set in full sentences to be able to extract all features for role labelling. However, in using sentence contexts, we potentially bias the standard labeller towards the label corresponding to the role which is implied by the syntactic structure. A reduced relative structure implies an object role, while a main clause structure implies a subject role. We therefore created both a reduced relative and a main clause sentence context for the verb-argument pairs (*The doctor cured by the ...* and *The doctor cured the ...*) and present the results for comparison. Recall that for our own model, we have never needed to specify the grammatical function linking verb and argument, which is the only model feature referring to the syntactic relationship between the verb and the argument. If no grammatical function is specified, the semantic model drops the

Train	Test	Model	Coverage	ρ		Labelling Cov.	Labelling F
PB	McRae	Baseline	–	–		100%	45.7
		SVM (subj)	100%	0.001,	ns	100%	45.7
		SVM (obj)	100%	-0.003,	ns	100%	45.7
		PB 3	98%	0.105,	ns	100%	**52.2**
	Padó	Baseline	–	–		100%	38.6
		SVM (subj)	100%	0.081,	ns	100%	54.6
		SVM (obj)	100%	0.061,	ns	100%	49.3
		PB 3	100%	**0.286**,	***	100%	**59.4**
FN	McRae	Baseline	–	–		100%	34.4
		SVM (subj)	88%	0.056,	ns	100%	40.6
		SVM (obj)	88%	0.116,	ns	100%	34.4
		FN 3	88%	**0.415**,	**	100%	**59.4**
	Padó	Baseline	–	–		100%	28.5
		SVM (subj)	99.5%	0.103,	*	100%	49.8
		SVM (obj)	99.5%	0.205,	***	100%	43.0
		FN 3	96.9%	**0.515**,	***	100%	**57.9**

Table 4.8.: Standard SVM role labeller and semantic model. Coverage, correlation strength (Spearman's ρ), labelling coverage and labelling F score for PB and FN data on the McRae and Padó test sets, best results in **bold face**.

grammatical function feature and bases its predictions only on the specified verb and argument (see Section 3.1.4).

Results and Discussion

Table 4.8 shows that our semantic model always performs as well or better than the SVM role labeller, both on the labelling and the judgement prediction task.

McRae set On the labelling task, our FrameNet semantic model outperforms the baseline and the role labeller by at least 18 points F score. For the PropBank data, the semantic model still numerically outperforms the standard labeller, even though the labelling performance of all models is more similar and close to the baseline. Due to the small number of stimuli in the McRae set, all differences in F score are however not significant.

The performance of the SVM labeller confirms the strong influence of syntactic features: On the PropBank test set, it assigns the *Arg0* label in the majority of cases if

the argument was presented as a subject and mostly the appropriate *ArgN* label if the argument was presented as an object. This results in F scores similar to the frequency baseline. On FrameNet, performance is better for the subject condition, where there is also a clear trend for assigning agent-style roles. (The object condition is less clear-cut.)

On the judgement prediction task, the labeller fails to achieve significant correlations with human data for either the PropBank or the FrameNet training set. Its role scores do not predict human plausibility judgements. Our FrameNet-based model reliably predicts the McRae test data, as above, and not surprisingly significantly outperforms the labeller at $p < 0.05$, one-tailed.

Padó set On the Padó test set, the labeller clearly profits from the larger amount of seen test stimuli. Despite the lower frequency baselines, its labelling F scores are better than those for the McRae set, and all the FrameNet and PropBank labeller models significantly outperform the baseline ($p < 0.001$, one-tailed). Both instances of the FrameNet-trained role labeller even achieve a significant correlation to human data on the judgement prediction task. It is likely that the FrameNet role labelling model makes greater use of the lexical features due to the general sparseness of all features. This would allow the FrameNet labeller to make use of the proportion of seen verb-argument pairs in the test data.

The semantic model still outperforms the labeller in both tasks, however. The PropBank model significantly predicts the human data where the labeller does not, and the FrameNet model's correlation ρ is significantly higher than the labeller's ($p \leq 0.001$, one-tailed). Even on the labelling task, both semantic models' F scores are significantly higher than those of the labellers with the object role bias (FN: $p < 0.01$, PB: $p < 0.05$, both one-tailed).

Discussion The role labeller's performance both on the judgement prediction task and on the labelling task is clearly worse than our semantic model's. The labeller's F scores are much lower on the McRae and Padó test sets than on the standard role labelling test sets because of its strong reliance on syntactic cues, which may be unreliable during the processing of local syntactic ambiguities, as they were for our test data. Since the labeller uses no features that efficiently take word-specific plausibility into account, it usually assigns the same role to both arguments of a verb, precluding a significant correlation with the human ratings. The reliance on global syntactic features also makes the role labeller unsuitable for incremental processing, as roles can only be assigned once the complete sentence is known.

The success of our semantic model both on the labelling and on the judgement prediction task stems partly from the absence of global syntactic features that bias the standard labeller strongly. Instead, our semantic model successfully relies on argument-specific plausibility estimates furnished by class-based smoothing. It is

also able to make predictions for incomplete sentence structures during incremental processing. Our joint probability model has the further advantage of being conceptually much simpler than the SVM labeller, which relies on a sophisticated machine learning paradigm, and of requiring the computation of only about one-fifth of the number of SVM features.

4.3. Comparison to Selectional Preference Models

A second task from computational linguistics that is related to plausibility prediction is the inference of verb selectional preferences. Selectional preference models characterise the role fillers or classes of role fillers that verbs prefer for each of their argument slots. This task exactly corresponds to estimating the plausibility of a verb-argument-relation triple once the relation has been specified. This suggests combining a selectional preference account with a role labeller to solve the judgement prediction task; in our comparison, we specify the gold target roles, which allows us to estimate the upper bound of such a combined model's performance.

Argument slots in selectional preference models are normally defined by grammatical functions, which we have argued above are too coarse-grained to capture the possible relations between verb and argument. For comparison to our semantic model, we train the selectional preference models using the more fine-grained thematic roles as labels for the argument slots.

Evaluation results against human data from Resnik (1996), Keller and Lapata (2003) and, for German, Brockmann (2003) have shown that selectional preference models can successfully predict human plausibility ratings when using large amounts of training data and defining verb slots as grammatical functions. In contrast, in our evaluation, we define verb slots by thematic roles, which are more fine-grained, and at the same time have less training data available. Our evaluation will show how well the selectional preference models perform at deriving preferences for more fine-grained relation distinctions from less training data. We test the selectional preference models on three test sets with different characteristics: We first compare model performance on the McRae and Trueswell test sets. The Trueswell plausibility manipulation rests on animacy, a high-level concept that defines a large set of plausible and a large set of implausible fillers and that is implicitly represented in the concepts of the WordNet hierarchy that the selectional preference models use for smoothing. This data set should therefore be easier to model for selectional preference models than the McRae data set, which requires the identification of much smaller classes of acceptable arguments which may be sparsely represented in the training data. Finally, we evaluate the role of data sparseness in model performance by testing on the Padó set with its larger amount of seen test stimuli.

4.3.1. Selectional Preference Models

Models of selectional preference learn which sets of arguments a verb prefers in each of its argument slots in order to determine how well a new argument fits into a specified slot. This amounts to evaluating the fit of a verb-argument-relation triple for a given relation. Generally, selectional preference models define a verb's preference for specific arguments over word classes instead of words, for reasons of efficiency and sparse data. Therefore, the first step in inferring preferences is to identify a relevant set of argument classes. Then, the verb's preferences over this set of classes are induced, which specify for each class how well its members fit the verb as arguments in a specified relation. Finally, the fit of a verb-argument-relation triple is computed by identifying the correct word class for the argument and then using the knowledge about relation- and verb-specific preferences for that class.

The majority of models uses the WordNet noun hierarchy to furnish the word classes, while differing in how the relevant subset of classes is determined. This is the class of models that we focus on here. One of the earliest and most influential approaches to modelling selectional preferences is Resnik (1996). To specify verbs' constraints over their arguments, Resnik first computes the selectional preference strength of a verb (i.e., the amount of constraints it puts on its arguments overall) and then specifies how much of that selectional strength applies to each possible argument class in WordNet.

The selectional preference strength S of a verb is quantified as the difference (in terms of the Kullback-Leibler divergence) between the prior distribution of argument classes when no verb is taken into account ($p(c)$) and the distribution of argument classes given the verb ($p(c|v)$), as shown in Equation 4.2.

$$\begin{aligned} S(v) &= D(p(c|v)||p(c)) \\ &= \sum_c p(c|v) log \frac{p(c|v)}{p(c)} \end{aligned} \tag{4.2}$$

A predicate that imposes strong constraints on its arguments will have a probability distribution over argument classes that strongly diverges from the prior distribution.

The selectional association $A(v, c)$ between a verb and an argument class is the ratio of the verb's selectional preference strength for this class normalised over the verb's overall selectional preference strength, as shown in Equation 4.3.

$$A(v, c) = \frac{p(c|v) log \frac{p(c|v)}{p(c)}}{S(v)} \tag{4.3}$$

The selectional association specifies how much of a verb's overall preference strength is contributed by the class c, and thus helps identify strongly preferred classes. The selectional association between a verb and a class can be negative, which indicates that

the class contains a dispreferred set of arguments.

The selectional preference between a verb and a specific argument head is taken to be the selectional association between the verb and the WordNet class that is the most strongly associated parent class of the argument. Refinements to this approach have been proposed by Abney and Light (1999) and, more successfully, Ciaramita and Johnson (2000), who address the handling of ambiguous nouns during the inference of the argument class distributions.

Li and Abe (1998) propose an alternative approach that focuses on pre-selecting a subset of WordNet classes for the computation of argument fit instead of considering all WordNet classes like Resnik's approach. The selection aims to balance the conflicting constraints of generality of the classes, which allows smoothing, and specificity, which is a bias towards retaining fine-grained class distinctions and thereby helps to avoid over-generalisation.

Li and Abe treat the WordNet class hierarchy as a tree and prune away subtrees, using the leaf classes of the resulting tree as the classes over which the verb's preferences are defined. The set of leaf classes to be chosen is determined by applying the information-theoretic principle of *Minimum Description Length* (Rissanen, 1978), which states that the best probability model given a data set is the model for which the encoding of the model and of the data is shortest in the number of bits used. The probability model in this case is the probability distribution over the set of leaf classes given a verb and relation. The description length of the data is minimal if the model is maximally specific, while the description length of the model is minimal if the class set is maximally general. The optimal set of tree classes given these conflicting constraints is found by simultaneously minimising the description length of both the model and the data, which leads to a solution that balances generality and specificity.

Once the relevant set of classes is found, the fit of a verb and argument head can be determined by computing the conditional probability of the argument's parent class in the set given the verb and relation, and uniformly distributing this probability over all nouns in the class.

Finally, we consider a method proposed by Clark and Weir (2002) which, unlike the other two, does not intend to accrue abstract information about which WordNet classes contain preferred relation fillers for a verb. Rather, it solely aims at estimating as accurately as possible $P(argument|verb, relation)$, the fit of verb slot and argument given the training data. Consequently, Clark and Weir use the WordNet hierarchy primarily for smoothing. They aim to select the most general class for each argument, with the constraint that the distribution of the class members must still well approximate the argument's co-occurrence with the verb slot. Since more general classes have more members and therefore are able to furnish more co-occurrence counts for probability estimation, $P(class_a|verb, relation)$ then is used as an estimate of $P(argument|verb, relation)$ which is less noisy and more reliable than the estimate based on the co-occurrence of only the argument lemma with the verb. The selected

class is not assumed to be the optimal level of abstraction at which to describe the verb's preferred arguments. It is only used because it allows the fit estimate between verb, argument and relation to be as faithful to the evidence from the training data as possible.

The optimal class is determined as follows: The algorithm moves up the hierarchy from the leaf class containing the argument to more general classes, and tests whether the member concepts in the current class are still distributed in the training data in a similar way as the argument. The search stops immediately once the concepts in the current class are distributed significantly differently, and the last class that was a good approximation of the argument's distribution is used to estimate the selectional preference of the verb for the argument in the specified relation. Both the statistical test used for significance testing and the assumed level of significance are parameters of this method.

4.3.2. Evaluation

Method

We compare the semantic model's performance on the judgement prediction task to the three WordNet-based selectional preference models introduced above, defining the verb relations as thematic roles. We use an adapted implementation of the three models from Brockmann (2003)[4], who was able to show that the selectional preference methods significantly predict human judgements for verb-argument-role triples extracted from a German corpus.

We train all models both on the FrameNet and PropBank corpora and evaluate using three test sets. The McRae and Trueswell test sets (see Sections 3.2.2 and 4.1.1, respectively) allow us to gauge the models' ability to capture selectional preferences of different fineness. The Trueswell data can be predicted by a model that differentiates between animate and inanimate arguments, while the McRae data can only be predicted by a model that captures preferences for smaller, more constrained classes of arguments. In addition, by using the Trueswell set we attempt to replicate results from Resnik (1996), who showed that his model significantly predicts the human ratings in this set. Finally, the Padó set (Section 4.1.2) shows how the models perform once more seen test stimuli are available.

The parameters of the Clark and Weir method were set on the McRae development set. We use the χ^2 test at a significance level of $p = 0.005$ for PropBank and $p = 0.3$ for FrameNet. As always, only verbs seen in the training data were tested.

Unlike our semantic model, the selectional preference models do not generate predictions for the verb's whole role set, but are trained to make predictions for a given role. They were consequently trained to predict the likelihood of the argument filling either

[4]Many thanks to Carsten Brockmann for kindly making his software available.

of the two rated roles. As above, we compare against the semantic model instances PB 3 and FN 3.

Results and Discussion

Table 4.9 shows that the selectional preference models achieve good coverage of all three test sets: The Trueswell set seems difficult to cover given the FrameNet training data, but using just carefully selected noun classes for smoothing, the selectional preference models generally cover only slightly fewer data points than our semantic model does due to GT smoothing. We have seen the semantic model reach similar coverage when using the WordNet top-level ontology as noun classes for smoothing, but at the cost of low prediction quality (see Section 3.4.2). The selectional preference methods in contrast do achieve significant predictions at least for some test sets. We now turn to evaluating their performance on the plausibility prediction task, first for the Trueswell and McRae sets and then for the Padó set.

Trueswell and McRae Sets Comparing the Trueswell and McRae results, all models clearly perform better on the Trueswell set, which only requires the identification of a preference for the large class of inanimate nouns. The Resnik model achieves significant correlations to this data set for both training corpora, and the Clark and Weir as well as the Li and Abe models significantly predict the human data when trained on the PropBank. The overall better performance of the PropBank-based models is probably partly due to the larger size of the PropBank resource (see also the discussion of performance on the Padó set below) and partly to the bias towards the object role in both the PropBank and the Trueswell data (recall the discussion in Section 4.1.1).

Our models also significantly predict the human judgements for the Trueswell data, with a higher correlation ρ for the FrameNet-based model, which suggests that they are able to cope better with the smaller training set and the absence of a general role bias than the selectional restriction models. The PropBank-trained semantic model performs much the same as the Clark and Weir and Li and Abe models, numerically below the Resnik model. The difference in correlation ρ between the semantic model predictions and the Resnik model results is however not significant for either training corpus.

Our evaluation on the Trueswell set replicates the evaluation in Resnik (1996), where the selectional preference model was trained on data from the Brown corpus (Francis and Kučera, 1964) and where argument slots were defined as grammatical functions. Predictions were found to be correlated to the human data with $r = 0.46, p < 0.02$, at about 90% coverage. Using the PropBank training data, Resnik's model achieves comparable results to the original evaluation, even though we tested on the finer-grained thematic role slot labels instead of grammatical functions. The Resnik model's somewhat lower performance using FrameNet again is probably due to FrameNet's

Training	Test	Model	Coverage	ρ	
FN	Trueswell	FN 3	81%	**0.523**,	***
		Resnik	61%	0.382,	*
		Clark&Weir	74%	0.154,	ns
		Li&Abe	76%	0.102,	ns
	McRae	FN 3	88%	**0.415**,	**
		Resnik	81%	0.025,	ns
		Clark&Weir	81%	-0.038,	ns
		Li&Abe	88%	-0.056,	ns
	Padó	FN 3	97%	**0.515**,	***
		Resnik	93%	0.031,	ns
		Clark&Weir	93%	0.165,	**
		Li&Abe	97%	0.112,	*
PB	Trueswell	PB 3	100%	0.334,	**
		Resnik	93%	**0.504**,	***
		Clark&Weir	97%	0.338,	**
		Li&Abe	96%	0.289,	*
	McRae	PB 3	98%	0.105,	ns
		Resnik	85%	0.047,	ns
		Clark&Weir	100%	0.157,	ns
		Li&Abe	89%	**0.202**,	ns
	Padó	PB 3	100%	**0.286**,	***
		Resnik	96%	0.227,	***
		Clark&Weir	98%	0.254,	***
		Li&Abe	100%	0.217,	***

Table 4.9.: Selectional preference methods and our semantic model (verb class set 3 PB/FN). Coverage and correlation strength (Spearman's ρ) for both training corpora on the McRae, Trueswell and Padó test sets, best result in **bold face**.

smaller size and the absence of a strong bias towards object roles as in PropBank. The model however still copes much better with these disadvantages than the other selectional preference models.

None of the selectional preference models, however, is able to capture the more fine-grained selectional restrictions used to manipulate filler plausibility in the McRae test set. Using PropBank for training, none of the four models achieves a significant correlation with the human data. Our FrameNet semantic model is the only one to significantly predict the McRae data. It significantly outperforms the selectional preference models, which all show correlation ρs around zero ($p < 0.05$, one-tailed).

Padó Set Examining the predictions of the selectional preference methods for the McRae and Trueswell sets item by item, we find that the models fail because they routinely predict the same noun class for both arguments of a verb, which causes them to make exactly the same plausibility prediction for the arguments. This is probably a problem of sparse data, which precludes the learning of verb preferences for small, highly constrained sets of role fillers. The results on the Padó set support this hypothesis: The Clark and Weir and Li and Abe models consistently perform well, independent of the training set. Since the Padó set does not show a pronounced role bias, this good performance is probably rather due to the relatively large portion of seen test stimuli present in the test set.

The relatively large proportion of seen test stimuli even allows the selectional restriction models to significantly predict the human judgements, even for the FrameNet data, despite the smaller size of the resource that seems to preclude similar performance for sparser test sets. However, the FrameNet-trained semantic model still significantly outperforms the Clark and Weir model, which is the best of the selectional restriction models trained on FrameNet ($p < 0.001$, one-tailed), while the PropBank-trained semantic model performs indistinguishably from the selectional preference models.

The Resnik model, which yielded comparably high correlation coefficients for the Trueswell set with both training corpora, however performs noticeably worse on the Padó data. For PropBank, its performance on the Padó set is close to that of the semantic model, but for FrameNet, the correlation ρ is around zero, as for the McRae data. This is probably the case for two reasons: First, the Padó set, like the McRae set, does not vary plausibility on the animacy level, but requires models to learn preferences for smaller sets of arguments. The Resnik model appears to have difficulty capturing selectional preferences at this level. Second, the Resnik model performs well on completely unseen test sets when the selectional restrictions are relatively coarse-grained or when the larger PropBank training set is used. This suggests that it is dependent on the availability of large amounts of training data, and more specifically that its estimation method profits more from a lot of training evidence for the verb's preferred argument classes than from having encountered a specific test stimulus during training.

Discussion The semantic model appears to make better use of the limited amounts of training data than the selectional preference models which perform noticeably better when trained on the larger PropBank training set than when trained on FrameNet. The reverse is true for the semantic model: It performs on a par with the selectional restriction models when trained on PropBank, but much better than the selectional restriction models when trained on FrameNet.

The semantic model's good performance is carried mainly by smoothing using induced verb classes. These appear to generalise better when induced from FrameNet data, which also explains the performance gap between the two versions of the semantic model. Unlike the selectional preference models, the semantic model hardly relies on noun generalisations, since we use only the lowest available level of noun classes that furnishes very sparse, but also very specific and reliable information. Using the noun classes increases the correlation ρs by only about 0.02. This underlines the effectiveness of the verb generalisations, but given that the selectional preference models reach good coverage and acquire coarse selectional preferences with a careful selection of noun classes only, it could be worthwhile to refine the noun class selection for our semantic model in future work.

We found that existing selectional preference methods show some ability to predict selectional preferences for large argument classes like animates, but that they are largely unable to identify more fine-grained selectional restrictions for verbs' role fillers. The most striking example for this result is the Resnik model, which achieves the highest correlation ρ of all PropBank models for the Trueswell data set, but performs much worse for the Padó and McRae sets given either training set. The most consistently performing selectional preference models are those by Li and Abe and Clark and Weir, trained on PropBank.

Since we have specified the correct thematic roles to the selectional preference models, these results constitute an upper bound for the performance of a combined plausibility model that uses a role labeller to specify which relations between a verb and argument a human might assume and a selectional preference model to evaluate the plausibility of the verb-argument-role triples. Clearly, such a model does not promise superior performance than our semantic model. In addition, the tasks of identifying a set of applicable thematic roles for a verb-argument pair and estimating their plausibility are mutually dependent, which means that it advantageous to employ a model like ours which solves them simultaneously.

4.4. Summary and Discussion

In this chapter, we provided comprehensive evaluation of the semantic model, showing that it reliably predicts human data across a range of different data sets and that it outperforms two existing related approaches on the judgement prediction task.

Throughout the evaluations, we observed that the semantic model performs much better when trained on FrameNet data than when trained on PropBank. Therefore, we will only use the FrameNet model below. This pattern of performance is not surprising given the insights from the evaluation of clustering features in Section 3.4.1, where we concluded that the FrameNet classes are based on more robust semantic generalisations than the PropBank classes. We predicted those to be susceptible to overfitting the development set due to their reliance on sparse and non-generalisable features.

Section 4.1 showed that the FrameNet model is able to predict argument roles from test sets with different characteristics as well as adjunct roles, while correlation ρs were generally higher for the argument roles, which were more frequent in the training data. We also ascertained that the model makes correct predictions for unseen verb-argument combinations, and that its performance does not rest on assigning invariably high probabilities to seen triples and invariably low probabilities to unseen triples, as could be expected from a purely probabilistic model. This behaviour is made possible by our class-based smoothing approach, which allows the semantic model broad coverage of unseen triples as long as the verb in the triple is known. Otherwise, the correct set of verb-specific roles cannot be determined.

We compared our semantic model to a standard role labeller in Section 4.2. Our model showed more ability to take plausibility into account and was less dependent on syntactic features, which allowed it to easily outperform the labeller on the judgement prediction task (again when trained on FrameNet). On the labelling task, the semantic model performed better or at least as well as the labeller, which did much worse on our test set than on an unseen split of the training data because of its reliance on global syntactic features and, in consequence, its inability to account for semantic plausibility of role assignments.

The comparison to a range of selectional preference methods in Section 4.3 showed that the semantic model also easily outperforms the selectional preference approaches when trained on FrameNet data, and still performs as well as the selectional preference approaches on PropBank data. The selectional preference methods showed some ability to learn coarse selectional preferences that differentiate between the large classes of animate and inanimate arguments, but they did not reliably learn preferences for smaller argument classes. Sparse data appeared to be the main reason for this problem, since the models generally performed better on the Padó data set that contains more seen stimuli. We therefore conclude that our model makes better use of the limited training data.

We have also argued that the conceivable strategy of combining a labelling and selectional restriction model cannot lead to better performance than that which we have seen for the selectional preference models alone, and that it is therefore no viable alternative to our model. In sum, the semantic model has proven to robustly predict human judgement and is clearly better suited as a tool for solving this task than either of the related approaches, alone or in combination.

5. The SynSem-Integration Model

Having introduced and evaluated the semantic model in Chapters 3 and 4, we describe in this chapter how it is combined with a syntactic parser to form the SynSem-Integration model proposed in Chapter 2. The SynSem-Integration model's syntactic model is realised by an incremental probabilistic parser, which we describe in Section 5.1. In Section 5.2, we outline how the semantic model is extended to enable the processing of whole sentences with multiple arguments for the same verb. We restrict ourselves to using the FrameNet-trained version of the semantic model, since it has proven superior the PropBank model during evaluation.

The SynSem-Integration model uses the syntactic and semantic models to evaluate possible analyses of the input, find a globally preferred analysis and explain observed processing difficulty. In Section 5.3, we formulate several possible implementations of the conflict and revision cost functions, which were outlined in Chapter 2. We then describe the parameter selection process for the SynSem-Integration model, in which we choose the interpolation factor for combining the syntactic and semantic evaluations of an analysis into a global preference ranking, as well as the best-performing combinations of the cost functions.

In Chapter 6, we will evaluate the SynSem-Integration model against experimental observations for four sentence processing phenomena: The Main Clause/Reduced Relative (MC/RR) ambiguity, the NP object/Sentential Complement (NP/S) ambiguity, the NP object/0 (NP/0) ambiguity and the PP Attachment ambiguity.

5.1. The Syntactic Model

In Chapter 2, we have proposed to instantiate the syntactic model within the SynSem-Integration model with a parser that uses a probabilistic context-free grammar (PCFG). This parser should incrementally assign syntactic analyses to the input and rank them by their probability, to provide an instance of a probabilistic grammar-based model (Jurafsky, 1996, Crocker and Brants, 2000). We first give a short introduction to relevant aspects of syntactic parsing in Section 5.1.1 before describing the parser we use (Section 5.1.2) and evaluating different model parametrisations in Section 5.1.3. There, we ascertain that the syntactic parser assigns correct analyses to unseen test data, and especially that it correctly treats the syntactic structures involved in the ambiguity phenomena we will model in Chapter 6.

S	→	NP VP	1.0	V	→	saw	.8
NP	→	DT N	.6	V	→	shot	.2
NP	→	DT N PP	.4	DT	→	the	1.0
VP	→	V NP PP	.3	N	→	cop	.4
VP	→	V NP	.7	N	→	crook	.3
PP	→	PRP NP	1.0	N	→	gun	.2
				N	→	telescope	.1
				PRP	→	with	1.0

Figure 5.1.: Example of a PCFG: LHS → RHS rules annotated with rule probabilities.

5.1.1. Syntactic Parsing

Currently, the majority of computational approaches to wide-coverage syntactic parsing relies on probabilistic context-free grammars (PCFGs). A PCFG consists of a set of terminal symbols (i.e., words) T, a set of non-terminal symbols (i.e., part-of-speech tags and phrase symbols) NT and a set of context-free rules of the form shown in Equation 5.1:

$$NT \rightarrow (NT \mid T)^{+} \tag{5.1}$$

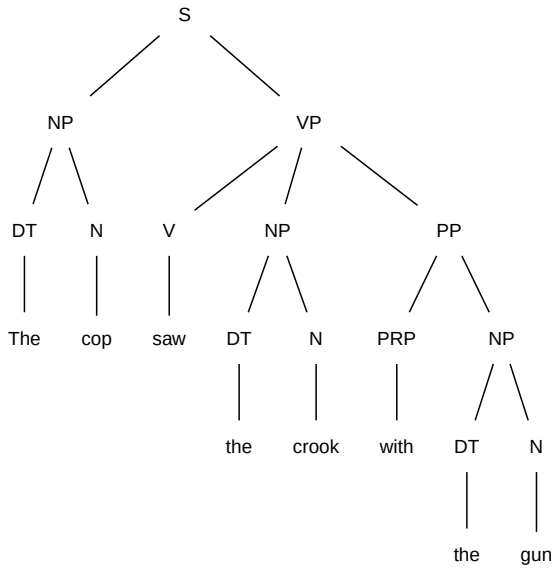
A non-terminal on the left-hand side (LHS) of a rule can be rewritten as a sequence of non-terminals and terminals on the right-hand side (RHS) of the rule. Each of these grammar rules is annotated with a probability $P(RHS|LHS)$. This probability represents the likelihood of expanding the category on the LHS to the categories on the RHS. In order to obtain a mathematically sound model, the probabilities for all rules with the same left hand side have to sum to one. An example for a PCFG covering a tiny fragment of English is given in Figure 5.1.

Given a grammar and an input sentence, a parser can derive a syntactic analysis of the input. Figure 5.2 shows two example tree structures for the input sentence *The cop saw the crook with the gun*. In the parse trees, the application of a grammar rule corresponds to a tree node and its daughter nodes: The $S \rightarrow NP\ VP$ rule is reflected in the top node S and its daughter nodes NP and VP. This correspondence can also be used to induce rule probabilities from syntactically annotated corpora using the MLE approach.

Each parse tree is associated with a probability that is derived from the grammar rules involved in creating the tree. The tree probability $P(T)$ is the result of multiplying up the probabilities of all applied rules, as shown in Equation 5.2:

$$P(T) = \prod_{rule \in T} P(rule) \tag{5.2}$$

$$\begin{aligned} P(T_1) &= 1.0 \cdot .6 \cdot 1.0 \cdot .4 \cdot .3 \cdot .8 \cdot .6 \cdot 1.0 \cdot .3 \cdot 1.0 \cdot 1.0 \cdot .6 \cdot 1.0 \cdot .2 \\ &= 0.0007 \end{aligned}$$



$$\begin{aligned} P(T_2) &= 1.0 \cdot .6 \cdot 1.0 \cdot .4 \cdot .7 \cdot .8 \cdot .4 \cdot 1.0 \cdot .3 \cdot 1.0 \cdot 1.0 \cdot .6 \cdot 1.0 \cdot .2 \\ &= 0.002 \end{aligned}$$

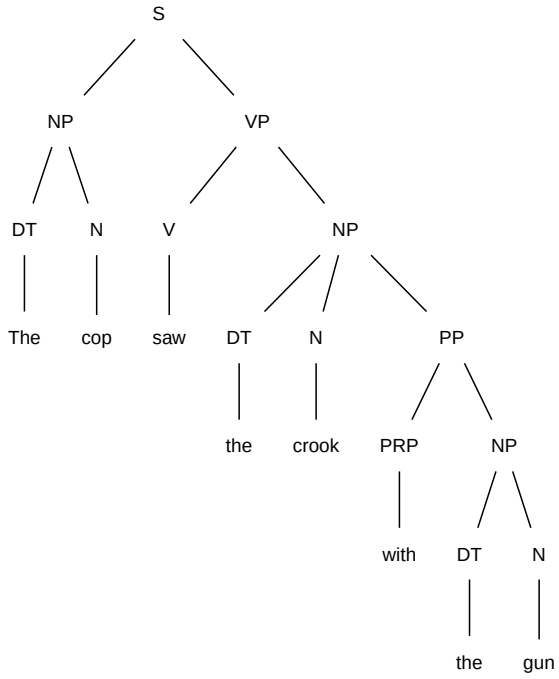


Figure 5.2.: Trees with tree probabilities generated by the example grammar

Figure 5.2 also gives the tree probabilities associated with the parse trees given the rule probabilities in Figure 5.1.

Structural Ambiguity and Human Preferences

With any grammar of practicable coverage, a large number of alternative parse trees can be derived for any input sentence. Using a probabilistic grammar, we can however predict which of the many syntactic analyses is the best one by simply selecting the tree with the highest probability, as shown in Equation 5.3.

$$Preferred\ Tree = \underset{tree}{argmax}\ \ P(tree) \tag{5.3}$$

The preferred tree is likely because it contains grammar rules with high probabilities. If the grammar rule probabilities were induced from an annotated corpus, high rule probability indicates a frequently used rule.[1] Choosing the tree with the highest probability therefore amounts to choosing a tree that uses structures which were frequent in prior language experience.

The fact that probabilistic grammars allow us to rank analyses and select a preferred one can be exploited for psycholinguistic modelling under the additional assumption that the rule probability information in the grammar correlates to human structural preferences. It may not be altogether unproblematic to make this assumption, since the grain size of grammatical frequency information that humans employ may not exactly correspond to that of PCFGs, which are restricted to accruing information solely on the level specified by the grammar rules and cannot easily take larger chunks of structure or non-structural information into account (recall Section 2.1 and also see Mitchell, 1987, for an early discussion). However, the success of PCFG-based models like Jurafsky (1996) or Crocker and Brants (2000) relativises this concern.

In order to make the rule probabilities of a PCFG-based parser as similar to human structural preferences as possible, grammar rules and probabilities are induced from large syntactically annotated corpora. This approach also has the advantages of quickly yielding grammars with good accuracy and large coverage of unseen data points, and has therefore become a default practice in computational linguistics, independent of psycholinguistic considerations.

Disambiguation through tree probabilities is illustrated by the two parse trees derived for the input sentence *The cop saw the crook with the gun* shown in Fig 5.2. The two readings (the correct one in T_2, where the PP modifies *the crook* and an incorrect one, where the PP modifies the verb) involve different grammar rules and therefore differ in their overall probabilities. The NP attachment reading T_2 is correctly predicted to be

[1] Note that, all else being equal, the number of grammar rules involved also plays an important role. It follows from Equation 5.2 that trees with fewer rule applications are more likely than larger trees.

preferred over the verb attachment reading T_1 because the grammar encodes a general bias against PP modification of the verb (P($VP \rightarrow V\ NP$) > P($VP \rightarrow V\ NP\ PP$)). Note that this relatively coarse bias does not consider preferences introduced by the lexical material involved, for example syntactic subcategorisation preferences of the verb (Jurafsky, 1996) or semantic plausibility of the analysis. The grammar would still prefer the NP attachment reading for input like *The cop shot the crook with the gun*, where verb modification is presumably preferred due to the verb's subcategorisation preferences, and it would also prefer the NP attachment for *The cop saw the gun with the telescope*, where plausibility introduces a strong bias towards the verb attachment.

Including Lexically-Specific Information

There are several strategies that make it possible to integrate lexically-specific information into PCFGs. Two of these are *lexicalisation* of the grammar (Jelinek, Laerty, Magerman, and Roukos, 1994, Collins, 1996) and the addition of *grammar subcategories* (Johnson, 1998, Klein and Manning, 2003). We apply the latter strategy primarily to addresses the lack of verb subcategorisation information, but can be used more generally to avoid making unjustified independence assumptions between rule applications in PCFGs. The former strategy annotates grammar categories with their head words and allows the general use of co-occurrence information between head words in the grammar. For English at least, it yields such an improvement over unlexicalised grammars that it has become a quasi-standard. Given coverage in the training data, head-head co-occurrence can be used to make plausible attachment decisions and can therefore possibly act as a substitute for semantic evaluation of the attachment. Recall, however, that at least for our test data in Chapter 4 direct co-occurrence of verb and argument head was extremely rare. This is underscored by the results of Gildea (2001), who found that eliminating bi-lexical dependencies from a lexicalised grammar hardly hurts performance, especially on test data that is different in genre from the training data. In sum, a lexicalised parser cannot be expected to make valid plausibility decisions based on head-head co-occurrences due to data sparseness. Therefore, the SynSem-Integration model employs an independent, carefully smoothed semantic model for the semantic evaluation of syntactic structures.

Figure 5.3 demonstrates the two strategies. Lexicalisation of the grammar (top) extends each grammar rule by adding the lexical head to each left-hand side and right-hand side non-terminal (only demonstrated for a subset of heads for concise presentation). This drastically increases the number of grammar rules, but allows the grammar to capture fine-grained lexical preferences for each head, including head-head co-occurrence.

The addition of subcategories (Figure 5.3, bottom) codes lexical preferences, for example for subcategorisation frames, in the pre-terminal grammar symbols. This strategy allows each verb in the example to select the verb category that encodes its

S[see]	→	NP[cop] VP[see]	1.0
S[shoot]	→	NP[cop] VP[shoot]	1.0
VP[see]	→	V[see] NP[crook] PP[gun]	.3
VP[see]	→	V[see] NP[crook]	.7
VP[shoot]	→	V[shoot] NP[crook] PP[gun]	.7
VP[shoot]	→	V[shoot] NP[crook]	.3
V[see]	→	saw	.3
V[see]	→	sees	.2
	⋮		
S	→	NP VP	1.0
VP	→	$V_{NP,PP}$ NP PP	.6
VP	→	V_{NP} NP	.4
$V_{NP,PP}$	→	saw	.3
V_{NP}	→	saw	.7
$V_{NP,PP}$	→	shot	.7
V_{NP}	→	shot	.3
	⋮		

Figure 5.3.: Extending PCFGs: Lexicalisation (top), addition of subcategories (bottom).

preferred subcategorisation frame. In the other rules, the grammar generalises across all verbs that show similar subcategorisation preferences, which is not immediately possible in the lexicalised grammar.

Both lexicalisation and the addition of grammar subcategories enlarge the number of grammar rules, which makes careful smoothing necessary. This is especially true for lexicalised grammars, where many head-head dependencies encountered in the input will be unseen in the training data. A standard approach to solving this problem is *back-off smoothing*, which allows the parser to back off to a less specific (e.g., unlexicalised) grammar rule in case no lexicalised rule exists. Another very efficient way of smoothing a grammar is Markovisation (Collins, 1997). This approach computes the probability of a rule's left-hand side given its right-hand side as a Markov chain of conditional probabilities, where for each daughter only the mother node and a specified number of sisters is considered, and the Markovian independence assumption is made with regard to the other sisters. In this way, grammar rules can be generated "on the fly" even for unseen sequences of daughters, as long as the sequence is reasonable given prior experience of partial daughter sequences.

5.1.2. The Parser

As outlined in Section 2.5, we wish the syntactic model to fulfil three requirements: Wide coverage of unseen utterances, the derivation of the model from language data and the ability to process input incrementally. A number of highly accurate, wide-coverage syntactic parsers that use corpus-derived grammars have been proposed and implementations are available. Two standard models are, for example, those of Collins (1997) and Charniak (2000). However, these parsers do not support word-by-word incremental processing, but rather rely on the availability of the complete input string when processing begins.[2] Therefore, we use the incremental parser proposed by Roark (2001)[3]. It is able to incrementally output syntactic analyses for partial input, and its grammar and lexicon are induced from a large corpus, which makes it both an experience-based model and gives it wide coverage of unseen text.

Roark's parser employs a *top-down* parsing strategy, where a parse tree is constructed by expanding the grammar rules left to right, starting with an S rule. With this parsing strategy, normally structure is generated predictively even if no input supports it yet. For example, the expansion of the $VP \rightarrow V\ NP\ PP$ rule from Figure 5.1 to parse a verb seen in the input predicts the existence of two syntactic arguments that are not yet supported by the input (and possibly never will be, if the input turns out not to contain a PP).

The parser achieves incremental processing without prediction of such unsupported structure by grammar factorisation, such that a rule expansion can leave the rightmost daughter or daughters unspecified until they are supported by more input. To return to the example, this means that the parser initially stipulates only the V child of the VP rule, effectively leaving the choice between the two VP rules underspecified until more input is encountered.

The top-down parsing strategy is known to get caught in an infinite loop if the grammar contains left-recursive rules like $NP \rightarrow NP\ PP$ which can be applied over and over again to expand their own right-hand sides. To avoid this problem, a second grammar transformation allows the parser to selectively treat recursive cases in the way a *left-corner* parser would, namely by combining the inference of structure from the input words with top-down structural prediction. This strategy limits the number of times a recursive rule can be applied because predicted top-down structure is immediately validated against the input.

Roark (2001) reports optimal parser performance when using a lexicalised grammar with Markovisation for smoothing. The left-corner transformation in practice only has to be applied to the grammar rules that cover NPs.

[2] Stolcke (1995) shows how such parsers can compute incremental output, but no implementation is available for the Collins (1997) and Charniak (2000) parsers.

[3] We would like to thank Brian Roark for kindly making his software available and even adding functionality for us.

From a psycholinguistic point of view, parsing strategies differ with regard to their cognitive plausibility. A plausible strategy should require little memory for processing structures that people read easily, and much memory for structures that are difficult for people. The human pattern of difficulty is that left- and right-recursive structures, such as examples (5.4) and (5.5), but that deeply centre-embedded structures like (5.6) are hard.

(5.4) (((Tom's) mother's) house)

(5.5) (This is the man who found (the cat that ate (the mouse that died)))

(5.6) (The man (whom the dog (that the cat chased) hated) slept)

While pure top-down parsing has high memory requirements for both left-recursive and centre-embedded structures, left-corner parsing[4] is a cognitively plausible parsing strategy because it encounters high memory load only for centre-embedding materials (Abney and Johnson, 1991, Resnik, 1992). By selectively using a left-corner strategy to avoid infinite memory requirements for left-recursion, Roark's top-down parser thus shows the same memory load profile as human readers.

For our experiments, we prefer to work with a parser that does not use head-head co-occurrence information to ensure that decisions about semantic plausibility are only made by the semantic model. No large drop in performance is expected in comparison with a fully lexicalised model according to Gildea (2001), since the the parser is trained on a newspaper genre training corpus, but finally used on psycholinguistic experimental items.

5.1.3. Evaluation

We compare two instantiations of a parser without bilexical dependencies: The *No Bilex* parser is partially lexicalised in that it uses head information of the current category when proposing non-terminal sister categories, but it contains no bilexical dependencies that would for example link to the heads of sister categories. This restricted lexicalisation still provides information about verb subcategorisation preferences and other information that conditions on the current head word.

The *Unlex SC* parser introduces verb subcategorisation information by adding relevant grammar subcategories. All verb POS tags in the training data are annotated with the currently realised subcategorisation frame before training a completely unlexicalised parser on the extended grammar. This strategy ensures that the parser has information about each verb's subcategorisation frame preferences via its tagging preference.

[4] More precisely, *arc-eager* left-corner parsing, which immediately links found and predicted nodes.

We also include a fully lexicalised parser (*Lex*) and an unlexicalised parser (*Unlex*) as an upper and lower bound for performance.

Training data

The standard training data for syntactic parsers consists of sections 2-21 of the Wall Street Journal section of the Penn Treebank (WSJ). We add the data from section 24, which is often used as a held-out development set, to gain as much lexically-specific information as possible.

A second training set, referred to as the *restricted* set below, was extracted from the full WSJ training data to control more carefully the structures covered by the grammar in order to reduce noise present in the large data set and grammar. The restricted training set was created using `tgrep2` queries (Rohde, 2001) to extract simple examples of the structures involved in the MC/RR, NP/S, NP/0 and PP Attachment ambiguities. We extracted main clauses with transitive and intransitive verbs, reduced relatives and sentences with initial adverbial clauses. No complex NPs or additional embedded clauses were allowed in the structures. The resulting training corpus contains only 12,600 sentences, but the frequency relations between the structures in the WSJ corpus are approximately maintained.

The Penn Treebank annotation allows the identification of of passive *by*-phrases (as opposed to, e.g., locative *by*-phrases) through the annotation of a function tag. This is helpful for the processing of the MC/RR ambiguity. Since the parser does not support the use of these function tags, we added a preterminal grammar category to retain the distinction between different types of *by*-phrases. Traces were removed because the parser has no special mechanism to account for them, and sentence-final punctuation was also deleted because the sentences in the restricted corpus are often subclauses and therefore show little evidence of sentence-final punctuation.

Test Data

Two different sets of test data are available. The first is Section 23 of the WSJ data set, the standard test data for syntactic parsers trained on the WSJ data. A second test set is made up of example experimental items for the ambiguities modelled below to ensure that the parser covers their syntactic structures correctly. We used the materials from McRae et al. (1998) (MC/RR ambiguity), Pickering et al. (2000) (NP/S and NP/0 ambiguities) and Rayner et al. (1983) (PP-Attachment). On this test set, we report the number of stimuli that are parsed correctly in all critical regions investigated in the literature, since only such analyses are useful for the SynSem-Integration Model.

	Full WSJ				Restricted WSJ			
Parser	Recall	Precision	F	Cov.	Recall	Precision	F	Cov.
Lex	86.47	86.65	86.49	100%	69.70	68.66	69.18	99.5%
No Bilex	86.17	86.31	86.29	100%	68.69	69.76	69.22	99.5%
Unlex SC	85.47	85.59	85.61	100%	67.73	67.54	67.63	99.5%
Unlex	85.48	86.01	86.15	100%	69.36	68.28	68.82	99.5%

Table 5.1.: Bracketing Recall, Precision, F and Coverage on WSJ Section 23 for different parser instances.

Results: Parsing F Score

We report the parsing F score, bracketing precision and recall over all sentences in section 23 of the WSJ. Recall from Section 4.2.2 that F score is defined as $F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$. For the recall measure, we count a parse as failed if all the lexical nodes are immediate daughters of the top tree node. All parser instances were trained both on the standard WSJ training set and the restricted corpus.

Table 5.1 shows results for all parsers trained on both the full WSJ training set as well as the restricted training corpus. All parsers trained on the full WSJ training set perform reasonably close to standardly used parser implementations (Charniak, 2000, $F = 89.5$), with complete coverage of the test data. The difference in performance between the lexicalised and No Bilex parsers on the full WSJ training set is very small, as expected.

The completely unlexicalised parser somewhat unexpectedly also performs very similarly to the lexicalised parsers, but the introduction of subcategorisation information into the unlexicalised grammar does not appear to increase performance. Instead, precision drops by about half a point F score for the Unlex SC parser in comparison to the Unlex parser.

When training on the restricted corpus, performance is overall much lower due to a lack of training data. However, coverage of the WSJ test corpus is still very high at 99.5%, which underscores the effectiveness of the smoothing strategies used in the parser. Using the restricted training corpus, the difference between lexicalised and unlexicalised parsers is again very small, which is not unexpected given a small training corpus. Under these conditions the No Bilex parser's precision is even higher than the lexicalised parser's, presumably because it uses only the least sparse level of lexicalisation. Extending the categories of the unlexicalised grammar however leads to an abrupt drop in performance, especially in recall. The reason is probably that a distribution over even more variables has to be learnt from extremely sparse data,

	Full WSJ				Restricted WSJ			
Parser	NP/0	NP/S	MC/RR	PP	NP/0	NP/S	MC/RR	PP
Total	50	32	80	24	50	32	80	24
Lex	70.6%	84.8%	56.8%	91.6%	0%	75.8%	51.9%	54.2%
No Bilex	70.6%	84.8%	56.3%	90.5%	0%	75.8%	51.9%	54.2%
Unlex SC	57.1%	78.8%	46.8%	87.5%	70.6%	75.8%	51.9%	62.5%
Unlex	54.9%	84.8%	56.8%	87.5%	0%	75.8%	53.1%	50.0%

Table 5.2.: Percentage of sentences correctly parsed throughout for different parser instances.

while smoothing is not optimised for this training strategy. Also, of course, many verbs in the test set are unseen in the training set, so that the assignment of a wrong fine-grained verb tag may do a lot of damage.

For our experiments, the lexicalised version of the parser without bilexical dependencies seems to be the best choice so far (given that we do not wish to use the fully lexicalised parser, which uses head-head co-occurrence information). It easily outperforms the unlexicalised parser with subcategorisation information, and, when using the restricted training corpus, even the lexicalised parser.

Results: Parsing Experimental Items

The second step of evaluation compares the parsers' performance on one set each of actual experimental items for the MC/RR, NP/0, NP/S and PP Attachment ambiguities. Table 5.2 presents the percentage of correctly parsed sentences for each test set, where a correctly parsed sentence is one with correct incremental analyses for all critical regions. An incremental analysis is correct if it corresponds to one of the (usually two) analyses that are assumed in the literature to give rise to the ambiguity. Again, the parser instances trained on the full and restricted WSJ corpus are compared. The first line in the table contains the total number of sentences parsed.

On the experimental items test set, the first interesting comparison is between parsers trained on the full and restricted corpus. The restricted corpus was constructed with the aim of ensuring noise-free and correct coverage of the structures involved in the ambiguity phenomena of interest. However, as for the WSJ test set, the parsers trained on the restricted corpus are generally at a disadvantage. Coverage is clearly lower than for the parsers trained on the full corpus, most drastically so for the PP and the NP/0 sentences. The striking success of the Unlex SC parser at parsing the NP/0 sentences arises only because it practically always predicts one fine-grained verb POS

tag which allows the assignment of correct structure. Given the size of the training set, it is more likely that this preference is due to sparse data than to subcategorisation preferences evident in the training data. Thus, the restricted training set appears to be too sparse to allow reliable coverage. The parser instances trained on this set frequently encounter unknown words in the sentences, which leads to parsing problems and incorrect assumptions about subcategorisation preferences.

When the full WSJ is used for training, the experimental items are generally parsed accurately, with the MC/RR set proving hardest. The lexicalised and No Bilex parsers are virtually indistinguishable and easily outperform the unlexicalised parsers. The Unlex SC parser again performs slightly worse than the unlexicalised parser.

Thus, on the experimental items just as on the WSJ test set, the No Bilex parser outperformed the Unlex SC parser. The grammar extension strategy probably did not interact well with the smoothing routines of the parser, which are not intended for this kind of manipulation. Also, the small, but clean restricted training corpus has not led to accurate parses of the experimental items as intended, but has rather proven too small for the induction of lexical subcategorisation preferences. The No Bilex parser trained on the full WSJ corpus is therefore used as the syntactic model in our overall architecture.

5.2. The Semantic Model: Extension to Multiple Arguments

We use one of the two most consistently-performing FrameNet-trained instances of the semantic model in the implemented SynSem-Integration model, namely the FN 3 model. It numerically outperforms the other well-performing model, FN 2, on the McRae test set, but in the context of the SynSem-Integration model, no advantage on the McRae data set is gained by using FN 3 rather than FN 2.

However, two extensions to the semantic model are necessary, because thus far, it has been used to make predictions only for isolated verb-argument pairs. In the experimental items to be modelled in Chapter 6 below, however, role assignments to all arguments of a verb must be evaluated. This section describes two extensions to the semantic model that allow it to process multiple arguments for one verb.

After extracting all verbs and their prospective arguments (NPs, sentential complements and PPs) along with their grammatical functions from the parser output, the semantic model finds the optimal sequence of role assignments to the verb-argument-grammatical function triples as described in Section 5.2.1. Section 5.2.2 describes the normalisation strategies that help the semantic model overcome its preference to assign as few roles as possible which is due to the probabilistic formulation of the model and the independence assumption for role assignments.

5.2.1. Dealing with Multiple Role Assignment

In sentence contexts, there are usually several arguments to each verb. In this situation, the semantic model assigns roles to each verb-argument pair independently, and the plausibility of the whole sentence is computed as the product of the plausibility ratings for the individual verb-argument pairs. However, we wish to further control the role assignments made by the model by positing two common-sense restrictions: First of all, we posit the *unique-sense constraint*, which requires all roles assigned by the same verb to be legitimate for the same verb sense. This constraint simply ensures that the same verb meaning is assumed for all role assignments to arguments of the same verb. The second constraint is the *unique-role constraint*, which states that each role can only be assigned to one argument.[5]

Given the unique-role constraint, the optimal role assignment for each argument is influenced by role assignments to additional arguments. For example, when people read a sentence like *The doctor cured by the dentist...*, they will initially assign *the doctor* the agent role of *cure*, but when the *by*-PP is encountered, they reassign this role to *the dentist* and assign *the doctor* the patient role instead. In order to correctly model human processing, the semantic model needs to do the same. Recall, however, that the semantic model makes role assignments only to one argument at a time without considering other arguments for reasons of sparse data, and that we also make an independence assumption between the arguments of a verb in computing the total plausibility of a sentence for the same reason.

Thus, when processing verbs with more than one argument at a time, we need to ensure that role assignments by the same verb assume the same sense, that each role is assigned only once and that the model outputs the optimal sequence of role assignments, that is the one that has the highest possible probability, while still making the assumption that each role assignment to a verb-argument pair is independent from other assignments.

The problem of verb sense consistency can be solved quite simply by determining the optimal role assignments individually for the set of roles licensed by each verb sense and then choosing the sense that allows the most likely set of assignments. To solve the latter two problems, we use a graph-based optimisation strategy to choose the globally optimal role assignment to each verb-argument-grammatical function triple given the unique-role constraint. That is, instead of choosing the role assignment for each triple that is most likely locally, we choose the role assignment that maximises the probability of the whole set of role assignments. This strategy allows us to keep the independence assumptions in role prediction and overall plausibility computation, because the probability associated with each role assignment is still computed independently of context.

[5]This constraint is somewhat relaxed in the annotation of FN and PB semantic roles to account for non-contiguous argument phrases, for example the message argument of a statement verb that may be split and surround the verb.

At the same time, this strategy allows for co-dependency between role assignments as motivated in the example sentence above. Using the global optimisation strategy, the semantic model is able to revise all role assignments to find the optimal assignments for the encountered arguments at each incremental processing step.

Our optimisation problem can be phrased as a *Linear Assignment* problem (LAP), which requires us to find the optimal matching for a bi-partite graph. Figure 5.4 shows an example bi-partite graph with two distinct node sets. A matching is an assignment of links from all nodes on one side of the bi-partite graph to the the nodes on the other side such that each node is linked to exactly one partner (bold-face links in Figure 5.4). Taking the set of all arguments as one set of graph nodes and the set of all roles proposed for any argument as the other, as shown in the figure, this constraint ensures that each role is assigned to exactly one argument. In addition, we can construct weighted edges between all nodes on either graph side, which allows us to add information about the likelihood of a role assignment to be made. If a role was not proposed for an argument, the edge weight will be zero, otherwise, it will have the value of the plausibility corresponding to the role assignment made by the semantic model. The algorithm that computes the optimal matching uses these weights to find the matching that yields the highest product of assignment probabilities.

It is possible in our problem setting that the two sides of the bi-partite graph have a different number of nodes (usually, the number of possible roles proposed by the semantic model is larger than the number of arguments). In this case, the smaller graph side is filled with dummy nodes which allow all role assignments with very low probability (i.e., which have low weights on the outgoing edges). Such a case is shown in Figure 5.4, where the role-assignment model also proposed the role *Affliction* for one of the arguments, which makes the set of proposed roles larger than the set of arguments.

We use the shortest augmenting path algorithm by Jonker and Volgenant (1987)[6] to solve the linear assignment problem and select the role assignments (graph connections) that optimise the overall score and at the same time obey the unique role constraint. Role optimisation is done incrementally and on a verb-by-verb basis. If a sentence contains several verbs, the probabilities of their optimised role assignments are multiplied.

5.2.2. Eliminating the Few Role Bias

When processing the test input incrementally, it is possible that some syntactic analyses allow the assignment of more roles than others. In this case, the semantic model has a strong bias towards assigning the fewest possible roles because longer role sequences

[6] An implementation by the authors is available at `http://www.magiclogic.com/assignment.html`.

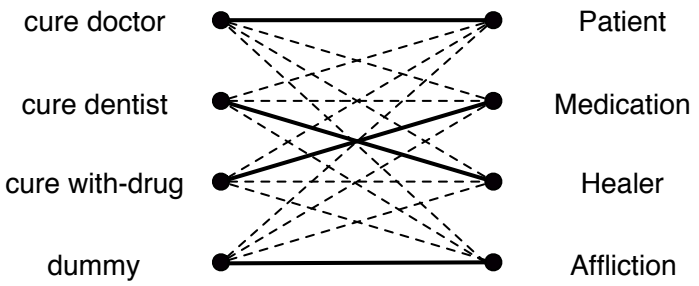


Figure 5.4.: A bipartite argument/role graph.

involve the multiplication of more assignment probabilities and therefore are less probable than short sequences. This behaviour however runs counter to the strong intuition that readers process utterances incrementally and aim to incorporate new material as soon as possible (Pritchett, 1992, Crocker, 1996). We therefore employ two strategies to normalise the probabilities and eliminate this bias.

The first strategy consists of filtering role combinations on the basis of their frequency of appearance with the verb. The prediction for each role combination is weighted by the probability of seeing this combination with the verb. We term this the *preferred role set bias*. It usually places assignments with few roles (e.g., just a role assigned to the the syntactic subject) at a disadvantage, because the training data contains more evidence for larger role sets for most verbs. For unseen combinations, a low, smoothed value is assumed because unseen combinations can either indicate a highly unlikely prediction or very sparse training data. Giving an advantage to analyses that support a verb's preferred role set both places an emphasis on assigning roles to incoming possible arguments until the verb is saturated and at the same time ensures that plausible role sets are preferred by the semantic model. This sets the preferred role set bias off from another possible bias that just puts more weight on role sets the larger they are.

In addition to introducing this bias, we also compute the geometric mean of the role probabilities that make up the overall prediction. This normalisation takes into account the number of roles in each prediction and therefore further reduces the advantage of predictions with few roles.

5.3. Parameter Setting in the SynSem-Integration Model

As outlined in Section 2.5, the SynSem-Integration model predicts processing difficulty as follows: For each syntactic analysis returned by the syntactic model, the model computes an overall probability which is an interpolation of the probabilities assigned

to it by the syntactic and semantic models.[7] As other parser-based models, we restrict the beam of available syntactic analyses to ensure computational tractability and model human memory restrictions. We set the beam dynamically by admitting only analyses with syntactic probability values within two orders of magnitude of the probability of the top-ranked analysis.

The syntactic analyses are then ranked by this overall probability, and the highest-ranked analysis (with its semantic interpretation), the *globally preferred structure*, is predicted to be the one preferred by people. Recall from Section 2.5.2 that the SynSem-Integration model characterises difficulty in processing by two kinds of cost: *Conflict cost*, which predicts difficulty due to conflict between syntactic and semantic preferences, and *revision cost*, which predicts difficulty due to the revision of the assumed analysis of the input. For each region, the amount of predicted difficulty is computed as the sum of predicted conflict and revision cost normalised across the number of stimuli.

In this section, we introduce several alternative ways of computing the cost terms and describe the selection of the best-performing cost functions as well as the setting of the syntax-semantics interpolation factor used to compute the global probabilities for syntactic analyses.

5.3.1. Interpolation Factor

The global plausibility score for the candidate analyses is computed by interpolating the syntactic and semantic scores, as shown in Equation 5.7.

$$Global\ score(s_i) = f Syn(s_i) \cdot (1 - f)\ Sem(s_i) \tag{5.7}$$

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, which is used to determine the globally preferred analysis.

Note that the interpolation of the probability estimates from the syntactic and semantic model yield only a score, not a probability distribution, because the event space of the two models differs. The syntactic model is defined over input strings, the semantic model over verb-argument pairs. This means that the semantic model does not make predictions for all input strings, but only for those containing at least one verb-argument pair according to the syntactic model.

5.3.2. Conflict Cost

Conflict cost is computed by ranking all structures separately in the syntactic and semantic models and then comparing the syntactic and semantic ranks of the globally

[7] The syntactic and semantic probabilities for all analyses are separately normalised to sum to 1.

preferred structure. Cost is incurred if one of the models prefers a different structure than the globally preferred one. Since the global ranking is computed on the basis of the syntactic and semantic rankings, this is equivalent to saying that the syntactic and semantic ranks of the globally preferred structure disagree. The second formulation instantiates the idea that processing takes longer when syntactic and semantic constraints disagree than when they agree, as supported by McRae et al. (1998).

There are several options for computing the amount of conflict cost. Take $rank_{syn}$ and $rank_{sem}$ to denote the syntactic and semantic rank of the globally preferred analysis gp. 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. In the order of fineness of granularity, the cost functions then are

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

  Fixed Cost predicts binary difficulty by assigning a cost of 1 if the rank of the globally preferred analysis differs in the syntactic and semantic models.

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

  Rank cost computes the 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 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. Thus, the cost function captures the strength of the disagreement between the models and thereby allows somewhat more graded predictions than the fixed function.

- *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 is the ratio of the probability assigned to the the highest-ranked structure in a disagreeing model (the locally preferred structure lp) and the probability assigned by that model to the globally preferred structure (gp). In this way, even more gradedness can be achieved than with the rank function, 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 locally 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.

5.3.3. Revision Cost

We also identify three revision cost functions that apply when the semantic interpretation of the globally preferred analysis at the current processing step has changed from the last 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. Note that we do not pay attention to the roles assigned to the verb-argument pairs, because role re-assignment does not incur cost as long as the syntactic structure remains the same (cf. *He loaded the truck*$_{Goal}$, which is easily reanalysed into *He loaded the truck*$_{Theme}$ *onto the boat*$_{Goal}$, upon encountering *onto the boat*, e.g., Pritchett, 1992). $set(gp_t)$ 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.

- *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 again 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.

- *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 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 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 function 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.

- *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 makes the amount of cost assigned by the if-worse function variable by assigning the ratio of the semantic probabilities of the last preferred analysis and the current preferred analysis. Cost is then scaled by the logistic function (see Ratio conflict cost) to avoid an explosion of cost if the current best analysis is much less likely than the last preferred analysis.

5.3.4. Method

The parameters of the SynSem-Integration model are set so that the model predicts an experimentally found pattern of human processing difficulty with maximal accuracy. We use the No Bilex syntactic parser model introduced in Section 5.1.3 as the syntactic model and the FrameNet-trained FN3 model with the extensions outlined in Section 5.2 as the semantic model.

Development Set

We use one of the available sets of experimental results on human processing difficulty as a development set for parameter setting. Since we have most data available for the NP/S ambiguity, we choose a set from this pool that shows significant effects of thematic fit, so that the SynSem-Integration model is optimised to predict a statistically significant difference in difficulty. An additional choice criterion is the number of stimuli that are processed correctly by the syntactic model and that contain verbs covered by the training data for the semantic model.

The data set with the largest number of processable stimuli is the data set corresponding to the equibiased verbs in the Garnsey et al. (1997) reading time study: We have 11 implausible and 12 plausible stimuli available that were parsed correctly by the syntactic model and for which the verbs in the ambiguous region are covered by the FrameNet training data for the semantic model. In this study, processing difficulty is identified by subtracting the reading times for the ambiguous stimuli from those for disambiguated versions. The difference is assumed to indicate processing difficulty caused by the processing of the ambiguity. See Section 6.3.3 for further details on the materials and results of Garnsey et al. (1997).

The development data set consists of four data points, namely measurements for two thematic fit conditions on two sentence regions. The experimental observations and the SynSem-Integration model's predictions are scaled (as proposed in Narayanan and Jurafsky, 2005) to indicate the percentage of difficulty contributed by each region, since our model does not intend to directly predict reading times or reading time differences, but more abstractly the occurrence of difficulty due to processing mechanisms.

Evaluation Metric

We evaluate the different parameter combinations 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. 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 are aware that only four data points are

Conflict Cost	Revision Cost	Good Predictions for f range
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

Table 5.3.: Best-performing interpolation factors for different cost function combinations. Best result in **bold face**. 1: syntax only, 0: semantics only, –: No correct predictions

not a sufficiently large basis for computing a correlation analysis and therefore do not report the significance level for the correlation).

The number of different options for implementing cost functions and the need to simultaneously select the best options and set the weighting parameter for syntax and semantics lead to a relatively large model space. Of the nine possible combinations of conflict and revision cost functions (3×3), we explore only a subset of five. We do not evaluate the combinations of the conflict and revision Ratio cost functions with any of the non-ratio cost functions, because the ratio functions make predictions of vastly different grain size. We evaluate each of the five combinations of cost functions with ten values for the weighting parameter (in 0.1 steps from 0 to 1).

5.3.5. Results and Discussion

Table 5.3 gives an overview over parameter values that allow good qualitative predictions of the pattern of difficulty in the development data. The conflict and revision cost functions are the ones introduced in Sections 5.3.2 and 5.3.3. Table 5.3 also gives the range of values for the interpolation factor f that lead to qualitatively correct predictions. All specified values of f lead to a correlation coefficient of Pearson's $r \geq 0.95$ between the predicted and observed data points. f values in bold face denote parameter settings that lead to especially good predictions (Pearson's $r > 0.99$).

Several observations are interesting:

- Only models using the probabilistic or If-Worse revision cost function make qualitatively correct predictions. This indicates that it is important to assign revision cost only if the new preferred semantic analysis is less plausible than the old one was.

- The probability ratio approach, though appealing through its fine grain size, does not allow us to predict the correct distribution of difficulty as well as the coarser-grained approaches: It predicts a somewhat too even distribution of difficulty to match the observations. This is possibly due to noise, as the development set is still relatively small at a maximum of fourteen stimuli. Note that the development set is however relatively large in comparison to other test sets used for evaluation below, so this problem is expected to persist for other test sets. Despite this disadvantage, the probability ratio cost functions however still allow qualitatively correct predictions.

- The well-performing parameter settings all favour the influence of syntax for the computation of the global ranking, with performance improving with the size of the weighting factor. For all successful model instances, predictions became more like the observed development data the larger the interpolation factor was. However, the range of f for which the non-probabilistic functions qualitatively predict the experimental observations is relatively wide. This shows that the model is quite robust as long as the syntactic model has more weight in deciding the global ranking.

Figure 5.5 shows the model predictions and experimental results for the best-performing parametrisation which uses the *Rank/If-Worse* combination of cost functions and $f = 1$. The SynSem-Integration model captures the general trend of more or less equal difficulty in both measured regions for the good object condition as well as the fact that in the bad object condition, less difficulty is observed at the main verb (MV).

In the evaluation in Section 6 below, we will plot the predictions of this model. To show that the model's predictions are robust across the well-performing model instances, we will also report numerical evaluation results for the other two successful parametrisations, *Fixed/If-Worse* with $f = 1$ and *Ratio/Ratio* with $f = 1$.

5.4. Summary and Discussion

In this chapter, we have introduced the incremental probabilistic parser used as the syntactic model, discussed two extensions that enable the semantic model to process sentences with multiple arguments per verb and described the parameter setting for the SynSem-Integration model.

We compared two strategies of introducing verb subcategorisation information into the probabilistic parser: Limited lexicalisation and the introduction of grammar subcategories. The lexicalised parser clearly outperformed the parser with the extended set of grammar categories both on the standard parsing test set and on the syntactic structures relevant for the evaluation in Chapter 6. It also became clear that the parser profits from large amount of training data. Using the standard parser training set led to

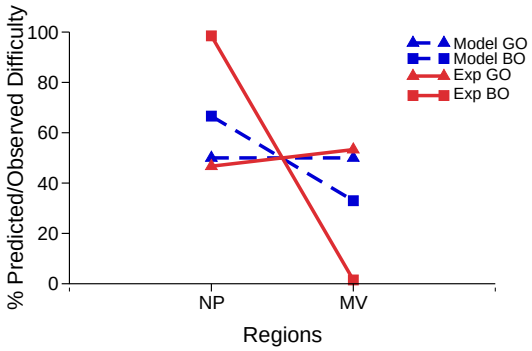


Figure 5.5.: Model predictions and experimental results for the best-performing parametrisation on the development set: Rank/If-Worse 1. Development set: Equibiased verbs from Garnsey et al. (1997). GO: Good Object, BO: Bad Object.

much better model performance than training on a small, but very clean data set that contained only structures relevant to the test phenomena.

The partially lexicalised parser model trained on the full WSJ training set is an incremental, experience-based model of human sentence processing. Parsing results on unseen test data underscore its wide coverage and the correctness of its structural predictions for unseen input.

The semantic model needed two extensions to allow it to process multiple arguments per verb and thus be fully usable within the SynSem-Integration model. We introduced a selection strategy for the best set of role assignments that ensures that two conditions are satisfied: The unique-frame constraint that demands the same verb sense to be used for role assignments to different arguments and the unique-role constraint, which allows each role to be assigned at most once. We also added a normalisation procedure for plausibility predictions to eliminate the bias towards small role sets inherent in the semantic model.

Finally, we proposed different implementations of the conflict and revision cost functions and selected the best-performing combinations. We also selected the optimal values for the interpolation factor f for each combination of conflict and revision cost function. Three cost function combinations allowed predictions that were qualitatively appropriate for the development data: The *Rank/If-Worse* combination, the *Fixed/If-Worse* combination and the *Ratio/Ratio* combination. The former two combinations performed somewhat better than the *Ratio/Ratio* combination, presumably due to

noise in the *Ratio/Ratio* predictions that had a large impact given the limited number of stimuli.

The different instances of the SynSem-Integration model performed quite consistently across different values for f. This indicates that the model is robust against small differences in the parameters. Crucially, however, all three successful model instances assign revision cost only if the semantic plausibility of the globally preferred analysis drops after a revision. Cost functions that always predict difficulty due to revision were not able to correctly predict the pattern of results of the development set.

6. Evaluation of the SynSem-Integration Model

We have selected three instances of the SynSem-Integration model in Chapter 5 that qualitatively predict the pattern of human processing difficulty on the development set. In this chapter, we turn to evaluating the SynSem-Integration model in detail by comparing the model's difficulty predictions to the patterns of human processing difficulty observed experimentally.

We test the model predictions for four sentence processing 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. We compare our model's predictions to the patterns of difficulty observed in reading-time studies, where processing difficulty is taken to cause lengthened reading times for the ambiguous and disambiguating regions. Difficulty is usually quantified by comparing the conditions with the interesting manipulation to unambiguous control conditions. Any difference in difficulty is then attributed to the manipulation. Comparing this difference across conditions or regions gives an impression of the distribution of processing difficulty across the utterance.

Section 6.1 describes the general method followed in the evaluations described in Sections 6.2 to 6.5. In these sections, we start by introducing the experimental findings for each ambiguity and then present the SynSem-Integration model's predictions in comparison to the findings. We carry out a quantitative analysis of all predictions made for each phenomenon, and in Section 6.6, also pool the data from all studies and compute the correlation for all predictions and observations. This is a stringent test for model performance, as we group together the observations from four different phenomena and a total of eight studies. Finally, we discuss the model's properties and evaluation results in Section 6.7.

6.1. Method

We evaluate the predictions of three instances of the SynSem-Integration model using three different combinations of conflict and revision cost functions. These are the *Rank/If-Worse*, *Fixed/If-Worse* and *Ratio/Ratio* combinations. We plot the predictions made using the *Rank/If-Worse* instantiation of the model that uses the rank-difference

conflict cost function and the if-worse semantic cost function, but we report numeric evaluation results also for the *Fixed/If-Worse* and the *Ratio/Ratio* models. We use the FN 3 semantic model from chapter 3 and instantiate the syntactic model with the *No Bilex* parser from Section 5.1.3.

From each reading-time study, we test only those items that can be processed correctly by the syntactic and semantic model to avoid making noisy predictions due to misanalyses. An item counts as being processed correctly by the syntactic model if in all critical regions, the best analysis assigned by this model corresponds to one of the analyses assumed in the literature.

An item can be processed correctly by the semantic model only if the verb in the ambiguous region is present in the FrameNet training data. We accept unknown verbs in the disambiguating regions to ensure that enough stimuli are covered. If a verb is unseen, the semantic model assigns the verb-argument pair a dummy role (since the admissible verb-specific role set is unknown) and predicts a smoothed probability estimate. Using a smoothed value for the role assignment made by an unknown main verb amounts to labelling the main verb-argument relation as not specifically plausible, but acceptable. This is enough for our purposes since the plausibility of main verbs and their arguments is never manipulated in the experimental studies and since there is no comparison of plausibility across stimuli, only within the analyses for one stimulus.

We generally require at least 10 covered stimuli per condition to reduce the likelihood of spurious predictions due to noise as much as possible, but in several cases, we have to relax this requirement slightly.

To show how the introduction of a semantic model affects the predictions and give a performance baseline, we also report the performance of a syntax-only model, namely the fully lexicalised *Lex* model introduced in Section 5.1.3 that makes use of head-head lexical dependencies and thereby is in principle able to use co-occurrence information to reflect the likelihood of the verb-argument pairs in the syntactic parses it constructs. This model predicts difficulty through the "flip" cost function (Crocker and Brants, 2000) that 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.

Different experimental paradigms yield different measures of reading times. For self-paced reading studies, in which the participants reveal new material by button-presses once they have finished reading, the time between button presses is reported. In eye-tracking studies, where reading time is established by tracking the participants' eye movements, a number of different measures is available. We compare the SynSem-Integration model's predictions to the results 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. Recall that the model's difficulty prediction for each critical region is the sum of conflict and semantic cost incurred in the region, normalised across all covered stimuli.

The experimental observations and the predictions of the SynSem-Integration model

as well as the syntax-only baseline 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 neither model intends to directly predict reading times or reading time differences, but more abstractly predicts the occurrence of relative difficulty due to processing mechanisms. We scale by summing the observed or predicted difficulty over all regions for each condition and by normalising 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 the SynSem-Integration model's predictions by comparing the predicted and observed patterns of difficulty quantitatively and qualitatively. For each individual data set, only the qualitative comparison is available, since each study only furnishes up to six data points for comparison. This means that the number of data points is generally too low for a meaningful correlation analysis of predicted and observed data. Therefore, we pool the data from all studies that investigated the same phenomenon and carry out a per-phenomenon analysis of the correlation of the predictions made by the SynSem-Integration model and by the syntax-only baseline to the observed pattern of difficulty.

6.2. Main Clause/Reduced Relative

The Main Clause/Reduced Relative (MC/RR) ambiguity arises for verbs which realise the simple past and past participle by the same form, like *cured* in sentences (6.1) and (6.2).

(6.1) The doctor cured the patient.

(6.2) The doctor cured by the treatment had invented it himself.

For the sentence prefix *the doctor cured*, there are two possible continuations: The main clause continuation as in (6.1), which interprets the verb being in the a simple past, and the reduced relative continuation as in (6.2), which interprets it as a past participle beginning a reduced relative clause that modifies *the doctor*. The two analyses also vary in the thematic roles assigned to the first NP: In the case of a reduced relative, *doctor* is assigned a patient role by *cured*, while in the case of the main clause continuation, it is assigned an agent role.[1] The ambiguity continues until the input is consistent with only one interpretation. This *point of disambiguation* which ends the *ambiguous region* is

[1] Different verbs of course assign different roles, but no verb assigns the same role to the first NP in both readings.

reached at *the patient* in (6.1) and in (6.2) at the latest at *had*, which is the true main verb of the sentence.

Note that the *by*-phrase in sentence (6.2) is often interpreted as already disambiguating the sentence towards the reduced relative reading. However, though a *by*-phrase undeniably introduces a strong bias for the reduced relative clause, its ability to disambiguate depends largely on the semantics of the embedded NP, because *by*-phrases can also be interpreted as specifying locations or manner.

For this ambiguity, there is a general syntactic bias towards the main clause interpretation over the much rarer reduced relative interpretation. Readers therefore experience processing difficulty at the *by*-phrase or the disambiguating main verb of stimuli like sentence 6.2. The experimental logic of most studies involves testing whether the influence of syntactic or semantic factors in the ambiguous region cancels out this processing difficulty or at least makes processing noticeably easier. The regions of interest are usually the ambiguous verb, where some studies already find effects of a plausibility manipulation of the first NP, and the point of disambiguation, which may lie at the completion of a *by*-PP or, at the latest, at the main verb.

6.2.1. Experimental Evidence

In the literature, a number of lexical and syntactic factors which influence the processing of this ambiguity have been identified. They can generally be grouped under the headings of syntactic factors, such as verb form frequency or subcategorisation preferences, and semantic factors, such as the influence of thematic fit or referential effects like NP definiteness and context (e.g., Crain and Steedman, 1985, Spivey and Tanenhaus, 1998). We are concerned with the the former three factors here.

Syntactic Factors

A first factor that influences the processing of the MC/RR ambiguity is lexical verb form preference: Trueswell (1996) showed that the frequency with which the ambiguous verb is used as a past participle influences readers' preference for adopting a reduced relative clause. Verbs with low past participle frequency, which presumably bias the reader towards a main clause interpretation, lead to difficulty at the point of disambiguation, when the main clause reading becomes impossible. Verbs with a high past participle frequency, however, cause as little disruption at the point of disambiguation as an unambiguous control verb. These verbs thus bias the reader towards a reduced relative interpretation early on.

Another factor is the availability of plausible alternative analyses licensed by different subcategorisation frames. MacDonald (1994) demonstrated that the number of acceptable syntactic analyses of the material up to the main verb influences processing difficulty before and at the disambiguating region. When many analyses are possible,

as in *the dictator fought*, which allows interpretation as a main clause with or without a direct object in addition to the reduced relative analysis, processing was difficult. In cases where fewer analyses were possible, readers experienced less difficulty at the point of disambiguation.

MacDonald also demonstrated the influence of post-verbal material in the ambiguous region. Material like a *by*-phrase immediately and strongly points towards a reduced relative analysis, but the presence of an adverbial phrase can delay constraining information, causing readers to entertain a strong main clause hypothesis for longer. This increases processing difficulty at the disambiguation.

Thematic Fit

Even more than syntactic influences, the influence of the thematic fit of the first NP with the verb has been intensively studied. The first study to ask whether thematic fit influences processing in the MC/RR ambiguity, Rayner et al. (1983), compared the processing of reduced relative sentences with good and bad agents for the verb to the processing of two unreduced control structures. Using eye-tracking, they found that both types of reduced relatives were harder than the unambiguous control sentences, and that the use of the implausible subject NP in one of the control structures did not cause difficulty, either. This evidence led them to conclude that thematic fit does not influence processing, at least not in its the early stages.

After these results, research was continued using animacy manipulation, a somewhat stronger variant of thematic fit manipulation. Ferreira and Clifton (1986), in another eye-tracking study, contrasted unreduced controls and reduced relative clauses with animate and inanimate first NPs. They found again that animacy information of the first NP does not suffice to cancel out the difficulty at the *by* region for reduced relatives in comparison to the unreduced relatives. However, they also found that readers had difficulty at the verb of a reduced sentence after reading an inanimate first NP which made the main clause interpretation less likely. Thus, even though readers did not seem to profit from the animacy information at the point of disambiguation, they obviously did react to it.

Trueswell et al. (1994) identified a number of problems with the items and presentation method employed in the Ferreira and Clifton study. Most importantly, the animacy manipulation in their materials in about 50% of cases did not exclude a plausible main clause continuation of the sentence. Trueswell et al. therefore manipulated animacy while ensuring that main clause continuations were excluded. Also using eye-tracking, they found a strong effect of animacy that practically eliminated difficulty at the *by*-phrase in items with inanimate first NPs. Re-processing effects (measured by second pass reading times) also showed numerically longer processing times at the verb for inanimate first NPs, similar to the effects observed in Ferreira and Clifton (1986). These results suggest that thematic fit information does play an early and important role

during the processing of the ambiguous region.

Clifton, Traxler, Mohamed, Williams, Morris, and Rayner (2003), in an eye-tracking replication of the Trueswell et al. experiment, did not find the same extreme effect of thematic fit. Instead, in their experiment, readers did show difficulty for reduced relatives with inanimate first NPs in comparison to the unreduced control. Still, reading times at the *by*-phrase for inanimate NPs were numerically, but not significantly, lower than for animate NPs. Additionally, using an eye-tracking measure that is sensitive to difficulty in initial processing and to difficulty in recovery from misanalysis, they found no difference between animate and inanimate first NPs in the *by*-phrase. Instead, significantly longer durations for animate first NPs on the main verb were found in one experiment and an interaction of animacy and ambiguity in the second experiment. Clifton et al. conclude that animacy information does not help readers to avoid processing difficulty, but that it does influence the time needed to recover from misanalysis.

In contrast to the Clifton et al. study, and in keeping with the Trueswell et al. study, McRae et al. (1998) found effects of thematic fit before the disambiguating main verb. They held animacy constant and varied general thematic fit, controlling the plausibility properties of their combinations of first NP nouns and verbs in a norming study. The McRae et al. self-pace reading study found an interaction of reduction and thematic fit in the verb+by region and at the disambiguating main verb, such that sentences with good agents were easier to read than good patient sentences at the verb+by region and harder at the disambiguating verb. Thematic fit did not eliminate difficulty at the verb+by region, but sentences with good patients were read as quickly as the unreduced controls at the disambiguating main verb. This clearly points towards an immediate use of thematic fit in processing that helps determine the assumed sentence structure.

This result is corroborated by a study by Tabossi, Spivey-Knowlton, McRae, and Tanenhaus (1994), who tested the influence of graded thematic fit on the processing of the MC/RR ambiguity. Using the same method and regions as McRae et al., they chose items with varying thematic fit within the larger classes of good agent and good patient items. Consequently, they did not find reliable interactions of reduction and thematic fit in reading times, but they were able to show that agenthood/patienthood ratings correlated with the size of the reduction effect in all three regions (marginally so at the verb+by). The difference between agenthood and patienthood ratings also predicted the reduction effect on the agent NP and at the main verb. This study indicates that relatively subtle changes in the thematic fit variable may have a graded influence on processing. The marginal prediction of the reduction effect by agenthood ratings at the verb again suggests that thematic fit is used rapidly.

Summary

A number of factors for processing the MC/RR ambiguity has been identified in the literature: Verb form frequency, subcategorisation preference and post-verbal context form a group of syntactic constraints. These factors can be modelled through lexical and grammatical preferences encoded in the probabilistic grammar of the syntactic model. Thematic fit is unanimously found to have some influence on processing, although the literature reviewed above does not agree on whether it is used to guide the construction of the initial analysis, as implied by the results of Trueswell et al. (1994), Tabossi et al. (1994), McRae et al. (1998) and also MacDonald (1994) or not, as argued in Rayner et al. (1983), Ferreira and Clifton (1986), and, more recently, Clifton et al. (2003). The influence of this factor is accounted for by the semantic model.

6.2.2. Data Selection and Evaluation Method

For each of the studies reviewed above, we determined (a) how many experimental stimuli were parsed correctly by our syntactic model and (b) how many of the correctly parsed stimuli were covered by FrameNet, such that the verb in the stimulus was seen in the training corpus. We required at least 10 stimuli per condition that were covered by the syntactic and semantic model. Two studies (nearly) met this requirement: MacDonald (1994) and McRae et al. (1998).

Note that we do not model the results of Tabossi et al. (1994), because they investigated the influence of a fine-grained thematic fit manipulation on reading times, while two of the three best-performing instances of our model make only relatively coarse-grained predictions for individual stimuli that cannot capture Tabossi et al.'s effects, and the third has been shown to suffer from noise for small test sets and therefore also for individual stimuli.

6.2.3. McRae et al. (1998)

Experimental Setup

We used the plausibility ratings for the verb-argument-role triples from McRae et al. (1998) in Chapter 3. Now, we are interested in the effects the plausibility manipulation has on reading times. The self-paced reading study presented two words at a time and measured reading times at three regions: At the verb and *by*, at the agent NP in the *by*-phrase and at the disambiguating main verb. This mode of presentation was chosen since Trueswell et al. (1994) found that short function words such as *by* are generally not fixated during reading, but processed *parafoveally*, that is, when fixating the preceding verb. Thus, during natural reading, the *by* may influence the analysis process already when reading the verb. Indeed, the presentation of *by* with the verb or on its own seems to modulate the appearance of a thematic fit effect in self-paced

reading (Burgess, 1991). Clifton et al. (2003) found no such effect in eye-tracking when varying the availability of parafoveal preview of *by*, but it is possible that the effect is strengthened in self-paced reading by the artificial segmentation of the input.

The study found an interaction of ambiguity and thematic fit at the verb+by and at the disambiguating main verb, such that ambiguous sentences with good agents were easier to read than ambiguous good patient sentences at the verb+by region and harder at the disambiguating main verb. At the agent NP, there was a main effect of thematic fit and one of ambiguity, but no interaction between thematic fit and ambiguity.

Materials

The materials consist of 80 stimuli, formed by 40 verbs paired with two first NPs each as shown in sentences (6.3) and (6.4).

(6.3) The patient cured by the treatment had been diagnosed as terminal.

(6.4) The doctor cured by the treatment had invented it himself.

One of the NPs is a good patient (but bad agent) of the verb as in (6.3), and one is the inverse, a good agent (but bad patient) as in (6.4). The plausibility of verb-argument pairs was established in a norming study, the results of which we have been using in Chapter 3. Unambiguous controls were created for the stimuli by inserting *that was* before the ambiguous verb.

24 of the 40 stimuli with a good agent first NP and 21 of those with a good patient first NP were parsed correctly. Of these, 17 good agent stimuli and 14 good patient stimuli were covered by the FrameNet training data and were used as the basis for predictions. The predictions of the lexicalised syntactic baseline model were based on 27 correctly parsed good agent stimuli and 24 correctly parsed good patient stimuli.

For this study, the average reading times per stimulus were kindly made available by Ken McRae. We can therefore use the average reading times for the covered stimuli instead of the average reading times for all stimuli to compute reading time differences. Both the full and the covered data set show a very similar pattern of results.

Results and Discussion

For this data set, we made predictions for the verb and for *by* separately, since both words contain cues for the processing system. The model (dashed blue lines in Figure 6.1) predicts that stimuli with good patients should be harder to read at the verb than stimuli with good agents, where there is a conflict between the syntactic preference for the main clause reading and the semantic preference for the reduced relative. At *by*, both conditions are equally difficult, but from the agent NP on, our model predicts more difficulty for the good agent sentences than for the good patients. This reflects the

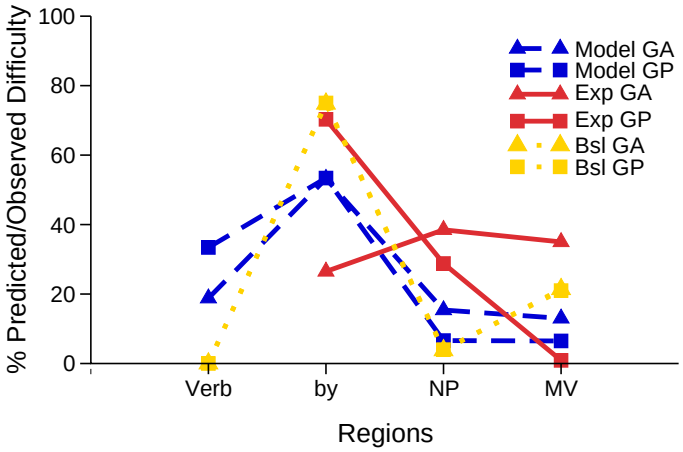


Figure 6.1.: McRae et al. (1998): Experimental results and model predictions for the MC/RR ambiguity. GA: Good agent first NP, GP: Good patient first NP

revision of the previously well-supported main clause readings as the disambiguating region unfolds.

We find these predictions mirrored in the experimental results (solid red lines), but one region late. First of all, 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 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), as predicted by our model. 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 does not allow to exactly time the onset of the difficulty with good agents, and second a *spillover effect*, a phenomenon frequently found with self-paced reading data, where effects show up a region or two after their hypothesised onset.

The syntactic baseline (dotted yellow lines) in contrast makes exactly the same predictions for both plausibility conditions. It 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 analysed as containing a reduced relative at *by*, and the remainder switches the preferred analysis towards a reduced relative at the disambiguating main verb.

In sum, our model correctly predicts the pattern of experimental results. The one-region lag between the model predictions and the observations can be explained by the choice of the regions and the use of self-paced reading as experimental method. The syntactic baseline fails to account for the influence of the thematic fit manipulation and therefore does not account for the experimental findings.

6.2.4. MacDonald (1994)

Experimental setup

MacDonald (1994), in her Experiment 2, varies both thematic fit of the first NP and the amount of information the post-verbal context in the ambiguous region yields with regard to the correct analysis. This makes the study very interesting for us, because it allows us to test whether the SynSem-Integration model correctly predicts the interplay of syntactic and semantic constraints. Sentences (6.5) to (6.8) show a complete item with all manipulations.

(6.5) The news stated that the microfilm concealed inside the secret passageway was discovered. Good/Good

(6.6) The news stated that the microfilm concealed most of the night was discovered. Good/Poor

(6.7) The news stated that the spy concealed inside the secret passageway was discovered. Poor/Good

(6.8) The news stated that the spy concealed most of the night was discovered. Poor/Poor

Thematic fit was varied by manipulating animacy: In a good thematic condition, an inanimate first noun pointed towards a reduced relative continuation as in (6.5) and (6.6), and in a poor thematic condition, an animate first noun pointed towards a main clause continuation as in (6.5) and (6.6).

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. Good materials as in (6.5) and (6.7) made this obvious at the first word. Poor materials as in (6.6) and (6.8) 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 a strong main clause hypothesis.

The experiment used self-paced reading, presenting the first NP, then the verb together with the post-verbal material and finally a two-word region starting with the disambiguating main verb and followed by the rest of the sentence. MacDonald found that a combination of good first NP and good post-verbal material (pointing towards the reduced relative) eliminated the difficulty at the disambiguating main verb. When the two information sources pointed into different directions, she found difficulty effects at the disambiguation that were significant only in one of the subject or items analyses. When both information sources pointed towards a main clause, readers had significant difficulty at the disambiguating main verb.

Materials

The stimuli consisted of 128 reduced relative sentences. Each item was made up of four versions of each sentence combining good and poor first NPs and post-verbal information, as in sentences (6.5) to (6.8).

Sentence (6.5) shows both good NP and post-verbal information, both of which points towards the ultimately correct reduced relative reading. In contrast, both types of information point towards the main clause interpretation and thus are poor indications of the ultimately correct interpretation in sentence (6.8). The other two stimuli combine one good and one poor indication of the reduced relative interpretation. Unambiguous controls were created for the items by inserting *that was* before the ambiguous verb.

Out of the four conditions, the two conditions with poor post-verbal material caused the syntactic model most problems. It appears that the post-verbal adverbial phrases used in the materials not only give information about the possible analyses late, but that they are also relatively infrequent and therefore poorly attested in the training data.

After we replaced most of the four-word phrases with two or three constructions that were parsed correctly, the parser correctly processed 17 stimuli in the Poor NP/Poor post-verb condition and 18 in the Good NP/Good post-verb condition. In the Poor post-verb conditions, 21 stimuli each with poor and good NPs were parsed correctly. Sparseness in the FrameNet training corpus excluded all but 9 stimuli in the Poor NP/Good post-verb condition and 10 in the Good NP/Good post-verb condition. 7 stimuli are covered in the Poor NP/Poor post-verb condition and 8 in the Good NP/Poor post-verb condition. This means that only one condition meets our requirements for the minimum number of stimuli, and generally, there are very few stimuli in each of the conditions that we can base predictions on. We will nonetheless present predictions for all four conditions, with the caveat that the Poor post-verb conditions are rather underrepresented.

The baseline predictions are based on 21 stimuli each in the Good post-verb conditions and 15 stimuli each in the Poor post-verb conditions.

Method

In the reading-time study, measurements were taken at the end of the ambiguous region (i.e., the last word of the post-verbal material), and at the two-word region starting with the disambiguating main verb. We sampled the model predictions at the same regions as for the McRae data, but summed the difficulty predictions from the first verb up to the last word of the post-verbal material to capture the individual predictions for all interesting sub-regions of the long ambiguous region. Below, the conditions are named according to the goodness of the first NP followed by the goodness of the post-verbal material.

Results and Discussion

Figure 6.2 shows the model predictions and the experimental results for the MacDonald study. The model (dashed blue) predicts that the Poor NP/Poor post-verb (P-P) condition where both the first NP and the post-verbal material point towards a main clause to be easiest during the ambiguous region, but hardest at disambiguation. Inversely, the G-G condition, where both information sources point towards a reduced relative, is hard in the ambiguous region and easy at disambiguation. The P-G condition is in the middle between the other two, but harder to process during the ambiguous region than at disambiguation.

All these predictions are correct: During the ambiguous region, processing is harder the more strongly the thematic fit information and post-verbal material point towards a reduced relative construction. At the disambiguation, readers have a more difficult time to accommodate the reduced relative reading the more thematic fit and post-verbal constraints pointed towards or at least allowed a main clause reading.

However, the model makes an incorrect prediction for the G-P condition, which it predicts to be hardest during the ambiguous region and easiest at disambiguation. Recall that the G-P condition yielded only 8 stimuli for analysis. For this relatively limited amount of stimuli, no difficulty at all was predicted at disambiguation, which leads to the extreme scaled predictions. While the predictions for the P-G condition are almost perfect despite the low number of available stimuli, the result for the G-P condition justifies our general requirement for a minimum number of stimuli by demonstrating that predictions made on the basis of few stimuli can be unreliable.

The predictions of the syntactic baseline are plotted separately for the Good NP and Poor NP conditions for clarity. Figure 6.3 shows that the baseline model predicts that the G-P condition should be harder to process during the ambiguous region than the G-G condition, which does not correspond to the observed pattern. For the Poor NP conditions plotted in Figure 6.4, the baseline model predicts that the P-P condition should be harder than the P-G condition in the ambiguous region and easier at the disambiguation, which is also not correct.

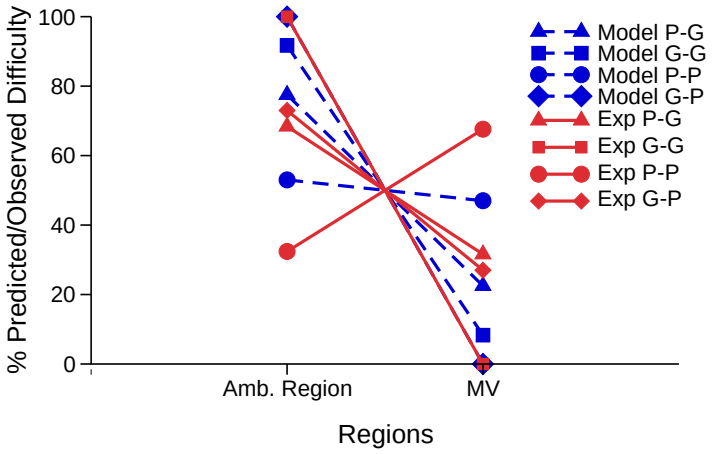


Figure 6.2.: MacDonald (1994): Experimental results and model predictions for the MC/RR ambiguity, all conditions. Thematic Fit–Postverbal Material: G: Pointing towards RR, P: Pointing towards MC

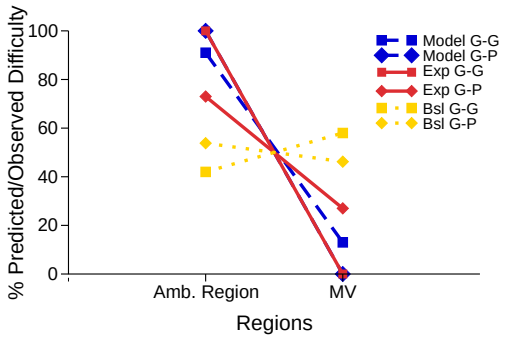


Figure 6.3.: MacDonald (1994): Experimental results and model predictions for the Good NP conditions.

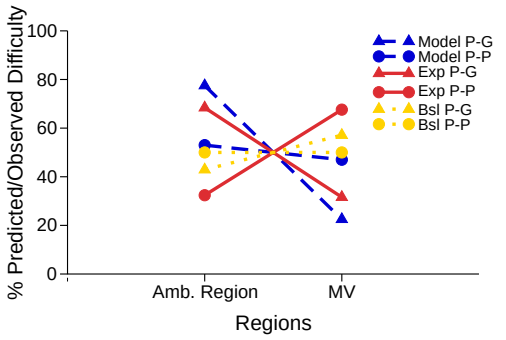


Figure 6.4.: MacDonald (1994): Experimental results and model predictions for the Poor NP conditions.

The syntactic baseline generally predicts that more changes to the preferred parse are made in the Poor post-verb conditions during the ambiguity and in the Good post-verb conditions at the disambiguation. This contradicts the intended effect of the post-verb manipulation, where good post-verb constraints are supposed to disambiguate towards a reduced relative reading during the ambiguous region, leaving no reason to change the preferred parse at the disambiguation, while poor post-verb constraints should postpone the disambiguation for as long as possible. The reason for this apparent contradiction is that many of the Poor post-verbal constraints allow two switches in preferred structure that both correspond to a main-clause reading, thus not excluding the main-clause reading but incurring difficulty by the "flip" measure. The prepositions used in the post-verbal materials can first be interpreted as a quantifier for an object NP (as in *about three kilos*) and then as beginning an adverbial preposition modifying an intransitive verb (as in *about a week ago*) before the disambiguation towards a reduce relative takes place at the main verb. Thus, two changes of preferred analysis are predicted during the ambiguous region versus just one at the disambiguation. For the good post-verbal conditions, only one switch takes place for some stimuli during the ambiguous region and possibly another one at the disambiguation.

In sum, the SynSem-Integration Model again makes correct predictions of difficulty observations. Incorrect predictions for one condition were linked to a sparseness of available stimuli. The syntactic baseline model predicts small differences between conditions based on the manipulation of post-verbal material, but the predicted pattern of difficulty points in the wrong direction, reflecting parser difficulty with the poor post-verbal stimuli.

6.2.5. Numerical Evaluation

In addition to the qualitative evaluation of the model's predictions of the McRae et al. and MacDonald results, we also evaluate quantitatively. Since each study on its own yields rather few data points for a meaningful correlation analysis, we pool the data for the two MC/RR studies. We correlate the scaled model predictions to the scaled experimental findings.

For the correlation analysis, we sum the model's predictions for the verb and *by* regions in the McRae et al. study to ensure that the number of predictions and observations corresponds and re-scale. We use Pearson's r to test for correlation, since the data are normally distributed according to a one-sample Kolmogorov-Smirnov test. The model predictions are significantly correlated to the experimental results with Pearson's $r = 0.792, p < 0.001$. The *Fixed/If-Worse* model performs similarly at $r = 0.818, p < 0.001$, and the *Ratio/Ratio* model does somewhat worse at $r = 0.641, p < 0.02$, but still reaches a robustly significant correlation. For comparison, the syntactic baseline does not achieve a significant correlation to the observations ($r = 0.199, ns$).

These results corroborate the qualitative analysis and show that our model, unlike the syntactic baseline, correctly and reliably predicts difficulty in processing the MC/RR ambiguity.

6.3. NP object/Sentence Complement

The NP/S ambiguity results from the possible interpretation of an NP as a direct object or the subject of an embedded sentence complement. In the example sentences (6.9) and (6.10), the ambiguity occurs at *The criminal confessed his sins*, where two continuations are possible.

(6.9) The criminal confessed his sins and reformed.

(6.10) The criminal confessed his sins harmed too many people.

In sentence (6.9), *his sins* is as a direct object in a main clause, but in the sentence complement reading shown in (6.10), the NP is not an immediate argument anymore, but becomes part of the embedded sentence, which as a whole is a complement of the verb. Accordingly, in the main clause reading, the NP receives a role directly from the first verb, but in the embedded sentence reading, it only forms part of a larger role-bearing argument of the matrix verb and directly receives a role only from the embedded verb, *harmed*.

The ambiguous region in this construction comprises the second NP (*his sins*), with the disambiguation towards the sentence complement reading following immediately at the next word. In this ambiguity, readers usually interpret the second NP as the direct object of the main verb and show difficulty at a disambiguation towards the sentential complement interpretation. The experimental logic is therefore to see whether information like verb subcategorisation or thematic fit can eliminate this difficulty partially or completely. Comparison is usually to versions of the experimental items where the complementiser *that* before the NP eliminates the ambiguity.

6.3.1. Experimental Evidence

The most important factors that have been found to influence the processing of the NP/S ambiguity are verb subcategorisation and the thematic fit of the NP as a direct object.

Verb Subcategorisation

The influence of verb subcategorisation preference, a bias of the verb for taking an NP or sentential complement (SC) argument is controversial. Among the studies that argue

against an effect of verb subcategorisation is Pickering et al. (2000). This eye-tracking study held verb bias constant at a strong preference for SC arguments. However, the results still showed a reaction to the plausibility of the ambiguous NP as a direct object: In eye-tracking measures that reflect later stages of processing, implausible object NPs in ambiguous stimuli were harder to read than plausible ones. Also, ambiguous stimuli with plausible object NPs caused difficulty at the main verb in one of five measures. This seems to indicate that verb preference does not rule out the construction of a direct object reading even for verbs that are strongly SC biased.

Ferreira and Henderson (1990) also found no effect of verb bias using eye tracking. Rather, readers adopted the NP object analysis regardless of verb type and experienced difficulty at the disambiguation for ambiguous stimuli. In a replication using self-paced reading, there was however an interaction between ambiguity and verb type in the total time measure for the disambiguation, such that ambiguous stimuli with SC-biased verbs were read faster than ambiguous stimuli with NP-biased verbs.

These somewhat conflicting findings may be explained by the observation in Hare et al. (2004) that for the items used in Ferreira and Henderson (1990), the assumed subcategorisation preferences of the verbs differ substantially from corpus counts that differentiate by verb sense. This means that their results may have been compromised by inaccurate estimates of verb preferences. However, a recount of verb preferences on the basis of the Hare et al. data did not reveal such a tendency for the Pickering et al. study, so this observation cannot explain the apparent momentary formation of a direct object analysis even for strongly SC-biased verbs.

An early piece of evidence *for* the influence of verb bias comes from an self-paced reading study reported in Holmes, Stowe, and Cupples (1989) which varied verb bias and thematic fit of the NP. In this study, readers clearly experienced difficulty at the disambiguation region for ambiguous stimuli with NP-biased verbs.

Trueswell et al. (1993) point out that the verb bias used by Ferreira and Henderson (1990) was relatively weak and many of the NPs were implausible objects for the NP-biased verbs. Also, they identify an alternative explanation for the longer reading times in ambiguous SC-biased stimuli: Since many SC-biased verbs have a strong preference for being followed by *that*, readers may be surprised or misled by the absence of the complementiser. Since the ambiguous NPs in the Ferreira and Henderson study were very short (one word), such a complementiser effect may well show up only in the next region, the disambiguation, obscuring any differences between SC- and NP-biased verbs due to misanalysis. No analysis of complementiser preference has been carried out for the Pickering et al. study, so the possibility exists that their effects were caused or enhanced by the absence of a complementiser when it was strongly expected, which may have cued readers to prefer the direct object analysis.

Trueswell et al. (1993) ran a self-paced reading with more strongly controlled verb biases and ensuring that NPs were plausible objects. This study showed that participants were sensitive to verb information in that they only showed processing difficulty at the

disambiguation after an NP-biased verb without a complementiser. This result was also replicated in an eye-tracking study. There was also a disruption effect on the NP for SC-biased verbs without a complementiser, the strength of which was correlated to the preference with which the verbs took a complementiser. Consequently, this effect (and also similar small effects in the eye tracking study) was attributed to the lack of an expected complementiser. First fixation durations however revealed the same pattern that was found with this measure in Ferreira and Henderson: Reading times were longer in the absence of a complementiser regardless of verb type.

Garnsey et al. (1997) note that in the Trueswell et al. (1993) items, the ambiguous NPs now were much more plausible objects for the NP-biased verbs than for the SC-biased verbs, which might have inflated the observed influence of verb bias. However, they generally replicated the Trueswell et al. (1993) results in an eye-tracking experiment that varied both verb bias and the plausibility of the NP as a direct object. There was a disruption effect at the disambiguating region for NP biased verbs only, with only hints at an influence of plausibility. For SC-biased verbs, neither ambiguity nor plausibility had an effect at the disambiguation. This pattern was reliable also in first fixation durations, which it had not been for Trueswell et al. However, the effect of complementiser preference found by Trueswell et al. wasn't replicated.

The effects at the disambiguation were replicated in a self-paced reading study. In that study, readers however were slower reading a plausible NP after EQ-biased verbs instead of faster, as expected and as shown in the eye tracking experiment. Garnsey et al. attribute this to the large number of sentential complement analyses in the experiment and the experimental paradigm used.

Pickering et al. (2000), who found that even for SC-biased verbs, an object interpretation seemed to be initially constructed, as noted above, argue that Garnsey et al. found no evidence for this for SC-biased verbs due to a lack of power and a short ambiguous region, while Pickering et al. found the clearest results on the words after the NP. Further, they note that the evidence even for plausibility effects with NP-preferring verbs is relatively weak (see below for more details). This implies that a similar effect for SC-preferring verbs may not have been detected by the experiment.

In sum, there is no consensus in the literature about the role of subcategorisation information. Holmes et al. (1989) and Pickering et al. (2000) find no influence of subcategorisation preference on readers' preferences to construct a direct object interpretation of the ambiguous NP, while Ferreira and Henderson (1990) find a facilitating effect of subcategorisation information on the disambiguation, and Trueswell et al. (1994) as well as Garnsey et al. (1997) find that readers do not construct a direct object interpretation for SC-biased verbs.

Thematic Fit

A number of the studies introduced above also varied thematic fit. The results clearly establish influence of thematic fit on processing the NP/S ambiguity. The only exception is Holmes et al. (1989), where no effect of plausibility was found, except on the ambiguous NP for NP-biased verbs only using the grammaticality judgement task. This task has however been criticised for changing readers' processing strategies and reactions, which is probably the reason why this effect did not show up in self-paced reading without the grammaticality judgement task.

Garnsey et al. (1997) found no effect of plausibility on SC-biased verbs, but verbs that were equibiased between an NP and SC preference were processed faster at the disambiguation when they had implausible objects that seemed to bias readers towards the SC interpretation. Also, at the NP of ambiguous items, there was processing difficulty for NP-biased verbs when the NP was implausible as a direct object, and total times showed some indication for difficulty at the disambiguation after reading plausible NPs.

Pickering et al. (2000) found effects of plausibility both at the ambiguous NP and at the disambiguation, so that implausible NPs in ambiguous items were hard to read and ambiguous stimuli with plausible object NPs caused difficulty at the verb, even though all verbs were SC-biased.

Finally, Pickering and Traxler (1998), in an eye-tracking experiment that only manipulated thematic fit, also found a clear influence of thematic fit in that readers found it harder to process an implausible object NP in ambiguous stimuli, and showed disruption at the embedded verb for ambiguous stimuli with plausible object NPs. If the NP was implausible, difficulty was found in only one measure at the disambiguation (total time). Thus, implausible object NPs did not completely eliminate processing difficulty at the disambiguation, but they greatly reduced it. While verb bias was not manipulated, it can be assumed that verbs were biased towards taking an NP argument, as this is the largest subgroup of verbs that can take both NP and sentence complement arguments.

Other Factors

In the studies cited above, some additional factors were found to be of importance: Holmes et al. (1989), in a third experiment, found an interaction of NP length with ambiguity, such that long NPs were harder to process at the disambiguation of reduced items containing SC-biased verbs. It is conceivable that there is a syntactic preference for subjects of sentential complements to be short, so that long NPs signal a direct object.

Summary

In sum, we have discussed several factors that influence the processing of the NP/S ambiguity: Verb subcategorisation, although disputed in the literature, thematic fit, and, as a side note, the length of the ambiguous NP. The SynSem-Integration model accounts for thematic fit by its semantic model and can model effects of NP length via preference information coded in the grammar of the syntactic model. The influence of verb subcategorisation information is accounted for by lexical preferences in the probabilistic grammar, while the SynSem-Integration model shows a bias for the direct object interpretation, because this analysis involves fewer grammar rule applications and is therefore more likely than the sentential complement analysis. This *small tree bias* is inherent in all PCFG-based parsing models. Following Crocker and Brants (2000), we propose to interpret this bias as implementing a preference for simple structures. The smaller direct object analyses have the additional advantage of being immediately semantically interpretable. When the parser adopts the sentential complement reading, the plausibility of a role assignment to the sentential complement has to be delayed until the sentential complement verb is encountered.

Due to the small tree bias, our syntactic model is prone to initially prefer a direct object analysis, as argued for by Holmes et al. (1989), Ferreira and Henderson (1990) and Pickering et al. (2000). On the other hand, strong lexical preferences can in principle override this bias and result in an immediate preference for the sentential complement analysis, as found by Trueswell et al. (1994) and Garnsey et al. (1997). Therefore, it is conceivable that our model can account for both opposing views, depending on the properties of the stimuli.

6.3.2. Data Selection and Evaluation Method

We model the experimental results from Garnsey et al. (1997) and from Pickering and Traxler (1998), because only for these two studies, the required amount of at least 10 stimuli per condition was covered by both the syntactic and semantic model. Note that for one of the three verb conditions in the Garnsey et al. study, the condition with DO-biased verbs, only seven stimuli were covered per plausibility condition. We present predictions for this condition alongside predictions for the SC-bias condition for the sake of completeness. Recall that the equibiased condition in this study was used as a development set for parameter setting (Section 5.3).

The required amount of stimuli was also covered for the Holmes et al. (1989) study, but since no effect of thematic fit was found there, we do not make a formal comparison of the model's predictions and their results. Our model does however predict the null effect of thematic fit as well as the effect of verb bias.

We make predictions for the critical noun and verb regions in the modelled experiments, ignoring post-nominal and post-verbal regions. These regions are used to

capture evidence of processing difficulty that is delayed for physiological or experimental design reasons, while our model predicts difficulty without delay. Furthermore, the syntactic model often has difficulty to correctly process post-nominal or post-verbal material, and the amount of covered stimuli can be increased by excluding these regions. Our predictions are for the last word in each of the critical regions. If no significant results were found in the critical verb and noun regions, we compare our predictions to the findings for the post-verbal or post-nominal region.

During processing, the semantic model has to decide whether the NP following the first verb is a direct object or the subject of a sentential complement. In the direct object case, the NP directly receives a role from the seen verb, but in the sentential complement case, it indirectly receives the role assigned to the sentential complement. We therefore compare the probability of the NP directly receiving an object role to the probability of assigning a sentential complement role with an as-yet unseen verbal head. This comparison allows us to take into account the verb's preferences for taking direct objects versus sentential complements. Since the head of the sentential complement is unseen, the probability prediction for this role assignment depends on smoothing. Recall, however, that smoothed plausibility predictions are the rule rather than the exception, since few role fillers are actually encountered with the verb in the training data.

6.3.3. Garnsey et al. 1997

Experimental Setup

Garnsey et al. (1997) varied the plausibility of the ambiguous NP and verb bias on three levels: 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). Preferences were established by a norming study.

An eye tracking study and a self-paced reading study were conducted. We model the total times measured in the eye-tracking experiment, using the reported length-corrected residual reading times. For this measure, Garnsey et al. found no effect of plausibility on SC-biased verbs, but verbs that were equibiased between an NP and SC preference were processed faster at the disambiguation when they had implausible objects. For DO verbs, there was difficulty at the disambiguation for ambiguous stimuli with plausible NPs (interaction of ambiguity and plausibility significant by participants). For all three conditions, processing was slowed at the NP for ambiguous stimuli, regardless of plausibility.

The EQ condition is the development set on which we optimised syntactic and semantic cost computation and the interpolation factor between syntax and semantics. We will therefore not present the model's predictions for this condition.

Materials

The materials consisted of 32 stimuli per verb condition, formed by 16 sentences with two versions of the ambiguous NP each. Verb biases and the plausibility of the NPs as objects and as subjects of the sentence complement were established through norming studies. Sentence (6.11) shows an example stimulus with a plausible ambiguous NP and sentence (6.12) one with an implausible ambiguous NP.

(6.11) The editor printed the article had been slanderous to him.

(6.12) The editor printed the media had been slanderous to him.

The example stimuli are for a DO-preferring verb. Verbs were classified as DO or SC if they were completed twice as often with the respective argument than with the other. EQ biased verbs occurred approximately the same number of times with both continuations.[2] The plausibility ratings for plausible and implausible object NPs differed at least by 2.5 points on a 7 point scale. The plausibility ratings for the NPs as subjects of embedded clauses also made a plausibility confound unlikely. Unambiguous controls were created for the items by inserting the complementiser *that* after the first verb.

Of the 32 stimuli in the DO condition, eight were correctly parsed and seven also covered by FrameNet for each plausibility condition. As mentioned above, we will still present the predictions for this condition for the sake of completeness. Of the 32 SC stimuli, 13 were correctly parsed for both conditions, and 12 also covered by FrameNet. The predictions of the syntactic baseline are based on 16 stimuli each in the SC conditions and 13 stimuli each in the DO conditions.

Results and Discussion

Figures 6.5 and 6.6 show the predictions and observed data for the SC and DO verb conditions. For both data sets, the baseline predicts the vast majority of the difficulty to be encountered at the disambiguating region, reflecting the parser's preference for the direct object interpretation. This prediction is the opposite of the findings for the SC data set. For the DO set, it also predicts the majority of the difficulty to lie at the disambiguation, but in addition, it predicts a small difference between the thematic fit conditions, such that the good object nouns to be somewhat harder to process than the bad object nouns, which again is the opposite of the findings.

In contrast, our model's predictions are a very good fit to the data. For SC verbs, where no significant effects were found, our model correctly predicts the complete absence of difficulty at the disambiguation for stimuli with implausible objects, and

[2] Garnsey et al. report that two verbs were misclassified with regard to their subcategorisation preferences. The stimuli in question were not covered by our training data, so we do not consider them.

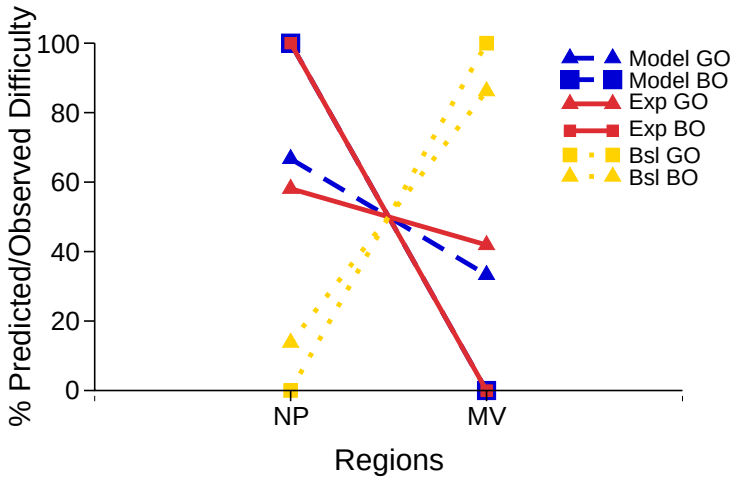


Figure 6.5.: Garnsey et al. (1997): Experimental results and model predictions for the NP/S ambiguity, SC verbs. GO: NP is plausible direct object, BO: NP is implausible direct object

medium amounts of difficulty in both regions for stimuli with plausible objects, with a tendency towards less difficulty at the disambiguation than at the NP.

For the DO verbs, our model (Figure 6.6) correctly predicts the clear crossover pattern in scaled observed difficulty, such that stimuli with plausible NPs were much easier to process at the NP than at the disambiguation, with a reverse pattern for stimuli with implausible first NPs. Again, we see very accurate predictions made on the basis of very few stimuli (as in Section 6.2.4).

6.3.4. Pickering and Traxler (1998)

Experimental setup

Pickering and Traxler (1998) varied only the thematic fit of the ambiguous NP as a direct object of the verb, without controlling verb bias. As discussed above, their eye-tracking study found a clear influence of thematic fit, replicated in a second study using the same materials and adding a one-sentence prior context to make reading more natural. We model total reading times per critical region in the eye-tracking study without context presentation. For this measure, effects were found both on the ambiguous NP and at the disambiguation. Implausible ambiguous NPs are harder to

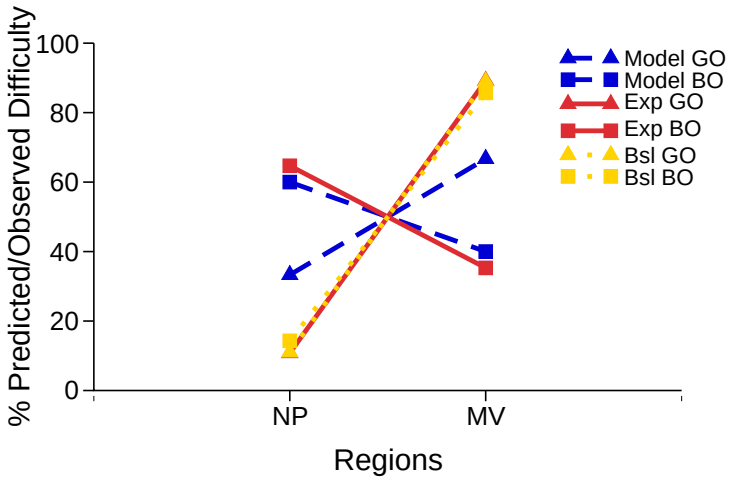


Figure 6.6.: Garnsey et al. (1997): Experimental results and model predictions for the NP/S ambiguity, DO verbs. GO: NP is plausible direct object, BO: NP is implausible direct object

read than plausible ones, both across all items and within the ambiguous items. The disambiguation is harder to read after seeing a plausible ambiguous NP than after seeing an implausible one, as expected. Also, ambiguous stimuli with implausible NPs are easier to read at the disambiguation than their unambiguous controls. The interaction between plausibility, ambiguity and region is significant by subjects.

Materials

The materials consisted of 48 stimuli. Verb-object pairs were constructed out of 24 verbs so that every verb was paired with a plausible and an implausible object, as shown in sentences (6.13) and (6.14).

(6.13) The criminal confessed his sins harmed too many people.

(6.14) The criminal confessed his gang harmed too many people.

Sentence (6.13) contains an plausible object NP, and sentence (6.14) an implausible object NP. In a norming study, plausible verb-object pairs elicited a judgement of 5.0 or higher on a 1-7 scale, and implausible verb-object pairs elicited a judgement of 2.0 or

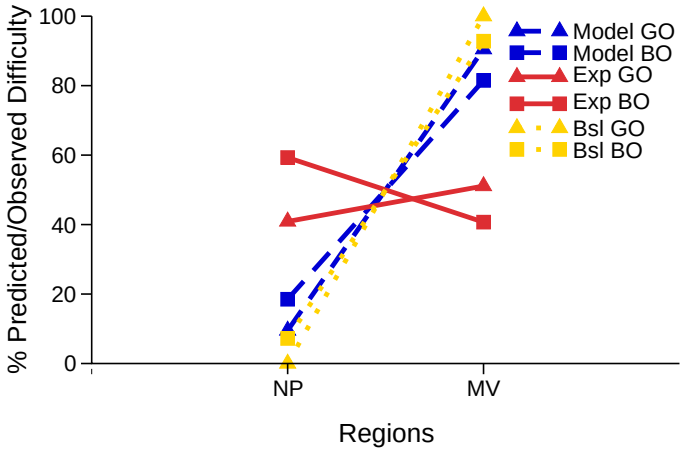


Figure 6.7.: Pickering and Traxler (1998): Experimental results and model predictions for the NP/S ambiguity. GO: NP is plausible direct object, BO: NP is implausible direct object

lower. A second norming study ensured that all objects made plausible subjects for the sentential complement continuations of the stimuli. Unambiguous controls were created for the items by adding the complementiser *that* after the first verb.

Of the 24 stimuli for each condition, 15 stimuli with plausible objects were parsed correctly, out of which 13 are covered by FrameNet. 17 stimuli with implausible objects were parsed correctly, yielding 15 stimuli covered by FrameNet. The syntactic baseline predictions are based on 14 stimuli in the good object condition and 16 stimuli in the bad object condition.

Results and Discussion

The SynSem-Integration Model predictions (dashed blue), baseline predictions (dotted yellow) and the observed difficulty (solid red) are presented in Figure 6.7. For this data set, the syntactic baseline predicts all the difficulty to be encountered at the verb, which reflects again the parser's preference for a direct object interpretation of the items. The baseline model also predicts a small difference between the thematic conditions, which however is at odds with the observations. For this data set, our model predictions are quite similar to the baseline predictions for this data set: The model clearly predicts

most of the difficulty to lie at the main verb, with somewhat more difficulty at the noun phrase for bad objects and at the disambiguation for good objects. The trend of these predictions is indeed correct, as the experimental data show bad objects to be harder to read at the noun phrase and easier at the disambiguation. However, the model obviously overestimates the amount of difficulty encountered at the disambiguation, and differs only slightly from the syntactic baseline.

A detailed analysis of the model predictions shows that the reason lies not so much with the model, but with a property of the items for this data set. In the Pickering and Traxler items, the disambiguation consists of a full verb in the majority of cases, whereas in the Garnsey et al. items, the disambiguation begins with an auxiliary. Many of the full verbs, unlike the auxiliaries, can be interpreted as the beginning of a reduced relative clause modifying the ambiguous NP. Thus, the stimulus *The biologist proved the theory explained* can be constructed to be consistent with the continuation *by his colleague was correct*, which is not intended in the experiment, as well as with *all the unclear data*, as intended by the experimenters. The semantic model systematically prefers the reduced relative reading over the embedded clause reading, reflecting a bias for the direct object interpretation of the NP. Consequently, the syntactic and semantic preferences often diverge at the point of disambiguation, causing the prediction of much difficulty in this region. Despite this effect, the number of stimuli that are interpreted as intended by the experimenters is high enough to predict the correct trend for the differences between the thematic fit conditions.

6.3.5. Numerical Evaluation

We again also present a quantitative analysis of the model's predictions for the NP/S ambiguity. Again, we pool the data from the two studies to have sufficient N for a correlation analysis. We enter scaled observed difficulty data for the DO and SC conditions from the Garnsey et al. study and for the Pickering and Traxler study into a correlation analysis with the model's predictions for these data sets. The resulting correlation coefficient of Pearson's $r = 0.688$ is significant at the $p < 0.02$ level (*Fixed/If-Worse*: $r = 0.780, p < 0.01$; *Ratio/Ratio*: $r = 0.737, p < 0.01$). This is true despite the model's qualitatively poor predictions for the Pickering and Traxler data, which suggests that the reliable correlation rests on the results for the Garnsey et al. data (but see also Section 6.6).

The baseline model again does not achieve a significant correlation to the observed data. At $r = -0.165$, the correlation coefficient is even negative due to the incorrect predictions for the Garnsey et al. results.

6.4. NP object/0

The NP/0 ambiguity shares some characteristics with the NP/S ambiguity, because again the interpretation of an ambiguous NP is affected. This NP can either serve as a direct object to the verb in an adverbial clause, as in (6.15), or as the subject in a main clause, as in (6.16).

(6.15) While the woman was editing the magazine it started to rain.

(6.16) While the woman was editing the magazine amused the reporters.

Accordingly, in the direct object reading, the NP receives a role from *editing*, but in the main clause reading, it only receives a role once the verb in the main clause, *amused*, is encountered. The difference to the NP/S ambiguity described above is that in the NP/S ambiguity, the NP as part of the verb's sentential complement is still in the argument domain or θ *domain* (Pritchett, 1992) of the verb. In the NP/0 ambiguity, the NP in the main clause is completely independent of the verb in the adverbial clause, which motivates the characterisation of this alternative as the 0 interpretation.

The ambiguous region in this structure comprises the NP, with the disambiguation following immediately at the next word. In this ambiguity, readers usually interpret the ambiguous NP as the direct object of the verb and show difficulty when it is disambiguated towards the subject of the main clause. The experimental logic is therefore usually to see whether information like verb subcategorisation preference or thematic fit can eliminate this difficulty partially or completely. Comparison is usually to versions of the experimental items where a comma (or non-object material) before the NP eliminates the ambiguity.

6.4.1. Experimental Evidence

In the literature, three factors were the focus of investigation for the NP/0 ambiguity: Verb bias, animacy of the subject NP (which can modulate verb bias) and the thematic fit of the ambiguous NP as an object of the verb. In addition, an effect of introducing disambiguating prepositional phrases or adverbials after transitive verbs has been found.

Verb Bias

An early, influential paper on the NP/0 ambiguity is Mitchell (1987). His self-paced reading study varied only verb subcategorisation bias, using transitive and intransitive verbs. Readers had more trouble at and after the disambiguation point after reading verbs with a transitive bias than after reading intransitive verbs: Apparently, they assumed the NP to be a direct object, and had trouble accommodating the main clause

subject reading at the disambiguation. However, reading the ambiguous NP after the intransitive verbs took longer than reading it after the transitive verbs, which indicates a possible reanalysis process after initially attaching the NP as an object to the intransitive verb.

These results were refuted by Adams, Clifton, and Mitchell (1998) in an eye-tracking study aimed at excluding an influence of presentation mode in the self-paced reading study, but replicated in a subsequent eye-tracking experiment by van Gompel and Pickering (2001), which used longer noun regions to catch spillover effects and which also analysed measures based on regressive eye movements.

Pickering et al. (2000) confirmed the interpretation that a direct object analysis is initially assumed even for verbs with a strong intransitive bias by showing in an eye-tracking study that the implausibility of the ambiguous NP as an object for the transitive-biased verbs disrupted processing (see also the discussion of thematic fit manipulations below).

Verb Bias Induced by Animate Subject

A number of studies have investigated the effect of varying animacy information of the subject NP with verbs that can be used as causative transitives or ergative intransitives, such as *stop*. The general result is that animate subject NPs like *police* bias readers towards adopting the causative, transitive reading of the verb, while inanimate subject NPs like *truck* bias them towards adopting the ergative, intransitive reading. Thus, they would prefer *the police stopped* to be continued by a direct object to *stopped*, while in the context of *the truck stopped*, *stopped* would be preferably interpreted as intransitive.

This effect was shown for example by Stowe (1989) in an influential study using self-paced reading combined with a grammaticality judgement task. Ambiguous NPs were plausible objects in a transitive reading of the verbs. Readers showed difficulty only if sentences were ambiguous and the subject NP was animate, showing clearly that animate subjects made the verbs behave like transitives and inanimate subjects made them behave like intransitives, for which there is no need for reanalysis at the disambiguation. Stowe did however not report any significant effects at the ambiguous NP, thus not replicating Mitchell's result regarding intransitive verbs.

Clifton (1993) replicated Stowe's result for the disambiguating region in an eye-tracking study. Animacy of the subject NP clearly influenced readers' preference to analyse the ambiguous NP as a direct object. For ambiguous NPs that were plausible direct objects, Clifton also replicated Mitchell's result: The NP was harder to read when the verb was interpreted as an intransitive (in inanimate subject conditions). This again probably indicates that readers did initially misanalyse the NP as a direct object, but recovered quickly. For this condition, Clifton even found some disruption at the disambiguation which shows that the object analysis was often not abandoned immediately. There was no evidence for this in either Mitchell's or Stowe's study.

Plausibility of the Ambiguous NP

Regarding the influence of plausibility of the ambiguous NP as a direct object for the verb, the picture emerging from the literature is again clear: Plausibility information has an effect, both on reading the ambiguous NP, and on reading the disambiguation.

At the Ambiguous NP Stowe (1989) crossed verb valency with the plausibility of the ambiguous NP as an object. At the ambiguous NP, it became clear that implausible ambiguous NPs are always difficult to read, a result which again can be interpreted as evidence that an object interpretation of the NP is constructed initially for both transitive and intransitive verbs. This is replicated for verbs with a transitive interpretation by two eye-tracking studies which both used transitive verbs (with optional intransitivity) and varied the plausibility of the ambiguous NP (Pickering and Traxler, 1998, Lipka, 2002). Findings by Pickering et al. (2000) furnish the replication for verbs with an intransitive bias.

Clifton (1993) came to slightly different results across two eye-tracking studies in which animacy information determined verb valency. When the ambiguous NP was an implausible object, he also found uniform difficulty for both verb conditions. However, in a second experiment in which the ambiguous NP was a plausible object, difficulty occurred for verbs that were treated as intransitive. This has not been found in any other study, and Clifton attributes his result to a reanalysis effect due to the insufficient plausibility of the ambiguous NP.

At the Disambiguation An effect of NP plausibility as an object is uniformly found at the disambiguation. In the Stowe (1989) study, readers had difficulty if the verbs were biased towards a transitive reading. There also was an interaction of animacy and plausibility, such that the transitive-implausible condition was faster than the transitive-plausible condition. Implausibility of the object appears to allow readers to recover from assuming the direct object reading early on. For intransitive verbs, reading times are slower after reading an implausible direct object. This finding is somewhat surprising, as the transitive verb bias together with the implausibility of the direct object interpretation should bias readers towards the ultimately correct new clause interpretation, making it easy to read the disambiguation.

Clifton (1993) found a more expected pattern: Readers showed difficulty for verbs with a transitive interpretation, but not for verbs with an intransitive interpretation, regardless of plausibility (since plausibility was varied not within, but between experiments, the direct influence of plausibility on difficulty cannot be assessed). Pickering and Traxler (1998) as well as Lipka (2002) also found that for transitive verbs, the disambiguation was harder to read after plausible objects, with only hints at difficulty for implausible objects at or after disambiguation.

For verbs with intransitive bias, Pickering et al. (2000) also found more disruption after plausible direct objects than after implausible ones, which is the reverse of Stowe's result.

In sum, it seems clear that implausible object NPs are harder to read after both transitive and intransitive verbs, while plausible object NPs are easy to integrate and cause no difficulty. At the disambiguation, the picture reverses and readers show difficulty after reading a plausible object NP. The difficulty is most pronounced for transitive verbs, but has also been shown to exist for intransitives.

Non-NP Material after the Verb

Some of the studies introduced above added non-NP material such as a prepositional phrase (PP) or an adverb after the verb of the adverbial clause (just before the ambiguous NP) to disambiguate their materials towards the new clause reading: PPs and adverbs typically do not intervene between a verb and its objects, so they all but rule out a transitive verb reading and indicate that the adverbial clause is closing. This strategy has however been shown to cause disruption for items with a transitive preference, presumably because an expectation for a direct object is disappointed.

Mitchell (1987), for example, added a PP after the verb to disambiguate his items. He found that the PP made it hard to read the ambiguous NP after transitive verbs. After reading intransitive verbs and the PP, processing of the NP was faster. Adams et al. (1998) found a similar effect with an adverb that was also used for disambiguation: There were both indications of difficulty for transitive verbs on reading the NP after the adverb and clear indications that the adverb itself was read more slowly following a transitive verb. Stowe (1989) also used prepositional phrases for disambiguation, but found no effects in the NP region.

Summary

We have seen that the thematic fit of the ambiguous NP as a direct object have direct bearing on the processing of the NP/0 ambiguity: Thematic fit of the ambiguous NP facilitates or hinders the integration of the NP as a direct object, as revealed by reading times both at the NP and at in the region of disambiguation. This semantic factor is accounted for in principle by our semantic model.

Verb subcategorisation preferences, be they global or invoked by a biasing subject NP as for ergative/causative verbs, seem to have a limited influence on the processing of the NP/0 ambiguity: A number of studies has found difficulty effects at the ambiguous NP even for intransitive verbs, suggesting that readers construct a direct object interpretation at least briefly. As for the NP/S ambiguity, our semantic model can account for this object preference through the small tree bias which prefers a direct object interpretation even if the verb is biased against this subcategorisation frame.

The SynSem-Integration model can in principle also account for the influence of the first NP on the subcategorisation bias for ergative/causative verbs through the plausibility evaluation of the semantic model.

In addition, introducing disambiguating prepositional phrases or adverbials after transitive verbs can cause processing difficulty in addition to disambiguating the ambiguity, possibly because an expectation for a direct is disappointed. In these cases, both the syntactic and semantic model predict lower probabilities for a transitive verb that is usually seen with an object to suddenly lack one, so our model is able to account for this effect.

6.4.2. Data Selection and Evaluation Method

We model the experimental results from Pickering and Traxler (1998) and from Pickering et al. (2000), again because only for these studies, the required amount of at least 10 stimuli per condition were available. We make predictions for the critical noun and verb regions in the modelled experiments, again ignoring post-nominal and post-verbal regions, as justified in Section 6.3.2.

6.4.3. Pickering and Traxler (1998)

Experimental Setup

Pickering and Traxler (1998) varied the thematic fit of the ambiguous NP as a direct object of the verb, using verbs that can be assumed to have a bias for the transitive reading. As discussed above, their eye-tracking study found a clear influence of thematic fit, generally replicated in a second study using the same materials and adding a one-sentence prior context to make reading more natural. We model total reading times per critical region in the initial study without context presentation. For this measure, effects were found both on the ambiguous NP and at the disambiguation. First, implausible ambiguous NPs were harder to read than plausible ones. The disambiguation was harder to read after seeing a plausible ambiguous NP than after seeing an implausible one, for both ambiguous and unambiguous sentences. After seeing an implausible ambiguous NP, there were no traces of disruption in the total time measure: Ambiguous items were read as fast as unambiguous ones. In addition, the three-way interaction between ambiguity, plausibility and region was significant.

Materials

The materials consisted of 48 stimuli. Verb-object pairs were constructed so that every object was paired with two verbs, one for which it was a plausible and one for which it was an implausible object, as demonstrated by sentences (6.17), which

shows the plausible verb-object combination, and (6.18), which shows the implausible combination.

(6.17) As the woman edited the magazine amused all the reporters.

(6.18) As the woman sailed the magazine amused all the reporters.

In a norming study, plausible verb-object pairs elicited a judgement of 5.0 or higher on a 1-7 scale, and implausible verb-object pairs elicited a judgement of 2.0 or lower. A second norming study ensured that all objects made plausible subjects for the sentential complement continuations of the stimuli. Verbs and objects were counterbalanced to create identical critical regions across conditions. Unambiguous controls were created by simply adding a comma after the verb in the adverbial clause.

To increase coverage of the syntactic parser, we to standardised the initial adverbial clause. The parser assigns the correct structure to sentences with adverbial clauses of the form *while the X was Ying...*, but does not correctly close the adverbial clause at the ambiguous NP or after disambiguation for other forms of adverbial clauses such as *after the X had Yed...*. We therefore standardised all adverbial clauses to the *while* type using the past progressive tense.

Note that the re-formulation does not affect the analyses yielded by the semantic model. While human readers show more difficulty at the disambiguation of past-tense NP/0 structures than past progressive structures (Frazier, Carminati, Cook, Majewski, and Rayner, 2005) and thus react to cues of tense, our semantic model only evaluates pairs of verb-argument lemmas and has no representation for the semantics of tense. It therefore makes the same predictions for the original and the adapted formulation of the items. The re-formulation also does not affect the syntactic analysis of the items other than allowing the parser to assign the correct sentence structure more often.

Of the 24 stimuli for each of the two conditions, 14 stimuli were parsed correctly, out of which 10 stimuli each were covered by FrameNet. The baseline predictions are derived from 13 stimuli in each condition.

Results and Discussion

For this data set, the syntactic baseline again predicts all difficulty to be encountered at the disambiguating main verb, with no difference in difficulty between the semantic conditions. As shown in Figure 6.8, our model only correctly predicts a difference in relative difficulty for the conditions in each region: On the ambiguous NP, it predicts the bad object NPs to be harder to read than the good object NPs. On the disambiguation, it predicts more difficulty after reading a good object NP. However, the difference between the predictions for each condition is very small, and the model drastically over-estimates the difficulty encountered at the disambiguation.

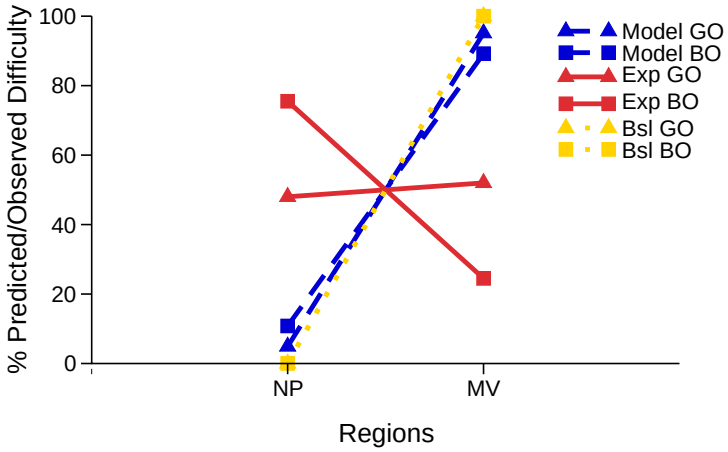


Figure 6.8.: Pickering and Traxler (1998): Experimental results and model predictions for the NP/0 ambiguity. GO: ambiguous NP is good object, BO: ambiguous NP is bad object.

This result again hinges partially on the formulation of the experimental items, just as for the Pickering and Traxler NP/S data set discussed in Section 6.4.3 above. Again, the items allow a reduced relative interpretation of the main verb that was intended to disambiguate the ambiguity, and the semantic model prefers this unintended reading over the correct analysis, causing the prediction of difficulty at the disambiguation. However, there is a second factor: For this ambiguity, the bias towards interpreting the ambiguous NP as a direct object is even stronger than for the NP/S ambiguity, as the semantic model has a bias towards preferring analyses that contain frequently seen role sets for each verb (recall Section 5.2.2). Since transitive verbs are seen more frequently with a subject and object in the training data than in an intransitive reading with just a subject, the semantic model is biased towards preferring analyses that assign an object role to the ambiguous NP instead of making it the subject of another transitive verb (momentarily) without an object.

Since the model thus strongly disprefers the correct analysis at the disambiguation for this data set, there are hardly any stimuli for which the intended analyses are preferred. This leads to the uniform prediction of extreme difficulty at the disambiguation, which is not mirrored in the data.

6.4.4. Pickering et al. (2000)

Experimental Setup

Pickering et al. (2000) used optionally transitive verbs with a strong intransitive bias. They varied the thematic fit of the ambiguous NP as a direct object of the verb. They did not create control items by inserting a comma, but used the plausible condition as a comparison for the implausible one and vice versa. We model the total time findings from their eye-tracking study for each region, which were as follows: On the NP, total time was longer for implausible objects, and on the verb, total time was longer for plausible object stimuli. The interaction of plausibility and region was also significant.

Materials

The materials consisted of 52 stimuli. 26 verbs were chosen that had a strong intransitive bias, both in free sentence production and in a gated completion task. Verb-object pairs were then constructed so that every verb was paired with two objects, a plausible and an implausible one, as demonstrated by sentences (6.19), which contains a plausible object, and (6.20), which contains an implausible object.

(6.19) While the pilot was flying the plane stood over by the fence.

(6.20) While the pilot was flying the horse stood over by the fence.

In a norming study, plausible verb-object pairs elicited a rating of more than 5.0 on a 1-7 scale, and implausible verb-object pairs elicited a rating of less than 2.0. A second norming study ensured that all objects made plausible subjects for the sentential complement continuations of the stimuli.

The materials contained a post-nominal region consisting of a two-to-four word noun modification (relative clause or prepositional phrase). Again, as for the materials in Section 6.4.3, the post-nominal material created difficulty for the parser and was deleted. Also, we again standardised all adverbial clauses to the *while* type using the past progressive tense. Recall that this re-formulation has no effect on the semantic analysis of the items by the semantic model or on the syntactic analyses (other than higher accuracy).

Of the 26 stimuli for each of the two plausibility conditions, 13 stimuli each were parsed correctly. 11 of these were also covered by FrameNet. The predictions of the baseline model are based on 14 stimuli in the good objects condition and 13 stimuli in the bad objects condition.

Results and Discussion

Since there are no unambiguous controls in this study, Pickering et al. compare the reading times for the good object conditions to the reading times for the bad object

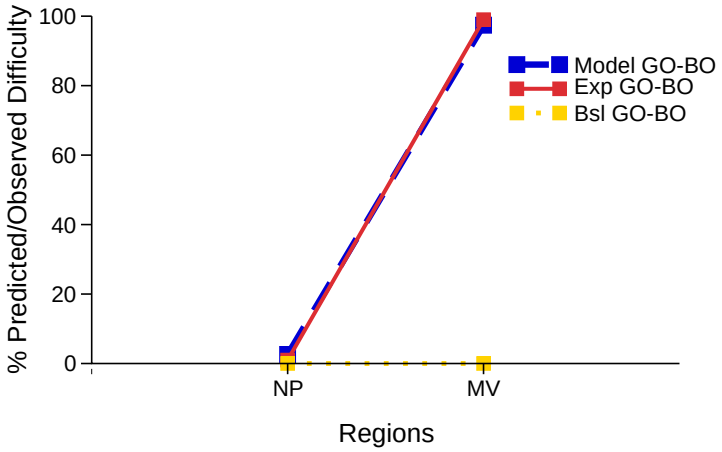


Figure 6.9.: Pickering et al. (2000): Experimental results and model predictions for the NP/0 ambiguity. GO-BO: Difficulty as difference in reading times between good objects (GO) and bad objects (BO).

conditions. The plot of observed and predicted difficulty in Figure 6.9 therefore shows the relative difficulty of good objects as opposed to bad objects. It was constructed by plotting the scaled results of subtracting the reading times for good object sentences from the reading times for bad object sentences.

For this data set, the syntactic baseline again predicts no difference in difficulty between the semantic conditions. This manifests as a straight line on the abscissa. Our model correctly predicts that good objects are easy to read in comparison to bad objects at the ambiguous NP, and that bad objects in contrast are hard to read in comparison with good objects at the disambiguation.

Note, however, that the Pickering et al. items again show the same characteristic as the Pickering and Traxler sets of items with regard to an unintended ambiguity arising at the point of disambiguation. In addition, the model again strongly disprefers the correct analysis of the items at the disambiguation. The correct predictions are reached again through a small number of stimuli that are processed as intended by the experimenters. For this data set, the large amounts of difficulty predicted for the disambiguating region in both conditions equal each other out when the predictions for the bad object condition are subtracted from those for the good object condition, which leads to the accurate model predictions plotted in Figure 6.9.

6.4.5. Numerical Evaluation

For this data set, a numerical evaluation is difficult even for the pooled data: the Pickering and Traxler study yields four data points, and the Pickering et al. study only two. For six data points, Pearson's r has to be at least 0.754 for the correlation to be significant. It is therefore not unexpected for the correlation between observed and predicted difficulty to be not significant ($p > 0.1$) for all models. In addition, however, r is low at about $r = 0.3$ for the *If-Worse* models ($r = 0.2$ for the *Ratio/Ratio* model) because of the models' failure to estimate the relative difference across regions for each condition for the Pickering and Traxler data. However, the syntactic baseline still does not outperform our model: The correlation to the observed data is also not significant, and at $r = -0.260$, the correlation coefficient is even negative, because the syntactic baseline predicts no difference at all between the conditions.

6.5. PP Attachment

A PP attachment ambiguity usually arises in utterances like (6.21) and (6.22), where the attachment of the PP is possible both to the main verb and to the object NP.

(6.21) The cop saw the crook with the binoculars.

(6.22) The cop saw the crook with gun.

In the example sentences, the attachment is disambiguated by semantic plausibility: Sentence (6.21) is more plausible as a verb attachment, while sentence (6.22) is more plausible as an NP attachment. There is no syntactic disambiguation for the PP attachment ambiguity, and if the semantics of the PP do not allow clear disambiguation, a PP attachment ambiguity can even remain completely unresolved, *globally ambiguous*, for example as in *The cop chased the crook in a car*.

The experimental logic of the studies reviewed below is to test whether an experimental factor (for example, verb subcategorisation preference) influences attachment preferences. Readers' attachment preferences are determined by comparing reading times for the alternative attachments (where the intended attachment is semantically disambiguated). Note that the PP Attachment ambiguity cannot incur revision cost in our model, because the set of verb-argument pairs for each analysis is only extended at the noun in the PP. If the the attachment previously assumed on the basis of only the verb and changes at this point, the change is therefore not reflected in the set of verb-argument pairs, and no revision cost is incurred.

6.5.1. Experimental Evidence

A large literature exists on the PP attachment ambiguity. In addition to the strong influence of semantic plausibility of the attachment, which is used for disambiguation in all studies reviewed below, the influence of several factors on attachment preferences have been discussed. The PP Attachment ambiguity has for example been used intensively to investigate effects of referential context (e.g., Altmann and Steedman, 1988, Rayner, Garrod, and Perfetti, 1992, Britt, 1994, Liversedge, Pickering, Branigan, and van Gompel, 1998), and effects of NP definiteness have been found (e.g., Spivey-Knowlton and Sedivy, 1995). We are more concerned here however with the proposed influence of parsing principles versus verb preferences and effects hinging on the argument versus adjunct status of the PP.

Parsing Principles versus Verb Preference

As proponents of the influence of parsing principles, Rayner et al. (1983) present evidence from an eye-tracking study that the parsing principle of *Minimal Attachment* (construct the analysis that requires the postulation of fewest syntactic nodes) determines initial attachment. They find an attachment preference to the verb, which is indeed the analysis requiring fewest nodes under the grammar assumed by Rayner et al. This result is confirmed by an eye-tracking and a self-paced reading study in Clifton, Speer, and Abney (1993).

However, there is evidence that this effect may have been caused by verb preference rather than by the application of a parsing principle. For example, Spivey-Knowlton and Sedivy (1995) analyse the set of temporally ambiguous *with*-PPs occurring in the Brown corpus and find that there is a reliable preference for verb attachment across all PPs, but that an analysis by verb type reveals a preference for NP attachment for some verb types, for example verbs describing psychological states or perception. For such verbs, a self-paced reading study confirmed a preference for NP-attachment also during on-line processing. Taraban and McClelland (1988) came to a similar result: While they replicated Rayner et al.'s results in a self-paced reading study, they also demonstrated in rating and completion studies that the Rayner et al. materials were biased towards the VP reading. They then constructed materials with verbs biased towards NP attachment and showed that for these materials, the pattern of results found by Rayner et al. was exactly reversed. These findings clearly argue against the application of a fixed processing strategy as proposed by Rayner et al. (1983) and Clifton et al. (1993).

Finally, van Gompel, Pickering, and Traxler (2001) add some interesting results for the case where the verb is unbiased with regard to PP attachment. They show that for cases of unbiased attachment, readers read a globally ambiguous condition fastest and show difficulty in both disambiguated conditions. These results also cannot be

reconciled with the existence of a general parsing principle that decides the attachment even when no subcategorisation preference is available. Instead, it is more plausible to assume that readers as a group initially decided for each attachment analysis about half the time in the absence of a verb bias. Thus, either semantic attachment disambiguation caused some difficulty across all readers, while the globally ambiguous analysis was always consistent with readers' initial decisions and thus caused no difficulty. These results however do not allow any conclusions about whether individual readers had overall initial attachment preferences or whether each reader decided for each of the analyses about half the time.

As a final observation regarding verb-specific attachment preferences, Taraban and McClelland (1988) also investigated whether violation of the expected thematic role of the PP has an effect over and above violation of the expected attachment. They found that reading a PP which conforms to the verb's preferred attachment but introduces an unexpected thematic role (e.g., a co-subject instead of an instrument as in *clean with the manager* instead of *clean with a broom*) causes as much difficulty as encountering a violation of the preferred verb attachment. When both the thematic role and the attachment of the PP go against the verb's preferences, little additional difficulty is observed. Taraban and McClelland take this to indicate that verbs not only specify their general preference for or against attachment, but also a specific preferred thematic role.

PP Argument Status

There are other arguments in the literature that assume that the verb-specific attachment preference effect is not caused by verb subcategorisation preference, but by expectations about the PP's argument status. This has been investigated for example by Clifton et al. (1993), Schütze and Gibson (1999) and Boland and Blodgett (2006) and Liversedge et al. (1998). Clifton et al. (1993), as discussed above, found an initial preference for attachment to the VP, but later in the sentence also observed that argument attachments to the VP and the NP were read faster than adjunct attachments. This implies that the status of the PP has an influence beyond the verb's subcategorisation preference. However, it appears that argument status, PP length and plausibility were not carefully controlled in their items.

A preference for arguments over adjuncts was also found by self-paced reading experiments in Schütze and Gibson (1999), who used items of a very similar form as Clifton et al., but unlike the latter controlled very carefully for argument status and plausibility. They found a preference for argument attachment (which was always to the NP) over modifier attachment (always to the verb), thus showing a preference for NP attachment instead of replicating the verb attachment preference found in Clifton et al. (1993). A large caveat with this study is that the verbs' attachment preferences were not pre-tested. Therefore, the NP (argument) attachment preference may also have been modulated by verb preference.

Boland and Blodgett (2006) in an eye-tracking study used verbs with a tendency to prefer PP attachments, and argument attachment could be either to the noun or to the verb and verbs had a preference for verb attachment of the PPs. They found that argument attachments were always easier to read than adjunct attachments, and some indications that verb attachments were still easier than noun attachments. This latter effect, if real, would indicate that verb preference has a separate influence from argument preference.

Another corroborating result for argument preference comes from an eye-tracking experiment in Liversedge et al. (1998), which used *by*-PPs that are ambiguous between an argument reading (agent in a passive clause) or an adjunct reading (locative). The study found a general preference to process a *by*-PP unambiguously attached to the VP as an argument (namely the agent) rather than an adjunct in isolated passive sentences.

Interestingly, the preference to interpret PPs as arguments rather than adjuncts also appears to depend on the preposition in the PP. In the PropBank corpus, which is a more representative sample of written English than FrameNet, role-bearing PPs appear approximately the same amount of the time as arguments and as adjuncts (c. 27,000 argument occurrences vs c. 24,000 non-argument occurrences). Two prepositions that are often used in the investigation of PP Attachment, *with* and *by*, are however seen vastly more often as arguments than as adjuncts: 85% of *by*-PPs and 64% of *with*-PPs are arguments.

Summary

We have reviewed the literature on the influence of a number of syntactic and semantic factors on the processing of the PP attachment ambiguity. The influence of parsing principles like Minimal Attachment has been clearly disproven by a number of experiments. Rather, it appears that verb-specific information such as verb subcategorisation preferences or even a preference for a specific thematic role is used. Furthermore, there appears to be preference to interpret PPs as arguments rather than adjuncts, but it is not quite clear how these factors interact. Since no syntactic disambiguation of the PP Attachment ambiguity exists, all studies use thematic fit to disambiguate the attachment.

Our model accounts for this semantic disambiguation by the plausibility predictions of the semantic model. Verb subcategorisation preferences are covered by the syntactic model. The preference for argument role assignment over adjunct role assignment can also be modelled to the extent that it is reflected in the training data. Recall that FrameNet focuses on providing corpus samples for argument roles, so that adjunct roles are generally infrequent in the corpus. Note that we do not account for role assignment by nouns, however, and therefore cannot account for NP complements.

6.5.2. Data Selection and Evaluation Method

We model the experimental results from two studies that varied PP attachment and verb preferences. Only the items from the studies by Rayner et al. (1983) and Taraban and McClelland (1988) were sufficiently covered by our syntactic and semantic models to allow modelling.

We make predictions for the last word in each of the critical regions in the modelled experiments. Since there are no syntactically unambiguous control materials in either study, we use the difference in reading times for verb and NP attachment as indication of difficulty.

6.5.3. Rayner et al. (1983)

Experimental Setup

Rayner et al. (1983) varied the plausibility of PP attachment, making either verb or NP attachment semantically more plausible. They measured reading times in two regions: The ambiguous region made up of the object NP and preposition, and the disambiguation, made up by the NP in the PP and following material. We model total times from their eye-tracking study. With this measure, they found an interaction of plausible attachment site and region: Readers took longer to read the noun in the PP if it was biased towards NP attachment rather than verb attachment. This also caused significantly longer reading times overall for NP-attachment biased sentences.

Materials

There were 48 stimuli, created from 12 sentences by manipulating the thematic fit of PP attachment (biasing towards verb attachment or NP attachment) and by manipulating the length of the NP. Thematic fit and verb attachment preference were not normed. Sentence (6.23) shows a bias for verb attachment and contains a short NP, while sentence (6.24) shows a bias for NP attachment.

(6.23) The spy saw the cop with binoculars, but the cop didn't see him.

(6.24) The spy saw the cop with a revolver, but the cop didn't see him.

We only used the 24 stimuli with short NPs, because no effect of NP length was found and the syntactic parser had previously shown problems with complex noun phrases (cf. the deletion of the post-nominal region for NP/S and NP/0 data, above).

From the twelve stimuli in each condition, ten each were parsed correctly. Nine stimuli each were covered by FrameNet. Again, this falls only slightly short of the minimum number of stimuli. The baseline predictions are based on eleven stimuli in the verb attachment condition and twelve in the NP attachment condition.

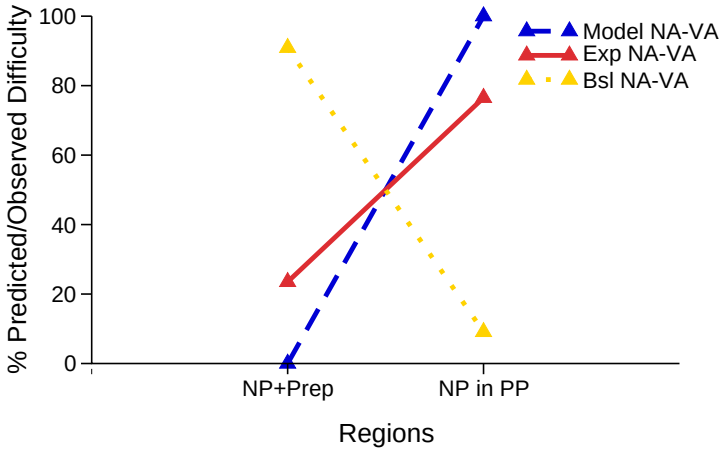


Figure 6.10.: Rayner et al. (1983): Experimental results and model predictions for the PP Attachment ambiguity. NA-VA: Observed/predicted difficulty as difference between NP attachment and verb attachment conditions.

Results and Discussion

Since there are no unambiguous controls, we use the difference between the attachment conditions as an indication of relative difficulty with the conditions. The plot in Figure 6.10 shows the difference between predicted or observed difficulty in the NP attachment condition and predicted or observed difficulty in the verb attachment condition.

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. Scaling moves both this negative difference and the difficulty prediction of zero for the NP+Prep higher into positive space. This prediction comes about because the parser only predicts difficulty for one stimulus at the noun in the PP in the verb attachment condition. Compared to no difficulty at all in all other conditions, this chance prediction means that the majority of difficulty is predicted for verb attachment.

The SynSem-Integration model similarly predicts that there should be little difference in difficulty between the conditions on the NP+preposition material that is identical in both conditions. Once the noun in the PP is read, the model however 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 thus correspond almost exactly to the pattern found in the data, while the baseline model's predictions are the opposite of the observations.

6.5.4. Taraban and McClelland (1988)

Experimental Setup

In their Experiment 1A, Taraban and McClelland (1988) replicate the results from Rayner et al. (presented above) in a self-paced reading task, and also test their own stimuli. Taraban and McClelland identify a bias towards verb attachment in a completion test of the Rayner et al. items, and therefore offset those items with an equal amount of items with verbs that show a bias towards NP attachment. This study thus varies verb attachment preference in addition to manipulating attachment plausibility. Attachment preference and plausibility were normed for all items, including the Rayner et al. items.

Taraban and McClelland replicate the results from Experiment 1A in their Experiment 1B, where the sentences are continued beyond the ambiguous PP. Note that since we are only interested in effects on the critical regions, and since the results of the experiments were comparable, there is no difference between the experiments for our purposes. We model reading times from Experiment 1A, where the Rayner et al. items were read faster when the PP attachment was disambiguated towards verb attachment, as in Rayner et al.'s study. The Taraban and McClelland items, however, were read faster when the PP attachment was disambiguated towards NP attachment. Measurements were taken only for the noun phrase in the ambiguous PP.

Materials

The items from Rayner et al. (see Section 6.5.3) were used in addition to 18 new stimuli with the same structure. The effect of the plausibility manipulation was pre-tested in a rating study, and the effect of the verb preference manipulation across the stimuli subsets was pre-tested in a gated completion pre-test.

Taraban and McClelland found that one of the stimuli from Rayner et al. was not interpreted as intended under the plausibility manipulation. The corresponding item was excluded from the study, such that only eleven Rayner et al. items were used.

From the 29 stimuli per plausibility condition, 20 verb-attachment stimuli and 19 NP-attachment stimuli were parsed correctly. Of those, 10 Taraban and McClelland NP attachment stimuli and 11 Taraban and McClelland verb attachment stimuli were covered by FrameNet. Nine stimuli each from Rayner et al. were covered by FrameNet.

The baseline model uses 14 Taraban and McClelland stimuli in each condition, and the same number of Rayner et al. stimuli as above.

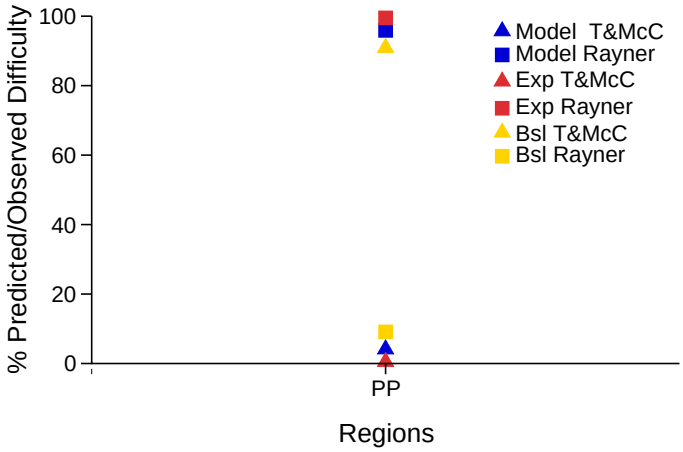


Figure 6.11.: Taraban and McClelland (1988): Experimental results and model predictions for the PP Attachment ambiguity. VA-NA: Observed/predicted difficulty as difference between verb attachment and NP attachment conditions

Results and Discussion

For the Taraban and McClelland (1988) study, we only have one set of measurements at the noun phrase in the ambiguous PP. We subtract the reading times (and difficulty predictions) for the verb attachment condition from those for the NP attachment condition to arrive at indications of relative difficulty. Scaling of observed and predicted difficulty for this data set is done across conditions: Predictions and observations sum to 1 over the values for the Taraban and McClelland and the Rayner et al. data sets.

Figure 6.11 shows that for both the SynSem-Integration model predictions and observed data, there is vastly more difficulty for the Rayner et al. stimuli (bias for noun attachment) than for the Taraban and McClelland stimuli (verb attachment bias). This is caused by more difficulty being encountered for verb attachment in the Taraban and McClelland set of items, which leads to a negative difference. For the Rayner et al. set of items, noun attachment is much harder than verb attachment, which leads to a positive difference. Scaling preserves the large difference in observed difficulty between the two conditions, but places the Taraban and McClelland results in positive space.

For this data set, the baseline model again makes the inverse prediction because it

Model	All Data		No Garnsey et al.	
Baseline	r=-0.223,	ns	r=-0.229,	ns
Rank/If-Worse	r=0.700,	***	r=0.690,	***
Fixed/If-Worse	r=0.737,	***	r=0.661,	***
Ratio/Ratio	r=0.639,	***	r=0.577,	**

Table 6.1.: Correlations between model predictions and observations for all studies and excluding the Garnsey et al. data points. $** : \ p < 0.01, *** : \ p < 0.001$

does not predict any difference between the semantically defined attachment conditions for the Taraban and McClelland stimuli. The negative difficulty predicted for the Rayner et al. stimuli is therefore large in comparison, and still points in the wrong direction. The baseline model's predictions thus again do not differentiate between the semantically defined attachment decisions, and the chance prediction of difficulty in a single condition leads to altogether incorrect predictions overall.

6.5.5. Numerical Evaluation

For this data set, a numerical evaluation is again impossible, because we only have four data points available: Two for the Rayner et al. data (one for each region) and another two for the Taraban and McClelland data (one for each set of items). For only four data points, no correlation analysis can be meaningfully computed. However, the qualitative data patterns clearly speak to the reliability of the model predictions, and the two unplotted models perform very similarly to the plotted *Rank/If-Worse* model.

6.6. Correlating All Predictions and Observations

As a final quantitative analysis, we pool the data from all experiments discussed above and correlate all predicted data to all observations. The results are given in Table 6.1. All three instantiations of the SynSem-Integration model make predictions that are strongly and highly significantly correlated to the observations. Over all 36 data points, the *Rank/If-Worse* model (which furnished the plotted data) and the *Fixed/If-Worse* model perform very comparably, while the *Ratio/Ratio* model lags behind a little, presumably due to the influence of noise over the relatively small sets of test stimuli. The correlation is significant for all three models on the $p \leq 0.001$ level. In contrast, the syntactic baseline model does not achieve a significant correlation with the observed data and even shows a negative correlation coefficient, which reflects the fact that the

baseline model predicted the exact opposite of the observed data for some test sets and failed to correctly predict the observation patterns for the other data sets.

One reservation about this analysis might be that it includes the Garnsey et al. SC and DO data sets, which come from the same study as the development set. One might argue that optimising on one data subset from a study makes it likely that the other data subsets from the study will also be indirectly optimised. Therefore, we repeated the correlation test for the pooled data without using the Garnsey et al. data sets (see the right side of Table 6.1). The correlation coefficients drop by about 0.04 for the *If-Worse* models, but the correlations remain significant on the $p < 0.001$ level. For the *Ratio/Ratio* model, the drop is a little larger, because this model makes extremely accurate predictions for the Garnsey et al. data sets, but the correlation also remains significant. Even without the data sets that are arguably most similar to the development data, the models thus still achieve a highly significant correlation to the observed data at little cost in correlation coefficient.

We have demonstrated that the SynSem-Integration model's difficulty predictions are reliable predictors of human processing difficulty, both for individual data sets (where they were large enough) and across all modelled data points. The *If-Worse* models performed very similarly, while the *Ratio/Ratio* model did somewhat worse due to noise. All three instantiations of the SynSem-Integration model however clearly outperform the baseline model based on a lexicalised parser, which does not achieve a correlation to the test data and even shows a negative correlation coefficient.

6.7. Discussion

The experiments in Sections 6.2 to 6.5 and the overall analysis in Section 6.6 have demonstrated that the SynSem-Integration model reliably predicts patterns of difficulty observed in human reading across a number of different ambiguity phenomena. The SynSem-Integration model's predictions are not only qualitatively plausible, but are also significantly correlated to the experimentally observed data. A baseline model making difficulty predictions based on the change in the preferred syntactic analysis found by a lexicalised syntactic parser was shown to make qualitatively implausible predictions, which was mirrored in the absence of significant correlations of the baseline predictions with human data.

6.7.1. Error Analysis

The SynSem-Integration model proved to perform less reliably for the NP/0 ambiguity than for the other phenomena, so that there is no significant correlation of the pooled NP/0 predictions to the experimental observations. We found that this is caused by a preference of the semantic model for analyses that are not predicted by the literature to

be constructed. This preference leads to a disproportional prediction of conflict cost, because these analyses always differ from those preferred by the syntactic model.[3]

Recall that the semantic model is by its implementation biased to prefer analyses that predict a verb to occur with its most frequent role combination (see Section 5.2.2). The *preferred role set bias* helps the semantic model overcome the preference for assigning as few roles as possible that is owed to its formulation. When comparing analyses for a transitive verb, the semantic model without the bias would prefer an analysis that only allows role assignment to one argument, since assignments with fewer arguments are more likely than those where the probabilities of several role assignments are multiplied to reach the final probability. The preferred role set bias instead leads to a preference for analyses that allow the transitive verb to assign roles to a syntactic subject and object, because transitive verbs are most frequently seen in this configuration in the training corpus. In general, the bias thus leads to a preference for analyses that allow the verb to fill all its preferred argument slots as early as possible and thus ensures that roles are assigned to incoming material as early as possible, in accordance with the processing assumptions made, e.g., by Pritchett (1992) or Crocker (1996).

However, when processing the NP/0 ambiguity, the preferred role set bias tends to lead the semantic model away from the assumed correct preference: The disambiguation of the NP/0 ambiguity leads to an analysis which contains two transitive verbs in intransitive readings. This is strongly dispreferred by the semantic model, since both verbs have to be assumed to occur with a dispreferred role set. This dispreference for the correct analysis can be seen as an instantiation of difficulty arising when a previously assigned thematic role has to be withdrawn and cannot be replaced with another thematic role from the same verb covering the previous role filler. Pritchett's theory explicitly predicts large processing difficulty in this case, and the NP/0 ambiguity (where this type of reanalysis has to take place) has been shown to be harder to process than the related NP/S ambiguity, where reanalysis can replace the role assigned to a direct object with one assigned to an embedded clause, and the former object remains part of an argument of the verb (Sturt et al., 1999). The preferred role set bias is thus theoretically plausible, but in our experiments still leads the SynSem-Integration model to make incorrect predictions for the NP/0 ambiguity.

To quantify the amount of unexpected analyses preferred by the semantic and syntactic models, we conducted an analysis of the model's cost predictions and classified the predictions into two categories: Predictions based on syntactic and semantic preferences that are consistent with literature assumptions about possible analyses of the experimental items, and predictions that are based on spurious preferences for other

[3]It might appear promising to use an NP/S data set for parameter setting to alleviate this problem. However, because the conflict cost stems from the semantic model's preference for a completely unintended structure in the NP/0 materials, and not from the magnitude of individual difficulty predictions or the definition of the preferred structure, this would not solve the problem of spuriously predicted conflict cost.

analyses. The error analysis shows that overall, 44% of all predicted cost (30% for the *Rank/If-Worse* and the *Ratio/Ratio* models) is due to conflict cost caused by spurious model preferences (there is no spuriously predicted semantic cost). The largest part of these spurious cost predictions, namely 72% (65% *Fixed/If-Worse*, 69% *Ratio/Ratio*), are those made for the two NP/0 data sets.

The single data set that contributes most to the remaining spurious predictions is the Pickering and Traxler NP/S data. For this data set, the formulation of the items often allows the disambiguating main verb to be interpreted as a reduced relative modifying the ambiguous NP. This interpretation was not intended by the experimenters, but the semantic model generally prefers it because it often appears to correlate better with the verb's preferred argument realisation. When the predictions made for the Pickering and Traxler NP/S data set are also disregarded, the *Rank/If-Worse* model predicts only 10% cost due to spurious preferences, and spurious cost predictions in the other two model instantiations go down to 12% (*Fixed/If-Worse*) and 8%(*Ratio/Ratio*). This means that the analyses preferred by the syntactic and semantic models generally correspond to one of the analyses involved in the ambiguity phenomena. Except for the problematic cases described above, the SynSem-Integration model therefore reliably bases its predictions on the alternative analyses assumed by the experimenters.

The error analysis reported above distinguishes between model preferences for analyses that are intended to be constructed for the experimental items and those that are not. It therefore accepts as correct model preferences for analyses that correspond to one of the two assumed alternative analyses of an ambiguous item, but that are not preferred at the current stage of incremental processing, according to literature assumptions. For example, a syntactic preference for a reduced relative analysis already at the ambiguous verb in the MC/RR ambiguity would count as correct, even though the experimenters assumed that the stimulus would be analysed as a main clause at this point. This may seem overly permissive. A second error, more rigid, analysis suggests itself: It would determine the percentage of semantic and syntactic preferences that are in accordance with the experimenters' assumptions.

However, it appears that comparing the model's preferences to the experimenters' assumptions is not informative for model development, because it does not necessarily lead to an improvement in the predictions of the experimental observations. An error analysis serves the purpose of identifying possible improvements to the model that will eventually allow the model predictions to become more similar to the observations in the test data. However, as we minimise divergence of the syntactic and semantic models' preferences from the literature assumptions, we are not guaranteed to predict the human data more correctly. Instead, the model's predictions will simulate more correctly the pattern of processing difficulty hypothesised in advance by the experimenters in accordance with the assumptions about preferred analyses of the experimental items. The theoretically assumed distribution of difficulty is however always much more clear-cut than the actual experimental findings, predicting clearly

defined difficulty to occur only in one or two clearly defined regions and nowhere else. These clear-cut expectations are usually not exactly mirrored in the experimental findings, however: If they appear, effects may be weaker than expected or exhibit a trend in an unexpected direction. Therefore, a model that exactly mimics the experimenters' assumptions about the properties of the items will not exactly predict the corresponding experimental findings. Even norming results, which establish that the experimenter's assumptions do indeed hold for the experimental items, do not always correctly predict the participants' reactions. For example, Binder (2001) created predictions for her eye-tracking experiment of discourse influences on the MC/RR ambiguity using the Constraint-Integration model (Spivey and Tanenhaus, 1998) and setting the constraints according to her norming results. The predicted effects were substantially larger than the effects that were actually observed.

Instead of aiming to model the experimenters' assumptions about item properties, the SynSem-Integration model is free to make some predictions that do not correspond to the assumed properties of the items. It thereby accounts for noise in the experimental data that can be due to several sources. One source is an unexpected reaction of the participants to some stimuli, possibly because they do not share the expected preferences (for example of verb subcategorisation or main clause interpretation). Such unexpected properties are predicted by the SynSem-Integration model if they are mirrored in the training data. Of course, we cannot guarantee that the preferences extracted from the training data exactly correspond to human preferences. However, recall from the discussion in Section 2.1 that it appears plausible to assume that corpus preferences match processing preferences to some degree. In addition, the SynSem-Integration model's success at predicting human processing data strongly suggests that the corpus resources show similar preferences to those employed by readers.

Likewise, the SynSem-Integration model can account for systematic confounding biases in the items, as long as these are due to a factor within the range of the model, for example to diverging verb subcategorisation preferences or the higher plausibility of one alternative analysis over the other in an ambiguous region. Finally, both the model and the experimental data contain some random noise: This noise is due to readers' individual preferences in the experimental data, while in the model it stems from the combination of two presumably noisy probabilistic models.

In sum, by allowing the preferences of the semantic and syntactic model to differ from the assumed preferences in the literature, the SynSem-Integration model accounts for the observed data more accurately than a model could that predicts exactly the preferences stipulated by the experimenters. The diverging preferences model true divergences of assumed and actual stimulus preferences as well as noise in the observed data that is due to readers' individual differences.

6.7.2. Theoretical Implications of Model Performance

The failure of the lexicalised parser baseline model to predict the patterns of human processing difficulty in our test data clearly demonstrates that a pure PCFG-based parser model is unable to capture the effects of semantic plausibility in the data. The parser model fails despite making use of head-head dependencies, which should in principle allow it to evaluate structural analyses with respect to the likelihood of seeing the verb-argument pairs they contain. However, the sparseness of the relevant head-head dependencies in the training data precludes their effective use.

These sparseness problems occur both for test data of the same genre and for psycholinguistic items: On an unseen portion of the Penn Treebank training data, the fully lexicalised parser model used as a baseline hardly outperformed the partially lexicalised parser (the SynSem-Integration model's syntactic model) in Section 5.1.3. Further, in Chapter 3, the coverage results of the unsmoothed semantic model trained on the PropBank (which adds thematic role annotation to the Penn Treebank) showed that hardly any verb-argument pairs in the test data were seen in the training corpus. To cover head-head relations sufficiently to make plausibility predictions, a fully lexicalised parser would therefore have to be trained on much larger reliably annotated corpora than are available today, and presumably will be available in the near future, given the high cost of human annotation. Further, it is not clear whether the evaluation of the frequency of head-head co-occurrence in syntactic structure is an appropriate way of modelling human plausibility evaluation, which can be expected to take place at least at the level of thematic roles (recall the discussion of grammatical functions and thematic roles as instantiations of verb-argument relations in Section 2.5.1).

This result suggests the need for a different source of semantic plausibility intuitions. Existing constraint-integration models have used human judgements (McRae et al., 1998, Narayanan and Jurafsky, 2002), which are however costly to elicit and thereby compromise the models' wide coverage. The SynSem-Integration model is the first to provide an independent, wide-coverage model of human plausibility intuitions that is derived from corpus data and requires no adaptations to process different phenomena.

The SynSem-Integration model proved able to reliably predict patterns of human difficulty in sentence processing. Its performance thus indicates that linking two sources of semantic plausibility and syntactic probability estimates by simple cost functions is sufficient to predict difficulty effects in human sentence processing. In contrast, existing constraint-integration models explicitly specify a large number of different constraints and use complex mechanisms for integration. Note that the SynSem-Integration model's syntactic and semantic models subsume many of the constraints usually posited by constraint-integration models, such as thematic fit, verb form or structural preferences. The SynSem-Integration model however does not require the definition of an individual hand-selected set of constraints and the manual setting of weights for each ambiguity phenomenon to be modelled. Instead,

its syntactic and semantic models are independently motivated as experience-based models of syntax and human plausibility intuitions, respectively, and the preferences relevant to processing different ambiguity phenomena simply fall out of these more general models. The use of general models of syntactic and semantic preferences ensures that in processing difficulty phenomena, all applicable constraints are always considered, and have consistent weights, which is difficult to ensure through manual constraint selection.

The SynSem-Integration model's predictions proved to be robust across three different combinations of cost functions. The cost functions based on probability ratios performed somewhat worse than those predicting fixed amounts of cost or cost based on a difference in syntactic and semantic rank. This is presumably because the probability ratio cost function is less robust to noise.

The probabilistic formulation of the SynSem-Integration model has two important consequences with regard to the model's predictions: One is the short tree bias of the syntactic parser model that overrides verb subcategorisation preferences in the NP/S and NP/0 ambiguities, allowing the model to make the correct prediction that humans initially prefer a direct object analysis even for verbs that generally prefer to take complement clauses. The other is the few role bias in the semantic model, which is not in accordance with the intuition that the human sentence processor tries to optimise the number of assigned roles to be able to interpret the input incrementally. We have applied normalisation strategies to avoid this bias, but these have in turn proven to dissuade the semantic model from preferring the intended interpretation after disambiguation for the NP/0 ambiguity.

Recall from the discussion in Section 2.5.3 that our claims about the cognitive reality of the architecture of the syntactic and semantic model are different. We do assume that the syntactic model's probabilistic basis reflects the human sentence processor's experience-based processing strategies. Against this background, we argue that the syntactic model's short tree bias reflects the human processor's preference for frequent, simple analyses that gave explanatory power to Frazier's (1978) Minimal Attachment principle.

The situation is different with the semantic model. While we have chosen an implementation based on linguistic experience, this choice was mostly one of practicability. It is much more plausible to assume that humans directly estimate plausibility from their experience in the real world than from the frequency of linguistic utterances about these experiences, even though we assume there is a correlation between the two (Section 2.5.1). Also, we have implemented a model that makes independence assumptions between the role assignments to different verb-argument pairs, due to data sparseness. This implementational choice causes the few role bias, which we only minimise, but not eliminate by normalisation procedures. However, there is no reason why events with many participants should be less plausible to people than events with few participants. Thus, our implementation and especially the independence

assumptions it makes between role assignments lead to a practicable approximation of the human reasoning system that should not be understood as making strong claims about the working of that system.

Another example of implementational choice that we do not construe as making strong claims about cognitive reality is the existence of two separate, feed-forward modules. We only assume that the human sentence processor constructs several analyses for the input in parallel, and that these analyses are semantically evaluated, so that both the syntactic and the semantic evaluations can be used to determine a globally preferred structure.

6.7.3. Summary

We have reviewed the performance of the SynSem-Integration model in an evaluation against human data and have found that the model clearly outperforms a syntactic parser baseline system. We have carried out an error analysis and have identified difficulties with one ambiguity phenomenon and two specific test sets, which are caused by the semantic model's preference for syntactic analyses that support a verb's argument slots to be filled over those that do not. We have shown that for all other data sets, the model bases its difficulty predictions on syntactic and semantic preferences that correspond to expected analyses of the experimental items. We have argued that where syntactic and semantic model's preferences deviate from the assumptions in the literature, the deviations capture noise and unexpected reactions to the data that are also present in the experimental results.

Overall, we have found that by combining preferences for syntactic analyses from two rich information sources by simple cost functions, the SynSem-Integration model achieves robustness, wide coverage and high reliability in the prediction of unseen human data, while not requiring the hand-selection of constraints and constraint weights. The poor performance of a lexicalised parser baseline model underscores the necessity of integrating a notion of semantic plausibility into our processing model.

Finally, we have again discussed the claims about the human sentence processing system that our probabilistic implementations of the syntactic and semantic model make. While we assume that the syntactic model mirrors the reliance of the human sentence processor on previous language experience, we consider the implementation of the semantic model a practicable approximation of human reasoning about plausibility that does not make strong claims about psychological reality.

7. Conclusions

In this thesis, we have identified four core desiderata for models of human sentence processing that are motivated by properties of the human sentence processor: *Wide coverage, incrementality, a probabilistic, experience-based architecture* and the *integration of semantic plausibility* on the level of verb-argument relations. The key contributions of this thesis are the proposal of the Syntax/Semantics (SynSem) Integration model, a sentence processing model that accounts for all four desiderata, and, as a precondition for this proposal, the introduction of a general model of human semantic intuitions about the plausibility of verb-argument relations that is used as the semantic component in the SynSem-Integration model.

This semantic model estimates the plausibility of a verb-argument pair connected by a specific thematic role by estimating the probability of seeing that triple in a corpus of annotated training data. It proceeds incrementally by verb-argument pairs, predicting thematic roles and plausibility estimates for individual pairs, and computes the plausibility of a syntactic structure on the basis of the plausibility estimates for the verb-argument pairs it contains. The semantic model in its naïve implementation faces a significant sparse-data problem, which we alleviate by combining two smoothing methods. In Chapter 3, we discussed how Good-Turing smoothing ensures verb-specific predictions about role preferences in case of an unseen argument, and how class-based smoothing exploits semantic generalisations to make more and better predictions that are specific to both the verb and the argument. The smoothed model's predictions are significantly correlated to human plausibility judgements for a variety of test sets (see Chapter 4). The semantic model achieves wide coverage of unseen data, but makes specific predictions only if the verb is known, because the verb defines the set of applicable thematic roles.

The existence of a general model of semantic plausibility allowed us to propose the SynSem-Integration model of human sentence processing. It integrates an incremental, probabilistic grammar-based syntactic model and the semantic model's preference predictions on the basis of the plausibility of the verb-argument relations in each syntactic structure. These plausibility predictions complement a syntactic probability of the structures in determining the globally preferred structure. The syntactic and semantic models also each identify a preferred structure based on their respective ranking of structures. Predictions of difficulty in human sentence processing are made by transparent cost functions defined over the preferences of both models in relation to the globally preferred structure. These predict difficulty whenever one model's

preferred structure conflicts with the globally preferred structure, and whenever the semantic interpretation of the globally preferred structure has to be revised and changes non-monotonically. Empirically, it has proven important that revision cost be assigned only if the new interpretation is less plausible than the old one (see Chapter 5).

The experience-based architecture of both models and the estimation of their parameters from corpora gives the SynSem-Integration model wide coverage both of structures that are processed effortlessly as well as those that cause disruption. This allows the model to cover a range of different psycholinguistic phenomena without requiring modifications. Finally, the SynSem-Integration model operates strictly incrementally, integrating each word into the syntactic representation immediately and making plausibility estimates as soon as a new verb-argument pair is encountered.

We have shown in Chapter 6 that the SynSem-Integration model successfully predicts difficulty in human sentence processing for four well-studied ambiguity phenomena. The model's success indicates that the simple and transparent combination of two preference rankings of the possible syntactic analyses of the input is enough to accurately predict the observed effects. This is relevant in comparison with constraint-based models, to which the SynSem-Integration model is most closely related. Both types of model combine preferences for different syntactic analyses that are computed based on different sources of information to arrive at a globally preferred analysis and to predict processing difficulty. However, unlike constraint-integration models, the SynSem-Integration model does not require the stipulation of individual sets of constraints for each construction type or the manual determination of weights. Instead, the parameters in the syntactic and semantic models are for the most part learnt automatically from corpus data, and remain constant across phenomena. Furthermore, the SynSem-Integration model's cost functions are a much simpler integration mechanism than those employed by constraint-integration approaches and similar models (e.g., Spivey and Tanenhaus, 1998, Narayanan and Jurafsky, 2002).

The SynSem-Integration model is also related to ranking parser models. It includes a syntactic parser model as one of its components and uses the concept of ranking structures according to their goodness. However, pure probabilistic parser models are limited to the prediction of syntactic preferences only, as demonstrated by the performance of the syntax-only baseline model, which could not predict the difficulty patterns in our test data: Even though we used a fully lexicalised probabilistic parser which had information about the frequency of head-head co-occurrences, for example of verbs and nouns, it did not predict plausibility effects. We have argued that this is due to the extreme sparseness of head-head co-occurrences in the available syntactically annotated training corpora. The SynSem-Integration model overcomes this limitation of pure probabilistic parser models by integrating an explicit source of semantic plausibility estimates, and thus is able to reliably predict the patterns of difficulty observed experimentally.

7.1. Future Work

A number of possible extensions and applications of the SynSem-Integration model and the semantic plausibility model suggest themselves for future work. One is the integration of additional sources of preferences that influence human sentence processing. This would allow an assessment of the scalability of the approach and would give the SynSem-Integration model wider coverage of human processing phenomena.

The most prominent candidate for this third information source is the influence of discourse context on the processing of ambiguities. Many studies have found that the processing especially of the PP Attachment and MC/RR ambiguities is influenced by preceding context. More specifically, these ambiguities appear to be preferentially resolved towards modifying an NP if this attachment disambiguates between several possible antecedents available in referential context (e.g., Altmann and Steedman, 1988, Spivey and Tanenhaus, 1998). A similar effect was found sentence-internally, depending on the definiteness of the NP to be modified (e.g., Crain and Steedman, 1985, Spivey-Knowlton and Sedivy, 1995). It is assumed that this preference is caused by differences in the amount of presuppositions that definite and indefinite modified NPs introduce and that have to be accommodated in the experimental null context.

Both these effects of existing or presupposed referential context can be captured by introducing a context representation and an evaluation mechanism that ranks input sentences by their acceptability given the preceding context. This ranking would prefer sentences which unambiguously define their referents and which force the accommodation of fewest presuppositions. The combination of this preference ranking from the discourse component with the preferences of the syntactic and semantic model would then serve to broaden the SynSem-Integration model's coverage of processing phenomena and to gauge the scalability of the model.

Results from Christianson, Hollingworth, Halliwell, and Ferreira (2001) suggest a second extension to the SynSem-Integration model. They show that after processing NP/0 sentences like *While Anna dressed, the baby lay on the bed*, readers often retain the impression that the sentences express the initially constructed direct object interpretation of the ambiguous NP, *Anna dressed the baby*. Simultaneously, they however also maintain the semantic interpretation of the disambiguated sentence. It therefore appears that humans do not always completely abandon their initial semantic interpretation of the input, even if it becomes untenable. The addition of an explicit memory component to the SynSem-Integration model would allow us to investigate the assumptions necessary to model this phenomenon. For example, slow decay of memory contents might allow ultimately incorrect interpretations to remain active enough at the end of the sentence to be recalled. Alternatively, an incomplete deletion of incorrect propositions during the update of the semantic interpretation might account for these effects.

A further strand of future work focuses on the semantic model. Here, one goal is to further improve its performance. We have employed class-based smoothing for verbs

and nouns in the semantic model, and have focused on optimising the verb classes. This strategy was clearly successful, in that the resulting model significantly predicts human intuitions both for seen and unseen verb-argument-role triples. However, the example of selectional preference models, which use only the WordNet noun hierarchy for smoothing, suggests that optimising the level of generalisation of the noun classes could yield another performance improvement. For example, we could adapt a reliably-performing selectional preference model to determine each noun's optimal class with regard to the verb before applying class-based smoothing. The method proposed by Clark and Weir (2002), for example, seems well-suited because it focuses on optimally using the WordNet hierarchy for smoothing. Automatic induction of word classes, as performed for the verbs, appears less promising because of the overall lower type frequencies of nouns in the training data.

Another goal is to investigate the question how much knowledge is needed to model human plausibility intuitions, which may lead to alternative formulations of the semantic model. Relying on corpus data with thematic role annotation, we have used a fairly knowledge-intensive approach for which relatively small amounts of training data exist. An alternative strategy would be to explore knowledge-lean models of lexical semantics, for example vector space models. Such models construct semantic representations using information about the distributions of words in context (Lund and Burgess, 1997, Landauer and Dumais, 1997). For knowledge-lean models, much larger training corpora are necessary for training, but since at most syntactic annotation is needed and annotation can be automatic, much more training data is also available. Vector space models have been shown repeatedly to predict relatedness effects between words, for example facilitated processing of words after a related word has been presented (*priming*). It is conceivable that vector space models can be used to distinguish between plausible and implausible pairs of words (see, e.g., Burgess and Lund, 1997, who show that the representations for the members of verb-argument pairs from studies that found a plausibility effect are more similar than those from studies that found no effect). However, it is not clear how vector space models can encode the relation between verb and noun, with respect to which plausibility must be judged. It is conceivable that models which also use the syntactic relations between words could allow a first step in that direction (e.g., Padó and Lapata, to appear).

Finally, the semantic model is of interest also for language processing approaches from computational linguistics, for example for lexicalised parsing models. Lexicalisation is used to allow lexically specific attachment decisions depending on a category's head, and if available, on co-occurrence information for example with the heads of sister categories. When comparing a lexicalised parser that uses such head-head dependencies with a lexicalised parser that does not in Chapter 5, we have however found that the two parsers performed almost indistinguishably, and in Chapter 6, we have seen that a fully lexicalised parser failed to predict the effect of thematic fit manipulations for which it could in principle account using head-head dependencies. Both

results indicate, in accordance with Gildea (2001), that head-head dependencies are not efficiently used in the lexicalised parser, due to data sparseness. In the semantic model, a similar sparseness problem is successfully addressed using class-based smoothing methods. If such smoothing models are applied in a lexicalised parser, they should increase the amount of head-head information available in lexicalised parsing. This in turn should increase performance by allowing parsers to make more specific, and in a sense more semantically motivated, attachment decisions.

Another domain of application for the semantic model in computational linguistics are tasks like selectional preference induction or thematic role assignment. We have shown that the FrameNet-trained semantic model outperforms selectional preference approaches on a number of test sets. It is especially interesting for these approaches that the model's reliance on verb classes allows accurate predictions despite the use of argument preferences from several verbs.

The semantic model also outperformed a standard role labeller on our test set. While the semantic model's performance on a standard role labelling test set, where syntactic features already allow accurate predictions, has not been tested, its plausibility predictions for verb-argument-role triples should certainly yield an interesting extension to standard role labelling systems. The machine-learning field of *ensemble learning* achieves improvements over the performance of single classifiers by combining systems with different *views* of the data and different error profiles (see Dietterich, 1999, for an introduction). Results from Erk and Padó (2005) suggest that semantic features may indeed provide such an independent view of the role-labelling problem beyond the view provided by the standard syntax-based features. This means that the combination of a standard role labeller and the preferred role predictions output by the semantic model could be profitable with regard to overall role labelling performance.

In sum, the SynSem-Integration model combines the wide-coverage, experience-based approach of current syntactic parsing accounts with a wide-coverage semantic model that evaluates the semantic plausibility of syntactic structures. As proposed in the constraint-based literature, the two information sources are integrated to determine a single preferred structure and to predict processing difficulty. The result is a wide-coverage, incremental, probabilistic model which explains the use of semantic and syntactic information in human sentence processing, as established by numerous experimental findings.

A. The Semantic Model: Training Data and Implementation

This appendix contains details relevant to the semantic model. Section A.1 describes the preparation of the training data and the extraction of features for verb class induction and model training. Section A.2 describes how verb classes were automatically induced from the training data to be used for class-based smoothing. Finally, Section A.3 contains the exact formulation of the backoff procedure to combine Good-Turing and class-based smoothing.

A.1. Training Data Preparation and Feature Extraction

This section gives details on how the training data for our models was prepared and how the information needed to estimate the model was extracted. From the annotated FrameNet and PropBank corpus data, we extracted information about which verbs, senses, roles, argument heads and grammatical functions co-occurred. Information about verb senses and roles are straightforward to extract from the annotations, but we need to describe in more detail how we arrive at the other features.

A.1.1. Verb and Argument Head Lemmas

The role-assigning verbs were extracted from the annotations and lemmatised using the MorphA tool (Minnen, Carroll, and Pearce, 2001). This rule-based tool gives reliable results when the correct part of speech of the input words is known, as was the case for the verbal predicates.

In order to extract a verb's co-occurrence with argument heads, we exploited the mapping of semantic role labels to syntactic phrases, from which we then extracted the head. For the PropBank data, manually assigned syntactic trees are available and role annotation is linked directly to phrases in the tree structure. The heads of the labelled arguments were determined heuristically, such that the first encountered verb is the preferred head of a verbal phrase or sentence, the first encountered adjective or adverb the head of an adverbial phrase etc. We defined the head of a prepositional phrase (PP) to be the head of the noun phrase to preserve as much of the semantic content of prepositional phrases as possible. Information about the PP status and preposition are

conserved in the grammatical function of the phrase (see below). In a noun phrase, the head is the last noun child.

The FrameNet annotation is made to character spans in the corpus sentences. To link this information to argument phrases, the corpus was automatically parsed with the Collins parser (Collins, 1997). We then extracted those phrasal nodes as argument phrases that spanned only the annotated words. If there was no correspondence of role span and syntactic phrase, due to parser errors or idiosyncrasies in the annotation, all nodes corresponding to the lexical items in the role span were extracted. The parser provides lexical heads for all constituents. Before extracting the argument heads, we lemmatised the corpora. Since only automatically assigned word class tags were available for the lexical items, we used the TreeTagger tool (Schmid, 1994), a standard lemmatisation and part-of-speech tagging tool trained on a large corpus of English.

A.1.2. Grammatical Functions

While both training corpora provide grammatical function annotation, the annotation style and the information contained in the grammatical function labels is very different. For example, the Penn Treebank function annotation for verbs' arguments focuses on marking phrases whose function cannot be immediately inferred. This means there are no tags for objects or sentence complements, but benefactive and dative PPs are marked, for example. The subject function is however annotated, and the NP in an agentive *by*-PP receives a tag for *logical subject*. In addition, there is a set of more semantic labels that usually attach to adverbials. These labels distinguish e.g. direction, location, manner and purpose.

On the other hand, the FrameNet function tagging annotates all arguments of a verb, but uses just three labels: *ext, obj* and *dep*. *Ext* marks arguments outside the maximal phrase headed by the verb. These are generally the syntactic subject, be it in a finite verb phrase, a governing verb construction or a passive clause. *Obj* marks direct objects of the verb, including those that are wh-extracted. Finally, *dep* marks adverbs, PPs and sentential complements, as well as the second object in double object constructions.

For reasons of model consistency, we wish to standardise the information contained in the grammatical function annotation. The intended role of grammatical functions in our model is to indicate the syntactic relation of the argument to the verb. For this end, the PropBank grammatical functions seem less adequate, since they do not apply to every argument of the verb and since they make quite fine-grained distinctions that may lead to sparse data in estimation. In addition, the functions assigned to adjuncts are semantic in nature and if specified would practically disambiguate role assignment. As a final technical consideration, it is easier to map more detailed annotation to coarser labels than the reverse. For these reasons, we will derive more reliable and useful annotations if we use FrameNet grammatical functions.

We make two adaptations to the FrameNet grammatical function labels to make

them more informative: First, we split the *ext* tag heuristically to account for passive constructions. We annotate NP subjects of verbs with a past participle POS tag that have no direct objects as *ext-pas* instead of *ext*. Second, for PPs, we add the preposition to the *dep* function to be able to distinguish between different kinds of PP dependents.

We map existing PropBank role and grammatical function annotation to FrameNet grammatical functions in the following way: We map the Penn Treebank subject tag to the *ext* function, which is subsequently split heuristically into *ext* and *ext-pas* as for the FrameNet data. Other NPs with PropBank role annotation are labelled as *obj*, with the second and subsequent objects receiving the label *dep*. All arguments with the PropBank *ArgM* label immediately receive the *dep* tag.

A.2. Inducing Verb Classes

This section describes in detail how verb classes were induced for class-based smoothing. Section A.2.1 describes the two clustering algorithms. In Section A.2.2, the selection of the optimal values for the three parameters *algorithm*, *smoothing within the clustering algorithm* and *number of clusters* is described.

A.2.1. Clustering Algorithms

We used an implementation of the Information Bottleneck (IB) and Information Distortion (ID) algorithms written by Zvika Marx (Marx, 2004). With both these soft clustering algorithms, clustering proceeds by dividing up one of the existing clusters at each time step. Over time, the clustering history forms a hierarchy of clusters, while at each time step, a new configuration of clusters is available. The clustering tool allows smoothing of the input features during the clustering process: Smoothing counts are added to the observed feature frequencies either uniformly or based on feature frequency. In practice, IB forms more even-sized clusters while ID converges faster.

Information Distortion Method

The Information Distortion (ID) method (Gedeon et al., 2003) takes its name from the idea of deriving a more compact, distorted representation of the input data that optimally represents the information contained in the input in the form of clusters and cluster membership. It minimises a cost term with two parts to arrive at the optimal clustering. One the one hand, it maximises the conditional entropy of clusters given verbs, thus discarding input information. In order to retain the relevant information in the input, it minimises the conditional entropy of features given clusters under the assumption that the features are related to the verbs and carry information about them. The two constraints are combined by a weighting factor β. The cost term that has

to be minimised during clustering thus is as shown in Equation A.1, if conditional independence between clusters and features given the data is assumed.

$$L = -H(Clusters|Verbs) + \beta H(Features|Clusters) \tag{A.1}$$

The size of β determines whether few or many clusters may form. For $\beta = 0$, there is just one cluster, which splits up again and again as β increases.

Information Bottleneck Method

The Information Bottleneck (IB) method (Tishby et al., 1999) is closely related to the Information Distortion method. It minimises a similar cost term, which is defined in terms of mutual information rather than conditional entropy, as shown in Equation A.2:

$$L = I(Clusters; Verbs) - \beta I(Features; Clusters) \tag{A.2}$$

Expressed in terms of conditional entropy and neglecting a constant factor, this equation becomes Equation A.3

$$L = (H(Clusters) - H(Clusters|Verbs)) + \beta H(Features|Clusters) \tag{A.3}$$

since $I(Clusters; Verbs) = H(Clusters) - H(Clusters|Verbs)$. The difference between the algorithms is therefore that the IB method uses the the entropy of the cluster distribution as a prior.

A.2.2. Parameter Setting

We have three parameters to explore for inducing verb classes: The clustering algorithm, the amount of smoothing applied within the clustering algorithms and the number of clusters within each clustering run. The evaluation of the clustering results is task-based: We choose the clustering configuration that produces optimal results in the judgement prediction task (see Section 3.2.1) on the development set. For each experiment, the same corpus was used to induce verb classes and to estimate the model terms.

Selecting Algorithm and Intra-Algorithm Smoothing We combined each of the algorithms with six levels of smoothing. We varied the intra-algorithm smoothing methods, namely assignment of smoothing counts by frequency, or uniform assignment, and three levels of intensity. We report the number of clustering runs (out of six) for each algorithm in which at least one cluster configuration allowed the model to make predictions that were significantly correlated to human judgements on the development set. Table A.1 gives a first indication that verb clusters formed from the FrameNet data

	All		NP	
Corpus	ID	IB	ID	IB
FN	4	1	–	–
PB	0	0	3	4

Table A.1.: Number of clustering configurations that allow significant correlations with human data for the two clustering algorithms (out of 6), for the FrameNet and PropBank training corpora.

work better for smoothing than PropBank clusters: For the FrameNet corpus, there was a significant correlation to human data for four ID algorithm runs. For the IB algorithm, one combination led to significant correlations.

On the PropBank training corpus, no significant correlations were found at all when features for all arguments were used as input to clustering. We also experimented with using just NP arguments, since semantic annotation in PropBank does not allow reliable generalisations about underlying role meaning beyond the *Arg1* role. By restricting the input to just NP arguments, we exclude many arguments that may lead the clusterer to making incorrect generalisations. Reducing the amount of information available to the clusterer did prove successful in that it produced a total of seven combinations of clustering algorithm and a smoothing level that led to significant correlations.

Additionally, we observed that for the PropBank data, one very large cluster and several small ones were formed. The large cluster overgeneralises semantic alternatives and made class-based smoothing ineffective. Therefore, we disregard the largest cluster for the PropBank data. No such effect was observed with the FrameNet clusters.

From the set of algorithm-smoothing combinations in Table A.1, we selected parametrisations for final evaluation based on the notion of *stability*. Recall that both algorithms increase the number of clusters by one at each iteration and that each parametrisation yields a series of cluster configurations as the number of iterations increases. We chose those parametrisations where a *stable* series of at least three consecutive cluster configurations returned significant correlations on the development set. This should be an indication of a generalisable success, rather than a fluke caused by peculiarities of the data.

We chose several parametrisations for both training corpora, because we want to be able to test whether evaluation results generalise to more than one algorithm-smoothing combination. We selected three parametrisations for each training corpus, choosing parametrisations with long series of significant results first. For FrameNet, two stable series were tied, so we chose the configuration that starts at a higher number of classes (assuming that more classes mean more fine-grained semantic distinctions).

Corpus	Verb Classes	Algorithm	Smoothing	No of Clusters
PB	PB 1	ID	ms 200	3
	PB 2	IB	s 300	11
	PB 3	IB	ms 200	7
FN	FN 1	ID	ms 100	11
	FN 2	ID	s 200	13
	FN 3	ID	s 300	13

Table A.2.: Selected clustering configurations for both training corpora. Reference number in thesis, algorithm, smoothing options and number of clusters.

The rejected IB series starts at three classes, while the chosen ID series starts at 13 classes.

Setting the Number of Clusters Having selected stable combinations of clustering algorithm and smoothing, the last remaining parameter to be set is the number of clusters to be used. Out of a stable series of configurations, we use the configuration that returned the first significant result, as this is the most general grouping of verbs. Any following configurations only make additional splits.

An overview of the stable configurations we use in all evaluations below is given in Table 3.3, with the smoothing parameters for the clustering algorithms and the number of clusters they contain. The identification string assigned to each set of verb classes is used in the body of the thesis to refer to the configuration. The row with intra-algorithm smoothing information contains the parameter switches for the clustering algorithm. *ms* signifies frequency-based smoothing, and *s* signifies uniform assignment of smoothing. The lower the smoothing indicator, the fewer additional counts were assigned to the input data.

A.3. Combining Good-Turing and Class-Based Smoothing

This section describes in detail how we combine GT and class-based smoothing in our final semantic plausibility model. We use pure GT smoothing for three of the four model terms. The sparsest $P(a|v_s, gf, r)$ term is first smoothed with class-based smoothing, and if that fails, we back off to a GT estimate. We use Katz' Backoff, a special case of linear interpolation (see Section 3.3.2), to combine the different smoothing strategies.

We back off in three phases: If $f(class_{arg}, class_{verb}, gf, role) > 0$, we use the class-based estimate $P(class_{arg}|class_{verb}, gf, role)$. Else, if $f(class_{arg}, class_{verb}, role) > 0$,

we drop the grammatical function information from estimation and attempt to use $P(class_{arg}|class_{verb}, role)$. As the treatment of sparse specified grammatical functions discussed in Section 3.1.4, this step assumes that the syntactic realisation of a verb-argument pair is flexible, and that dropping the grammatical function from the estimation altogether may yield useful information. As a last resort, we assign the GT estimate for unseen events from the second backoff distribution.

To ensure that the backoff procedure returns a probability distribution, we need to discount each backoff distribution so that probability mass is left over for the next backoff step, and we need to scale the output of the next backoff step so that it only takes up the left-over probability mass. We first use the GT approach to determine the amount of probability mass that each backoff distribution distribution should reserve for unseen events, and to discount it accordingly, so that probability mass is indeed freed. The probability mass reserved by the second backoff distribution $P(class_{arg}|class_{verb}, role)$ is split equally among all possible unseen events and assigned in the third backoff step.

Finally, the output of the second backoff distribution has to be scaled so that it only takes up the probability mass left over by the first distribution. The total probability mass taken up by the third backoff step by definition is equal to the mass left over by step two, so no additional scaling is necessary there.

The scaling factor α for the second backoff distribution apportions the probability mass assigned to unseen events in step 1 fairly over the probability mass we expect step 2 to assign to these events (see e.g., Dagan, Pereira, and Lee, 1994). Since we are dealing with conditional probability distributions in our backoff steps, the scaling factor depends on the conditioning events $class_{verb}$ and $role$.

The factor is computed as shown in Equation A.4: All probability mass that backoff distribution 1 assigns to seen events (those where the backoff distribution does not assign a zero probability) is summed up. Subtracting this sum from 1 results in the probability mass reserved for unseen events. This formulation is chosen because it is generally easier to sum over seen events than over (a potentially large number of) unseen events. The mass reserved for unseen events is then divided by the mass that backoff distribution 2 assigns to the events that are unseen in distribution 1. This mass is computed as 1 minus the sum of all probabilities that backoff distribution 2 assigns to the events covered by distribution 1.

$$\alpha_{class_{verb},role} = \frac{1 - \sum_{class_{arg}:P_{bo1}(class_{arg}|class_{verb},gf,role)>0} P_{bo1}(class_{arg}|class_{verb}, gf, role)}{1 - \sum_{class_{arg}:P_{bo1}(class_{arg}|class_{verb},gf,role)>0} P_{bo2}(class_{arg}|class_{verb}, role)} \tag{A.4}$$

Combining the discounted probability distributions by this backoff factor ensures that the semantic model outputs a probability distribution for the $P(class_{arg}|class_v, r, gf)$ term and cleanly combines class-based smoothing with a GT estimate for cases that are still assigned zero probability after smoothing was applied.

B. Plausibility Rating Materials

This Appendix contains the verb-argument pairs rated in our study, with FrameNet and PropBank thematic role annotation and ratings. Each verb is listed once with its FrameNet fillers and once with its PropBank fillers. The seen relation between verb and argument is presented first.

Ratings for the role commonly assigned to the syntactic subject position were elicited by asking *How common is it for an Argument to Verb?*, while ratings for the role commonly assigned to a syntactic object were elicited by asking *How common is it for an Argument to be Verbed?*. If necessary, *something* was added to the subject elicitation, for example for the verb *tell*. Ratings were on a scale of $1-7$, with 1 meaning *very uncommon* and 7 meaning *very common*.

We chose the verbs according to the roles they assign in VerbNet. The first six verbs in each table assign an *experiencer* role, the next six assign a *recipient* role and the last six assign a *patient* role.

Table B.1.: Verbs and FrameNet arguments with FrameNet and PropBank role and rating

Verb	Argument	FN Role	PB Role	Rating
resent	woman	Experiencer	Arg0	5.6
resent	woman	Content	Arg1	4.9
resent	group	Experiencer	Arg0	5.2
resent	group	Content	Arg1	5.2
resent	individual	Experiencer	Arg0	6.0
resent	individual	Content	Arg1	5.6
resent	presence	Content	Arg1	1.4
resent	presence	Experiencer	Arg0	5.0
resent	contribution	Content	Arg1	1.2
resent	contribution	Experiencer	Arg0	3.5
resent	intrusion	Content	Arg1	1.2
resent	intrusion	Experiencer	Arg0	6.1
hear	girl	Perceiver_passive	Arg0	6.8
hear	girl	Phenomenon	Arg1	6.5
hear	man	Perceiver_passive	Arg0	6.7
hear	man	Phenomenon	Arg1	6.5
hear	ear	Perceiver_passive	Arg0	6.8
hear	ear	Phenomenon	Arg1	1.1

Table B.1.: Verbs and FrameNet arguments (continued)

Verb	Argument	FN Role	PB Role	Rating
hear	sound	Phenomenon	Arg1	6.8
hear	sound	Perceiver_passive	Arg0	1.2
hear	voice	Phenomenon	Arg1	6.5
hear	voice	Perceiver_passive	Arg0	3.5
hear	knock	Phenomenon	Arg1	6.2
hear	knock	Perceiver_passive	Arg0	1.1
see	friend	Perceiver_passive	Arg0	6.7
see	friend	Phenomenon	Arg1	6.6
see	viewer	Perceiver_passive	Arg0	6.8
see	viewer	Phenomenon	Arg1	3.8
see	pupil	Perceiver_passive	Arg0	6.6
see	pupil	Phenomenon	Arg1	6.3
see	name	Phenomenon	Arg1	6.0
see	name	Perceiver_passive	Arg0	1.0
see	movement	Phenomenon	Arg1	6.0
see	movement	Perceiver_passive	Arg0	1.2
see	face	Phenomenon	Arg1	6.8
see	face	Perceiver_passive	Arg0	2.4
encourage	government	Speaker	Arg0	4.2
encourage	government	Addressee	Arg1	3.8
encourage	vicar	Speaker	Arg0	5.6
encourage	vicar	Addressee	Arg1	4.7
encourage	affiliate	Speaker	Arg0	3.7
encourage	affiliate	Addressee	Arg1	4.9
encourage	pupil	Addressee	Arg1	5.9
encourage	pupil	Speaker	Arg0	3.9
encourage	boy	Addressee	Arg1	5.7
encourage	boy	Speaker	Arg0	5.2
encourage	technician	Addressee	Arg1	4.8
encourage	technician	Speaker	Arg0	4.0
embarrass	government	Experiencer	Arg1	4.9
embarrass	government	Stimulus	Arg0	3.8
embarrass	executive	Experiencer	Arg1	4.9
embarrass	executive	Stimulus	Arg0	4.8
embarrass	intervener	Experiencer	Arg1	2.9
embarrass	intervener	Stimulus	Arg0	4.1
embarrass	importunity	Stimulus	Arg0	2.4
embarrass	importunity	Experiencer	Arg1	1.2
embarrass	book-keeping	Stimulus	Arg0	2.6
embarrass	book-keeping	Experiencer	Arg1	1.3

Table B.1.: Verbs and FrameNet arguments (continued)

Verb	Argument	FN Role	PB Role	Rating
embarrass	lyric	Stimulus	Arg0	3.6
embarrass	lyric	Experiencer	Arg1	1.2
confuse	baby	Experiencer	Arg1	6.0
confuse	baby	Stimulus	Arg0	3.7
confuse	sense	Experiencer	Arg1	3.8
confuse	sense	Stimulus	Arg0	4.0
confuse	computer	Experiencer	Arg1	4.1
confuse	computer	Stimulus	Arg0	5.4
confuse	comment	Stimulus	Arg0	5.8
confuse	comment	Experiencer	Arg1	1.5
confuse	boatman	Stimulus	Arg0	2.1
confuse	boatman	Experiencer	Arg1	3.2
confuse	equipment	Stimulus	Arg0	4.9
confuse	equipment	Experiencer	Arg1	1.4
promise	parent	Addressee	Arg1	5.8
promise	parent	Speaker	Arg0	6.5
promise	customer	Addressee	Arg1	6.4
promise	customer	Speaker	Arg0	3.6
promise	company	Addressee	Arg1	5.8
promise	company	Speaker	Arg0	5.2
promise	government	Speaker	Arg0	6.4
promise	government	Addressee	Arg1	5.0
promise	administration	Speaker	Arg0	6.0
promise	administration	Addressee	Arg1	4.9
promise	sun-god	Speaker	Arg0	1.2
promise	sun-god	Addressee	Arg1	2.7
advise	customer	Addressee	Arg1	6.0
advise	customer	Speaker	Arg0	3.8
advise	designer-gardener	Addressee	Arg1	3.8
advise	designer-gardener	Speaker	Arg0	5.8
advise	biologist	Addressee	Arg1	2.4
advise	biologist	Speaker	Arg0	5.0
advise	official	Speaker	Arg0	6.2
advise	official	Addressee	Arg1	5.9
advise	doctor	Speaker	Arg0	6.8
advise	doctor	Addressee	Arg1	4.0
advise	expert	Speaker	Arg0	6.5
advise	expert	Addressee	Arg1	3.8
inform	public	Addressee	Arg1	5.9
inform	public	Speaker	Arg0	3.8

Table B.1.: Verbs and FrameNet arguments (continued)

Verb	Argument	FN Role	PB Role	Rating
inform	employee	Addressee	Arg1	5.6
inform	employee	Speaker	Arg0	5.2
inform	police	Addressee	Arg1	6.0
inform	police	Speaker[1]	Arg0	5.8
inform	secretary	Speaker	Arg0	6.2
inform	secretary	Addressee	Arg1	5.8
inform	center	Speaker	Arg0	4.9
inform	center	Addressee	Arg1	4.0
caution	friend	Addressee	Arg2	5.0
caution	friend	Speaker	Arg0	5.6
caution	woman	Addressee	Arg2	5.7
caution	woman	Speaker	Arg0	5.7
caution	judge	Addressee	Arg2	3.7
caution	judge	Speaker	Arg0	5.4
caution	police	Speaker	Arg0	6.6
caution	police	Addressee	Arg2	2.2
caution	lady	Speaker	Arg0	5.0
caution	lady	Addressee	Arg2	5.0
caution	rain	Speaker	Arg0	1.2
caution	rain	Addressee	Arg2	1.3
ask	doctor	Addressee	Arg2	6.7
ask	doctor	Speaker	Arg0	6.5
ask	police	Addressee	Arg2	6.2
ask	police	Speaker[1]	Arg0	6.5
ask	state	Addressee	Arg2	3.8
ask	state	Speaker	Arg0	3.8
ask	prosecutor	Speaker	Arg0	6.6
ask	prosecutor	Addressee	Arg2	6.3
ask	charity	Speaker	Arg0	5.9
ask	charity	Addressee	Arg2	5.2
tell	reporter	Addressee	Arg2	6.7
tell	reporter	Speaker	Arg0	6.7
tell	therapist	Addressee	Arg2	6.7
tell	therapist	Speaker	Arg0	6.9
tell	court	Addressee	Arg2	6.6
tell	court	Speaker	Arg0	6.8

[1]This verb-argument pair was seen in both the Addressee and Speaker relation.

Table B.1.: Verbs and FrameNet arguments (continued)

Verb	Argument	FN Role	PB Role	Rating
tell	department	Speaker	Arg0	5.3
tell	department	Addressee	Arg2	5.0
tell	senator	Speaker	Arg0	6.1
tell	senator	Addressee	Arg2	6.3
tell	patient	Speaker	Arg0	5.9
tell	patient	Addressee	Arg2	6.4
hit	opponent	Victim	Arg1	5.7
hit	opponent	Agent	Arg0	5.3
hit	creature	Victim	Arg1	5.0
hit	creature	Agent	Arg0	5.0
hit	baby	Victim	Arg1	3.6
hit	baby	Agent	Arg0	6.2
hit	man	Agent	Arg0	5.6
hit	man	Victim	Arg1	4.1
hit	mummy	Agent	Arg0	2.3
hit	mummy	Victim	Arg1	2.2
hit	brother	Agent	Arg0	4.7
hit	brother	Victim	Arg1	4.1
increase	amount	Attribute	Arg1	5.5
increase	amount	Cause	Arg0	2.4
increase	rate	Item	Arg1	5.8
increase	rate	Cause	Arg0	4.7
increase	number	Attribute	Arg1	5.5
increase	number	Cause	Arg0	1.6
increase	authority	Agent	Arg0	6.0
increase	authority	Item	Arg1	2.9
increase	industry	Agent	Arg0	5.7
increase	industry	Item	Arg1	3.0
increase	model	Cause	Arg0	2.6
increase	model	Item	Arg1	3.5
kill	man	Victim	Arg1	5.4
kill	man	Killer	Arg0	3.4
kill	mother	Victim	Arg1	3.8
kill	mother	Killer	Arg0	2.1
kill	lion	Victim	Arg1	4.9
kill	lion	Killer	Arg0	2.7
kill	girl	Killer	Arg0	2.2
kill	girl	Victim	Arg1	4.1
kill	rebel	Killer	Arg0	4.5
kill	rebel	Victim	Arg1	5.0

Table B.1.: Verbs and FrameNet arguments (continued)

Verb	Argument	FN Role	PB Role	Rating
kill	shark	Killer	Arg0	2.1
kill	shark	Victim	Arg1	3.9
eliminate	club	Theme	Arg1	2.8
eliminate	club	Agent	Arg0	3.6
eliminate	barrier	Theme	Arg1	4.9
eliminate	barrier	Agent	Arg0	3.9
eliminate	need	Theme	Arg1	5.0
eliminate	need	Agent	Arg0	2.5
eliminate	intruder	Agent	Arg0	3.3
eliminate	intruder	Theme	Arg1	2.8
eliminate	law	Agent	Arg0	4.6
eliminate	law	Theme	Arg1	3.0
eliminate	therapy	Agent	Arg0	4.7
eliminate	therapy	Theme	Arg1	2.7
raise	price	Attribute	Arg1	6.0
raise	price	Cause	Arg0	1.7
raise	rate	Attribute	Arg1	5.7
raise	rate	Cause	Arg0	1.8
raise	dividend	Item	Arg1	4.7
raise	dividend	Cause	Arg0	1.6
raise	government	Agent	Arg0	6.0
raise	government	Item	Arg1	1.2
raise	bank	Agent	Arg0	6.0
raise	bank	Item	Arg1	1.8
raise	country	Agent	Arg0	5.2
raise	country	Item	Arg1	1.8
eat	meal	Ingestibles	Arg1	6.9
eat	meal	Ingestor	Arg0	1.9
eat	lunch	Ingestibles	Arg1	6.9
eat	lunch	Ingestor	Arg0	1.1
eat	egg	Ingestibles	Arg1	6.4
eat	egg	Ingestor	Arg0	1.0
eat	villager	Ingestor	Arg0	6.8
eat	villager	Ingestibles	Arg1	1.7
eat	local	Ingestor	Arg0	6.7
eat	local	Ingestibles	Arg1	1.5
eat	group	Ingestor	Arg0	6.0
eat	group	Ingestibles	Arg1	1.1

Table B.2.: Verbs and PropBank arguments with FrameNet and PropBank role and rating

Verb	Argument	FN Role	PB Role	Rating
resent	wife	Experiencer	Arg0	5.8
resent	wife	Content	Arg1	5.0
resent	viewer	Experiencer	Arg0	4.0
resent	viewer	Content	Arg1	2.7
resent	firm	Experiencer	Arg0	2.7
resent	firm	Content	Arg1	5.2
resent	cost	Content	Arg1	1.1
resent	cost	Experiencer	Arg0	5.6
resent	transfer	Content	Arg1	1.3
resent	transfer	Experiencer	Arg0	3.9
resent	product	Content	Arg1	1.2
resent	product	Experiencer	Arg0	4.2
hear	court	Hearer	Arg0	6.4
hear	court	Message	Arg1	5.2
hear	board	Hearer	Arg0	5.4
hear	board	Message	Arg1	2.8
hear	committee	Hearer	Arg0	5.8
hear	committee	Message	Arg1	4.1
hear	case	Message	Arg1	5.8
hear	case	Hearer	Arg0	1.4
hear	appeal	Message	Arg1	6.4
hear	appeal	Hearer	Arg0	1.7
hear	moan	Phenomenon	Arg1	5.8
hear	moan	Perceiver_passive	Arg0	1.1
see	return	Phenomenon	Arg1	5.0
see	return	Perceiver_passive	Arg	1.2
see	effect	Phenomenon	Arg1	6.0
see	effect	Perceiver_passive	Arg0	1.3
see	drop	Phenomenon	Arg1	3.8
see	drop	Perceiver_passive	Arg0	1.0
see	executive	Perceiver_passive	Arg0	6.8
see	executive	Phenomenon	Arg1	6.0
see	analyst	Perceiver_passive	Arg0	6.5
see	analyst	Phenomenon	Arg1	5.4
see	investor	Perceiver_passive	Arg0	5.9
see	investor	Phenomenon	Arg1	4.2
encourage	executive	Speaker	Arg0	5.0
encourage	executive	Addressee	Arg1	5.6
encourage	uncertainty	Speaker	Arg0	2.5
encourage	uncertainty	Addressee	Arg1	5.5

Table B.2.: Verbs and PropBank arguments (continued)

Verb	Argument	FN Role	PB Role	Rating
encourage	purpose	Speaker	Arg0	3.8
encourage	purpose	Addressee	Arg1	3.6
encourage	investor	Addressee	Arg1	5.5
encourage	investor	Speaker	Arg0	5.1
encourage	company	Addressee	Arg1	4.0
encourage	company	Speaker	Arg0	4.0
encourage	client	Addressee	Arg1	5.1
encourage	client	Speaker	Arg0	3.7
embarrass	conservative	Experiencer	Arg1	3.7
embarrass	conservative	Stimulus	Arg0	3.9
embarrass	board	Experiencer	Arg1	3.9
embarrass	board	Stimulus	Arg0	3.8
embarrass	official	Experiencer	Arg1	5.0
embarrass	official	Stimulus	Arg0	4.9
embarrass	treatment	Stimulus	Arg0	4.0
embarrass	treatment	Experiencer	Arg1	1.0
embarrass	information	Stimulus	Arg0	4.8
embarrass	information	Experiencer	Arg1	1.2
embarrass	revelation	Stimulus	Arg0	5.3
embarrass	revelation	Experiencer	Arg1	1.2
confuse	shareholder	Experiencer	Arg1	5.0
confuse	shareholder	Stimulus	Arg0	3.3
confuse	community	Experiencer	Arg1	4.9
confuse	community	Stimulus	Arg0	3.5
confuse	situation	Experiencer	Arg1	5.1
confuse	situation	Stimulus	Arg0	6.0
confuse	insistence	Stimulus	Arg0	4.2
confuse	insistence	Experiencer	Arg1	1.4
confuse	statement	Stimulus	Arg0	6.0
confuse	statement	Experiencer	Arg1	2.5
confuse	condition	Stimulus	Arg0	4.6
confuse	condition	Experiencer	Arg1	2.5
promise	state	Addressee	Arg1	3.9
promise	state	Speaker	Arg0	5.6
promise	station	Addressee	Arg1	2.2
promise	station	Speaker	Arg0	1.9
promise	foundation	Addressee	Arg1	5.1
promise	foundation	Speaker	Arg0	4.9
promise	company	Speaker	Arg0	5.8
promise	company	Addressee	Arg1	5.2

Table B.2.: Verbs and PropBank arguments (continued)

Verb	Argument	FN Role	PB Role	Rating
promise	plan	Speaker	Arg0	5.1
promise	plan	Addressee	Arg1	1.2
promise	fund	Speaker	Arg0	3.8
promise	fund	Addressee	Arg1	1.9
advise	client	Addressee	Arg1	6.6
advise	client	Speaker	Arg0	3.7
advise	business	Addressee	Arg1	5.8
advise	business	Speaker	Arg0	5.3
advise	investor	Addressee	Arg1	6.3
advise	investor	Speaker	Arg0	6.0
advise	hospital	Speaker	Arg0	6.0
advise	hospital	Addressee	Arg1	4.0
advise	planner	Speaker	Arg0	6.3
advise	planner	Addressee	Arg1	3.9
advise	banker	Speaker	Arg0	6.0
advise	banker	Addressee	Arg1	5.0
inform	committee	Addressee	Arg1	5.3
inform	committee	Speaker	Arg0	5.9
inform	reader	Addressee	Arg1	5.9
inform	reader	Speaker	Arg0	4.1
inform	network	Addressee	Arg1	2.6
inform	network	Speaker	Arg0	4.4
inform	system	Speaker	Arg0	4.6
inform	system	Addressee	Arg1	1.9
inform	administration	Speaker	Arg0	6.0
inform	administration	Addressee	Arg1	5.6
inform	foundation	Speaker	Arg0	4.2
inform	foundation	Addressee	Arg1	3.4
caution	bank	Addressee	Arg2	2.3
caution	bank	Speaker	Arg0	4.2
caution	leader	Addressee	Arg2	5.0
caution	leader	Speaker	Arg0	5.9
caution	user	Addressee	Arg2	5.4
caution	user	Speaker	Arg0	3.8
caution	developer	Speaker	Arg0	3.8
caution	developer	Addressee	Arg2	5.0
caution	professional	Speaker	Arg0	5.0
caution	professional	Addressee	Arg2	5.1
caution	trader	Speaker	Arg0	3.2
caution	trader	Addressee	Arg2	5.1

Table B.2.: Verbs and PropBank arguments (continued)

Verb	Argument	FN Role	PB Role	Rating
ask	court	Addressee	Arg2	5.8
ask	court	Speaker	Arg0	6.0
ask	department	Addressee	Arg2	5.2
ask	department	Speaker	Arg0	4.0
ask	congress	Addressee	Arg2	6.0
ask	congress	Speaker	Arg0	5.8
ask	official	Speaker	Arg0	6.1
ask	official	Addressee	Arg2	6.6
ask	union	Speaker	Arg0	6.0
ask	union	Addressee	Arg2	5.3
ask	firm	Speaker	Arg0	5.3
ask	firm	Addressee	Arg2	5.1
tell	reporter	Addressee	Arg2	6.7
tell	reporter	Speaker	Arg0	6.7
tell	analyst	Addressee	Arg2	6.0
tell	analyst	Speaker	Arg0	6.5
tell	investor	Addressee	Arg2	6.0
tell	investor	Speaker	Arg0	6.0
tell	chairman	Speaker	Arg0	6.6
tell	chairman	Addressee	Arg2	6.2
tell	officer	Speaker	Arg0	6.3
tell	officer	Addressee	Arg2	6.0
tell	executive	Speaker	Arg0	6.6
tell	executive	Addressee	Arg2	5.7
hit	stock	Impactee	Arg2	2.7
hit	stock	Impactor	Arg1	2.0
hit	market	Impactee	Arg2	3.7
hit	market	Impactor	Arg1	1.7
hit	ball	Impactee	Arg2	6.1
hit	ball	Impactor	Arg1	5.7
hit	quake	Impactor	Arg2	3.4
hit	quake	Impactee	Arg1	1.2
hit	player	Agent	Arg0	5.6
hit	player	Victim	Arg1	5.7
hit	earthquake	Impactor	Arg2	2.7
hit	earthquake	Impactee	Arg1	1.3
increase	sale	Item	Arg1	2.7
increase	sale	Cause	Arg0	5.6

Table B.2.: Verbs and PropBank arguments (continued)

Verb	Argument	FN Role	PB Role	Rating
increase	rate	Attribute	Arg1	5.8
increase	rate	Cause	Arg0	4.7
increase	revenue	Item	Arg1	5.8
increase	revenue	Cause	Arg0	4.0
increase	company	Agent	Arg0	5.2
increase	company	Item	Arg1	2.6
increase	bank	Agent	Arg0	6.0
increase	bank	Item	Arg1	2.4
increase	move	Cause	Arg0	4.0
increase	move	Item	Arg1	2.0
kill	item	Victim	Arg1	1.3
kill	item	Cause	Arg0	2.7
kill	people	Victim	Arg1	6.2
kill	people	Killer	Arg0	3.0
kill	cell	Victim	Arg1	5.7
kill	cell	Cause	Arg0	2.8
kill	antibody	Cause	Arg0	5.9
kill	antibody	Victim	Arg1	4.1
kill	group	Killer	Arg0	3.4
kill	group	Victim	Arg1	3.0
kill	house	Cause	Arg0	1.7
kill	house	Victim	Arg1	1.0
eliminate	job	Theme	Arg1	5.0
eliminate	job	Agent	Arg0	3.0
eliminate	barrier	Theme	Arg1	4.9
eliminate	barrier	Agent	Arg0	3.9
eliminate	need	Theme	Arg1	5.0
eliminate	need	Agent	Arg0	2.5
eliminate	policy	Agent	Arg0	4.2
eliminate	policy	Theme	Arg1	4.2
eliminate	act	Agent	Arg0	4.9
eliminate	act	Theme	Arg1	2.4
eliminate	provision	Agent	Arg0	3.1
eliminate	provision	Theme	Arg1	4.1
raise	question	None1	Arg1	6.4
raise	question	None2	Arg0	4.2
raise	rate	Attribute	Arg1	5.7
raise	rate	Cause	Arg0	1.8
raise	stake	Item	Arg1	5.2
raise	stake	Cause	Arg0	2.1

Table B.2.: Verbs and PropBank arguments (continued)

Verb	Argument	FN Role	PB Role	Rating
raise	congress	Agent	Arg0	5.3
raise	congress	Item	Arg1	1.2
raise	bank	Agent	Arg0	6.0
raise	bank	Item	Arg1	1.8
raise	firm	Agent	Arg0	5.3
raise	firm	Item	Arg1	2.0
eat	apple	Ingestibles	Arg1	6.5
eat	apple	Ingestor	Arg0	1.0
eat	debt	Ingestibles	Arg1	1.4
eat	debt	Ingestor	Arg0	1.3
eat	pizza	Ingestibles	Arg1	6.8
eat	pizza	Ingestor	Arg0	1.1
eat	people	Ingestor	Arg0	6.9
eat	people	Ingestibles	Arg1	1.5
eat	husband	Ingestor	Arg0	6.7
eat	husband	Ingestibles	Arg1	1.3
eat	cost	Ingestor	Arg0	1.0
eat	cost	Ingestibles	Arg1	1.8

Bibliography

Steven Abney and Mark Johnson. Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*, 20(3):233–250, 1991.

Steven Abney and Marc Light. Hiding a semantic hierarchy in a Markov model. In *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing at the Annual Meeting of the Association for Computational Linguistics (ACL)*, 1999.

Beverly C. Adams, Charles Clifton, and Don Mitchell. Lexical guidance in sentence processing? *Psychonomic Bulletin and Review*, 5(2):265–270, 1998.

Jean Aitchison. *Words in the Mind: An Introduction to the Mental Lexicon*. Basil Blackwell, Oxford and New York, third edition, 2003.

Gerry Altmann and Mark Steedman. Interaction with context during human sentence processing. *Cognition*, 30:191–238, 1988.

Elizabeth Bates and Brian McWhinney. Functionalism and the competition model. In B. McWhinney and E. Bates, editors, *The crosslinguistic study of sentence processing*, pages 157–193. Cambridge University Press, 1989.

Katherine S. Binder. The effects of thematic fit and discourse context on syntactic ambiguity resolution. *Journal of Memory and Language*, 44:297–324, 2001.

Julie Boland. The relationship between syntactic and semantic processes in sentence comprehension. *Language and Cognitive Processes*, 12(4):423–484, 1997.

Julie Boland and Allison Blodgett. Argument status and PP-Attachment. *Journal of Psycholinguistic Research*, 35:385–403, 2006.

Thorsten Brants and Matthew W. Crocker. Probabilistic parsing and psycholinguistic plausibility. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, 2000.

M. Anne Britt. The interaction of referential ambiguity and argument structure in the parsing of prepositional phrases. *Journal of Memory and Language*, 33:251–283, 1994.

Carsten Brockmann. Evaluating and combining approaches to selectional preference acquisition. In *Proceedings of the Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, Budapest, 2003.

Aljoscha Burchardt, Katrin Erk, Andrea Kowalski, and Sebastian Pado. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*, 2006.

Curt Burgess. *Interaction of semantic, syntactic and visual factors in syntactic ambiguity resolution*. PhD thesis, University of Rochester, Rochester, NY, 1991.

Curt Burgess and Kevin Lund. Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, 12(2/3):177–210, 1997.

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

Greg Carlson and Michael Tanenhaus. Thematic roles and language comprehension. In W. Wilkins, editor, *Thematic Relations*, volume 21 of *Syntax and Semantics*. Academic Press, 1988.

Xavier Carreras and Lluís Márquez. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, 2004.

Xavier Carreras and Lluís Márquez. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, 2005.

Eugene Charniak. A Maximum-Entropy-inspired parser. In *Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2000.

Wanxiang Che, Min Zhang, and Ting Liu. A hybrid convolution tree kernel for semantic role labeling. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2006.

Kiel Christianson, Andrew Hollingworth, John F. Halliwell, and Fernanda Ferreira. Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42:368–407, 2001.

Massimiliano Ciaramita and Mark Johnson. Explaining away ambiguity: Learning verb selectional preferences with Bayesian networks. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2000.

Stephen Clark and David Weir. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206, 2002.

Charles Clifton. Thematic roles in sentence parsing. *Canadian Journal of Experimental Psychology*, 47(3):222–246, 1993.

Charles Clifton, Shari Speer, and Steven Abney. Parsing arguments: Phrase structure and argument structure as determinants of initial parsing decisions. *Journal of Memory and Language*, 30:251–271, 1993.

Charles Clifton, Matthew Traxler, Mohamed Taha Mohamed, Rihana Williams, Robin Morris, and Keith Rayner. The use of thematic role information in parsing: Syntactic autonomy revisited. *Journal of Memory and Language*, 49:317–334, 2003.

Michael Collins. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 1996.

Michael Collins. Three generative, lexicalised models for statistical parsing. In *Proceedings of the joint Annual Meeting of the Association for Computational Linguistic and Meeting of European Chapter of the Association for Computational Linguistics (ACL/EACL)*, 1997.

Ann Copestake, Alex Lascarides, and Dan Flickinger. An algebra for semantic construction in constraint-based grammars. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2001.

Stephen Crain and Mark Steedman. On not being led up the garden path: The use of context by the psychological syntax processor. In D. Dowty, L. Karttunen, and A. Zwicky, editors, *Natural Language Parsing*, pages 320–358. Cambridge University Press, 1985.

Matthew W. Crocker. *Computational psycholinguistics: An interdisciplinary approach to the study of language*. Kluwer, 1996.

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

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

Fernando Cuetos, Don Mitchell, and Martin Corley. Parsing in different languages. In Manuel Carreiras, José García-Albea, and Núria Sebastían-Gallés, editors, *Language processing in Spanish*, pages 156–187. Lawrence Erlbaum, Hillsdale, NJ, 1996.

Ido Dagan, Fernando Pereira, and Lillian Lee. Similarity-based estimation of word cooccurrence probabilities. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1994.

Timothy Desmet, Marc Brysbaert, and Constantijn de Baecke. The correspondence between sentence production and corpus frequencies in modifier attachment. *Quarterly Journal of Experimental Psychology*, 55A(3):879–896, 2002.

Timothy Desmet, Constantijn de Baecke, Denis Drieghe, Marc Brysbaert, and Wietske Vonk. Relative clause attachment in Dutch: On-line comprehension corresponds to corpus frequencies when lexical variables are taken into account. *Language and Cognitive Processes*, 21(4): 453–485, 2005.

Thomas G. Dietterich. Machine learning research: Four current directions. *AI Magazine*, 18(4): 97–136, 1999.

Jeffrey Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

Jeffrey Elman. Distributed representations, simple recurrent networks and grammatical structure. *Machine Learning*, 7:195–225, 1991.

Katrin Erk and Sebastian Padó. Analyzing models for semantic role assignment using confusability. In *Proceedings of the Joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2005.

Christiane Fellbaum, editor. *WordNet – An electronic lexical database*. MIT Press, 1998.

Fernanda Ferreira and Charles Clifton. The independence of syntactic processing. *Journal of Memory and Language*, 25:348–368, 1986.

Fernanda Ferreira and John Henderson. Use of verb information in syntactic parsing: Evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16:555–568, 1990.

Todd Ferretti, Ken McRae, and Andrea Hatherell. Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44:516–547, 2001.

W. Nelson Francis and Henry Kučera. *Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers*. Brown University, Providence, RI, 1964.

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

Lyn Frazier. Sentence processing: A tutorial review. In M. Coltheart, editor, *Attention and Performance XII: The Psychology of Reading*. Lawrence Erlbaum Associates, 1987.

Lyn Frazier, Maria Nella Carminati, Anne E. Cook, Helen Majewski, and Keith Rayner. Semantic evaluation of syntactic structure: Evidence from eye movements. *Cognition*, 99:B53–B62, 2005.

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

Tomas Gedeon, Albert Parker, and Alexander Dimitrov. Information distortion and neural coding. *Canadian Applied Mathematics Quarterly*, 10(1):33–70, 2003.

Edward Gibson, Carson T. Schütze, and Ariel Salomon. The relationship between the frequency and the processing complexity of linguistic structure. *Journal of Psycholinguistic Research*, 25: 59–92, 1996.

Daniel Gildea. Corpus variation and parser performance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2001.

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

Ana-Maria Giuglea and Alessandro Moschitti. Knowledge discovery using FrameNet, VerbNet and PropBank. In *Proceedings of the Workshop on Ontology and Knowledge Discovering at the European Conference on Machine Learning (ECML)*, 2004.

Ana-Maria Giuglea and Alessandro Moschitti. Semantic role labeling via FrameNet, VerbNet and PropBank. In *Proceedings of the joint International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistic (COLING/ACL)*, 2006.

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

John Hale. 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)*, 2001.

John Hale. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2):101–123, 2003.

Mary Hare, Ken McRae, and Jeffrey Elman. Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language*, 48:281–303, 2003.

Mary Hare, Ken McRae, and Jeffrey Elman. Admitting that admitting verb sense into corpus analyses makes sense. *Language and Cognitive Processes*, 19(2):181–224, 2004.

Virginia Holmes, Laurie Stowe, and Linda Cupples. Lexical expectations in parsing complement-verb sentences. *Journal of Memory and Language*, 28(6):668–689, 1989.

Frederick Jelinek, John Laerty, David Magerman, and Salim Roukos. Decision tree parsing using a hidden derivation model. In *Proceedings of the 1994 Human Language Technology Workshop*, 1994.

Mark Johnson. PCFG models of linguistic tree representations. *Computational Linguistics*, 24: 613–632, 1998.

R. Jonker and A. Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38:325–340, 1987.

Dan Jurafsky. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Rens Bod, Jennifer Hay, and Stefanie Jannedy, editors, *Probabilistic Linguistics*, pages 39–96. MIT Press, 2003.

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

Marcel A. Just and Patricia A. Carpenter. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87:329–354, 1980.

Frank Keller and Mirella Lapata. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484, 2003.

Frank Keller and Christoph Scheepers. Context effects on frame probability independent of verb sense ambiguity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2006.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. Class-based construction of a verb lexicon. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2000.

Dan Klein and Christopher Manning. Accurate unlexicalised parsing. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, 2003.

Anna Korhonen, Yuval Krymolowski, and Zvika Marx. Clustering polysemic subcategorization frame distributions semantically. In *Proceedings of the Annual Meeting of the Association for Computational Linguistic (ACL)*, 2003.

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

Maria Lapata, Frank Keller, and Sabine Schulte im Walde. Verb frame frequency as a predictor of verb bias. *Journal of Psycholinguistic Research*, 30(4):419–435, 2001.

Beth Levin. *English Verb Classes and Alternations*. Chicago University Press, Chicago, 1993.

Roger Levy. *Probabilistic models of word order and syntactic discontinuity*. PhD thesis, Stanford University, 2005.

Hang Li and Naoki Abe. Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244, 1998.

Sigrid Lipka. Reading sentences with a late closure ambiguity: Does semantic information help? *Language and Cognitive Processes*, 17(3):271–298, 2002.

Ken Litkowski. 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*, 2004.

Simon Liversedge, Martin Pickering, Holly Branigan, and Roger van Gompel. Processing arguments and adjuncts in isolation and context: The case of by-phrase ambiguities in passives. *Journal of Experimental Psychology*, 24(2):461–475, 1998.

Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, 28:203–208, 1997.

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

Maryellen MacDonald, Neal Pearlmutter, and Mark Seidenberg. The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101:676–703, 1994.

Scott A. MacDonald and Richard C. Shillcock. Eye movements reveal the on-line computation of lexical probabilities. *Psychological Science*, 14:648—652, 2003.

Christopher Manning and Hinrich Schütze. *Foundations of statistical language processing*. MIT Press, Cambridge, MA, 1999.

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

Zvika Marx. *Structure-based computational aspects of similarity and analogy in natural language*. PhD thesis, Hebrew University, Jerusalem, 2004.

Marshall R. Mayberry. *Incremental nonmonotonic parsing through semantic self-organization*. PhD thesis, University of Texas at Austin, 2003.

James McClelland, Mark St. John, and Roman Taraban. Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes*, 4(3/4):287–335, 1989.

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

Douglas L. Medin and Cynthia Aguilar. Categorization. In Robert A. Wilson and Frank C. Keil, editors, *The MIT Encyclopedia of the Cognitive Sciences*, pages 104–105. MIT Press, 1999.

Paola Merlo. A corpus-based analysis of verb continuation frequencies for syntactic processing. *Journal of Psycholinguistic Research*, 23(6):435–457, 1994.

Guido Minnen, John Carroll, and Darren Pearce. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–223, 2001.

Don Mitchell. Lexical guidance in human parsing: Locus and processing characteristics. In M. Coltheart, editor, *Attention and performance XII*. Erlbaum, 1987.

Don Mitchell and Marc Brysbaert. Syntax and Semantics 31: Challenges to recent theories of crosslinguistic variation in parsing: Evidence from Dutch. In Dieter Hillert, editor, *Sentence Processing: A crosslinguistic perspective*, pages 313–344. Academic Press, San Diego, CA, 1998.

Srini Narayanan and Daniel Jurafsky. A Bayesian model predicts human parse preference and reading time in sentence processing. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 59–65. MIT Press, 2002.

Srini Narayanan and Daniel Jurafsky. A Bayesian model of human sentence processing. MS, `http://www.icsi.berkeley.edu/~snarayan/newcog.pdf`, 2005.

Srini Narayanan and Daniel Jurafsky. Bayesian models of human sentence processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 1998.

Stefan Oepen, Dan Flickinger, Kristina Toutanova, and Chris Manning. LinGO Redwoods: A rich and dynamic treebank for HPSG. In *Beyond PARSEVAL. Workshop at the Third Conference on Language Resources and Evaluation (LREC)*, 2002.

Sebastian Padó and Maria Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, to appear.

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

Martin Pickering and Matthew Traxler. Plausibility and recovery from garden paths. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24, 1998.

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

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

Bradley Pritchett. *Grammatical Competence and Parsing Performance*. The University of Chicago Press, 1992.

Trivellore Raghunathan. An approximate test for homogeneity of correlated correlations. *Quality and Quantity*, 37:99–110, 2003.

Keith Rayner, Marcia Carlson, and Lyn Frazier. 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, 1983.

Keith Rayner, Simon Garrod, and Charles A. Perfetti. Discourse influences during parsing are delayed. *Cognition*, 45:109–139, 1992.

Philip Resnik. Left-corner parsing and psychological plausibility. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Nantes, France, 1992.

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

Philip Resnik. Selectional preference and sense disambiguation. In *Proceedings of the ACL Siglex Workshop on Tagging Text with Lexical Semantics, Why, What and How?*, 1997.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

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

Douglas Rohde. *A Connectionist Model of Sentence Comprehension and Production*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2002.

Douglas Rohde. Tgrep2 user manual. `tedlab.mit.edu/~dr/Tgrep2/`, 2001.

Doug Roland and Daniel Jurafsky. Verb sense and verb subcategorisation probabilities. In P. Merlo and S. Stevenson, editors, *The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues*, pages 325–346. John Benjamins, Amsterdam, 2002.

Douglas Roland and Daniel Jurafsky. How verb subcategorization frequencies are affected by corpus choice. In *Proceedings of the joint International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, 1998.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, and Christopher R. Johnson. *FrameNet: Theory and Practice*. e-book, 2005.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLap)*, 1994.

Sabine Schulte im Walde and Chris Brew. Inducing German semantic verb classes from purely syntactic subcategorisation information. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

Carson Schütze and Edward Gibson. Argumenthood and English prepositional phrase attachment. *Journal of Memory and Language*, 40:409–431, 1999.

Wojciech Skut, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit. Annotating unrestricted German text. In *Fachtagung der Sektion Computerlinguistik der Deutschen Gesellschaft für Sprachwissenschaft*, 1997.

Michael Spivey and Michael Tanenhaus. 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, 1998.

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

Michael Spivey-Knowlton and Julie Sedivy. Parsing attachment ambiguities with multiple constraints. *Cognition*, 55:227–267, 1995.

Mark St. John and James McClelland. Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46:217–257, 1990.

Mark Steedman. Connectionist sentence processing in perspective. *Cognitive Science*, 23(4): 615–634, 1999.

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

Laurie Stowe. Thematic structures and sentence comprehension. In G. Carlson and M. Tanenhaus, editors, *Linguistic structure in language processing*, pages 319–357. Kluwer Academic Publishers, 1989.

Patrick Sturt, Martin Pickering, and Matthew W. Crocker. Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, 40:136–150, 1999.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. Using predicate-argument structures for information extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2003.

Whitney Tabor and Michael K. Tanenhaus. Dynamical models of sentence processing. *Cognitive Science*, 23(4):491–515, 1999.

Whitney Tabor, Cornell Juliano, and Michael Tanenhaus. Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes*, 12(2/3):211–271, 1997.

Patrizia Tabossi, Michael Spivey-Knowlton, Ken McRae, and Michael Tanenhaus. Semantic effects on syntactic ambiguity resolution: Evidence for a constraint-based resolution process. In C. Umilta and M. Moscovitch, editors, *Attention and Performance XV*, pages 589–615. Lawrence Erlbaum Associates, 1994.

Michael Tanenhaus, Michael Spivey-Knowlton, and Joy Hanna. Modeling thematic and discourse context effects with a multiple constraints approach: Implications for the architecture of the language comprehension system. In M. W. Crocker, M. Pickering, and C. Clifton, editors, *Architectures and Mechanisms for Language Processing*, pages 99–118. Cambridge University Press, 2000.

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

Naftali Tishby, Fernando C. Pereira, and William Bialek. The Information Bottleneck method. In *Proceedings of the Annual Allerton Conference on Communication, Control and Computing*, 1999.

John Trueswell, Michael Tanenhaus, and Christopher Kello. 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, 1993.

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

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

Roger van Gompel and Martin Pickering. Lexical guidance in sentence processing: A note on Adams, Clifton and Mitchell (1998). *Psychonomic Bulletin and Review*, 8(4):851–857, 2001.

Roger van Gompel, Martin Pickering, and Matthew Traxler. Reanalysis in sentence processing: Evidence against current constraint-based and two-stage models. *Journal of Memory and Language*, 45:225–258, 2001.

Nianwen Xue and Martha Palmer. 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)*, 2004.

Alexander Yeh. More accurate tests for the statistical significance of result differences. In *Proceedings of the International Conference on Computational Linguistics*, 2000.