Probabilistic grammars as models of gradience in language processing

Matthew W. Crocker
Saarland University

Frank Keller
University of Edinburgh

Abstract[1]

This article deals with gradience in human sentence processing. We review the experimental evidence for the role of experience in guiding the decisions of the sentence processor. Based on this evidence, we argue that the gradient behavior observed in the processing of certain syntactic constructions can be traced back to the amount of past experience that the processor has had with these constructions. In modeling terms, linguistic experience can be approximated using large, balanced corpora. We give an overview of corpus-based and probabilistic models in the literature that have exploited this fact, and hence are well placed to make gradient predictions about processing behavior. Finally, we discuss a number of questions regarding the relationship between gradience in sentence processing and gradient grammaticality, and come to the conclusion that these two phenomena should be treated separately in conceptual and modeling terms.

1 Introduction

Gradience in language comprehension can be manifest in a variety of ways, and have various sources of origin. Based on theoretical and empirical results, one possible way of classifying such phenomena is whether they arise from the **grammaticality** of a sentence, perhaps reflecting the relative importance of various syntactic constraints, or arise from **processing**, namely the mechanisms which exploit our syntactic knowledge for incrementally recovering the structure of a given sentence. Most of the chapters in this volume are concerned with the former: how to characterize and explain the **gradient grammaticality** of a given utterance, as measured, for example, by judgments concerning **acceptability**. While the study of gradient grammaticality has a long history in the generative tradition (Chomsky 1975, 1964), recent approaches such as the minimalist program (Chomsky 1995) do not explicitly allow for gradience as part of the grammar.

In this chapter, we more closely consider the phenomena of **gradient performance**: how can we explain the variation in processing difficulty, as reflected for example in word-by-word reading times? Psycholinguistic research has identified two key sources of processing difficulty in sentence comprehension: local ambiguity and processing load. In the case of local, or temporary ambiguity, there is abundant evidence that people adopt some preferred interpretation immediately, rather then delaying interpretation. Should the corresponding syntactic analysis be disconfirmed by the sentence's continuation, reanalysis is necessary, and is believed to be an important contributor to observable difficulties in processing. It is also the case, however, that processing difficulties are found in completely unambiguous utterances, such as center embedded structures. One explanation of such effects is that, despite being both grammatical and unambiguous, such sentences require more cognitive processing resources (such as working memory) than are available.

While these phenomena have been well studied, both empirically and theoretically, there has been little attempt to model relative processing difficulty: why some sentences are more difficult than others, and precisely how difficult they are. Quantitative models, which can predict real-valued behavioral measures are even less common. We argue, however, that one relatively new class of models offers considerable promise in addressing this issue. The common distinguishing feature of the models we discuss here is that they are **experience-based**. The central idea behind experienced-based models is that the mechanisms which people use to arrive at an incremental interpretation of a sentence are crucially dependent on relevant prior experience. Generally speaking, interpretations which are supported by our prior experience are preferred to those which are not. Furthermore, since experience is generally encoded in models as some form of relative likelihood, or activation, it is possible for models to generate real-valued, graded predictions about the processing difficulty of a particular sentence.

We begin by reviewing some of the key psycholinguistic evidence motivating the need for experience-based mechanisms, before turning to a discussion of recent models. We focus our attention here on probabilistic models of human sentence processing, which attempt to assign a probability to a given sentence, as well as to alternative parse interpretations for that sentence. Finally, we will discuss the relationship between probabilistic models of performance (gradient processing complexity), and probabilistic models of competence (gradient grammaticality). A crucial consequence of the view we propose is that the likelihood of a (partial) structure is only meaningful relative to the likelihood of competing (partial) structures, and does not provide an independently useful characterization of the grammaticality of the alternatives. Thus we argue that a probabilistic characterization of gradient grammaticality should be quite different from a probabilistic performance model.

2 The role of experience in sentence processing

People are continually faced with the problem of resolving the ambiguities that occur in the language they hear and read (Altmann 1998). Computational theories of human language comprehension therefore place much emphasis on the **algorithms** for constructing syntactic and semantic interpretations, and the **strategies** for deciding among alternatives, when more than one interpretation is possible (Crocker 1999). The fact that people understand language **incrementally**, integrating each word into their interpretation of the sentence as it is encountered, means that people are often forced to resolve ambiguities before they have heard the entire utterance. While it is clear that many kinds of information are involved in ambiguity resolution (Gibson and Pearlmutter 1998), much attention has recently been paid to the role of **linguistic experience**. That is to say, to what extent do the mechanisms underlying human language comprehension rely on previous linguistic encounters to guide them in resolving an ambiguity they currently face?

During his or her lifetime, the speaker of a language accrues linguistic experience. Certain lexical items are encountered more often than others, some syntactic structures are used more frequently, and ambiguities are often resolved in a particular manner. In lexical processing, for example, the influence of experience is clear: high frequency words are recognized more quickly than low frequency ones (Grosjean 1980), syntactically ambiguous words are initially perceived as having their most likely part of speech (Crocker and Corley 2002), and semantically ambiguous words are associated with their more frequent sense (Duffy et al. 1988).

Broadly, we define a speaker's **linguistic experience** with a given linguistic entity as the number of times the speaker has encountered this entity in the past. Accurately measuring someone's linguistic experience would (in the limit) require a record of all text or speech that person has ever been exposed to. Additionally, there is the issue of how experience is manifest in the syntactic processing mechanism. The impracticality of this has lead to alternative proposals for approximating linguistic experience, such as norming experiments or corpus studies.

Verb frames are an instance of linguistic experience whose influence on sentence processing has been researched extensively in the literature. The frames of a verb determine the syntactic complements it can occur with. For example, the verb *know* can appear with a sentential complement (S frame) or with a noun phrase complement (NP frame). Norming studies can be conducted in which subjects are presented with fragments such as (1) and complete them to form full sentences.

(1) The teacher knew ___ .

Subjects might complete the fragment using *the answer* (NP frame) or *the answer was false* (S frame). Verb frame frequencies can then be estimated as the frequencies with which subjects use the S frame or the NP frame (Garnsey et al. 1997). An alternative to the use of completion frequencies is the use of frequencies obtained in a free production task, where subjects are presented only with a verb, and are asked to produce a sentence incorporating this verb (Connine et al. 1984).

An alternative technique is to extract frequency information from a **corpus**, a large electronic collection of linguistic material. A **balanced** corpus (Burnard 1995, Francis et al. 1982), which contains representative samples of both text and speech in a broad range of genres and styles, is often assumed to provide an approximation of human linguistic experience. In our examples, all instances of *know* could be extracted from a corpus, counting how often the verb occurs with the NP and the S frame.

Additionally, however, there is the issue of how experience is manifest in the syntactic processing mechanism. A simple frequentist approach would mean that all our experience has equal weight, whether an instance of exposure occurred ten seconds ago, or ten years ago. This is true for the kinds of probabilistic models we discuss here. Thus an interesting difference between corpus estimates and norming studies is that the former approximates the experience presented to a speaker, while the latter reflects the influence of that experience on a speaker's preferences. Results in the literature broadly indicate that frame frequencies obtained from corpora and norming studies are reliably correlated (Lapata et al. 2001, Sturt et al. 1999). It should be borne in mind, however, that corpus frequencies vary as a function of the genre of the corpus (Roland and Jurafsky 1998 compared text and speech corpora) and also verb senses play a role (Roland and Jurafsky 2002).

Once language experience has been measured using norming or corpus studies, the next step is to investigate how the human language processor uses experience to resolve ambiguities in real time. A number of studies have demonstrated the importance of **lexical** frequencies. These frequencies can be categorical (e.g. the most frequent part of speech for an ambiguous word, Crocker and Corley 2002), morphological (e.g. the tendency of a verb to occur in a particular tense, Trueswell 1996), syntactic (e.g. the tendency of a verb to occur with a particular frame, as discussed above, Ford et al. 1982, Garnsey et al. 1997, Trueswell et al. 1993), or semantic (e.g. the tendency of a noun to occur as the object of a particular verb, Garnsey et al. 1997, McRae et al. 1998, Pickering et al. 2000). It has been generally argued that these different types of lexical frequencies form a set of interacting constraints that determine the preferred parse for a given sentence (MacDonald 1994, MacDonald et al. 1994, Trueswell and Tanenhaus 1994).

Other researchers (Brysbaert and Mitchell 1996, Mitchell et al. 1996) have taken the stronger view that the human parser not only makes use of lexical frequencies, but also keeps track of **structural** frequencies. This view, known as the **tuning hypothesis**, states that the human parser deals with ambiguity by initially selecting the syntactic analysis that has worked most frequently in the past (see figure 1).

The fundamental question that underlies both lexical and structural experience models is the **grain problem**: What is the level of granularity at which the human sentence processor 'keeps track' of frequencies? Does it count lexical frequencies or structural frequencies (or both), or perhaps frequencies at an intermediate level, such as the frequencies of individual phrase structure rules? The latter assumption underlies a number of experience-based models that are based on probabilistic context free grammars (see figure 2 for details). Furthermore, at the lexical level, are frame frequencies for verbs forms counted separately (e.g. *know, knew, knows, . . .*) or are they combined into a set of total frequencies for the verb's base form (the lemma KNOW) (Roland and Jurafsky 2002)?

INSERT FIGURE 1 HERE

----

3 Probabilistic models of sentence processing

Theories of human syntactic processing have traditionally down played the importance of frequency (Fodor and Frazier 1978, Marcus 1980, Pritchett 1992), focusing rather on the characterization of more general, sometimes language universal, processing mechanisms (Crocker 1996). An increasing number of models, however, incorporate aspects of linguistic experience in some form or other. This is conceptually attractive, as an emphasis on experience may help to explain some of the rather striking, yet often unaddressed, properties of human sentence processing:

- **Efficiency:** The use of experience-based heuristics, such as choosing the reading that was correct most often in the past, helps explain rapid and seemingly effortless processing, despite massive ambiguity.
- **Coverage:** In considering the full breadth of what occurs in linguistic experience, processing models will be driven to cover more linguistic phenomena, and may look quite different from the toy models which are usually developed.

- **Performance:** Wide-coverage experience-based models can offer an explanation of how people rapidly and accurately understand most of the language they encounter, while also explaining the kinds of **pathologies** which have been the focus of most experimental and modeling research.
- **Robustness:** Human language processing is robust to slips of the tongue, disfluencies, and minor ungrammaticalities. The probabilistic mechanisms typically associated with experience-based models can often provide sensible interpretations even in the face of such noise.
- **Adaptation:** The human language processor is finely tuned to the linguistic environment it inhabits. This adaptation is naturally explained if processing mechanisms are the product of learning from experience.

Approaches in the literature differ substantially in how they exploit linguistic experience. Some simply permit heterogeneous linguistic constraints to have 'weights' which are determined by frequency (MacDonald et al. 1994, Tanenhaus et al. 2000), others provide probabilistic models of lexical and syntactic processing (Crocker and Brants 2000, Jurafsky 1996), while connectionist models present yet a further paradigm for modeling experience (Christiansen and Chater 1999, 2001, Elman 1991, 1993).

Crucially, however, whether experience is encoded via frequencies, probabilities, or some notion of activation, all these approaches share the idea that sentences and their interpretations will be associated with some real-valued measure of **goodness**: namely how likely or plausible an interpretation is, based on our prior experience. The appeal of probabilistic models is that they acquire their parameters from data in their environment, offering a transparent relationship between linguistic experience and a model's behavior. The probabilities receive a cognitive interpretation; typically a high probability is assumed to correlate with a low processing effort. This suggests that the human sentence processor will prefer the structure with the lowest processing effort when faced with a syntactic ambiguity (see figure 1 for an example). Before considering probabilistic models of human processing in more detail, we first quickly summarize the ideas that underlie probabilistic parsing.

3.1 Probabilistic grammars and parsing

A probabilistic grammar consists of a set of symbolic rules (e.g. context free grammar rules) annotated with rule application probabilities. These probabilities can then be combined to compute the overall probability of a sentence, or for a particular syntactic analysis of a sentence. The rule probabilities are typically derived from a corpus – a large, annotated collection of text or speech. In cognitive terms, the corpus can be regarded as an approximation of the language experience of the user; the probabilities a reflection of language use, i.e., they provide a model of linguistic performance.

Many probabilistic models of human sentence processing are based on the framework of probabilistic context free grammars (PCFGs, see Manning and Schütze 1999, for an overview). PCFGs augment standard context free grammars by annotating grammar rules with rule probabilities. A rule probability expresses the likelihood of the lefthand side of the rule expanding to its righthand side. As an example, consider the rule VP → V NP in figure 2a. This rule says that a verb phrase expands to a verb followed by a noun phrase with a probability of 0.7.

In a PCFG, the probabilities of all rules with the same lefthand side have to sum to one:

(2) $\forall i \sum_{j} P(N^{i} \rightarrow \zeta^{j}) = 1$

where $P(N^i \rightarrow \zeta^j)$ is the probability of a rule with the lefthand side N^i and the righthand side ζ^j. For example, in figure 2a the two rules VP → V NP and VP → VP PP share the same lefthand side (VP), so their probabilities sum to one.

The probability of a parse tree generated by a PCFG is computed as the product of its the rule probabilities:

(3) $$P(t) = \prod_{(N \rightarrow \zeta) \in R} P(N \rightarrow \zeta)$$

where R is the set of all rules applied in generating the parse tree t. It has been suggested that the probability of a grammar rule models how easy this rule can be accessed by the human sentence processor (Jurafsky 1996). Structures with greater overall probability should be easier to construct, and therefore preferred in cases of ambiguity. As an example consider the PCFG in figure 2a. This grammar generates two parses for the the sentence *John hit the man with the book*. The first parse t_1 attaches the prepositional phrase *with the book* to the noun phrase (low attachment), see figure 2b. The PCFG assigns t_1 the following probability, computed as the product of the probabilities of the rules used in this parse:

(4) $$P(t_1) = 1.0 \times 0.2 \times 0.7 \times 1.0 \times 0.2 \times 0.6 \times 1.0 \times 1.0 \times 0.5 \times 1.0 \times 0.6 \times 1.0 \times 0.5 = 0.00252$$

The alternative parse t_2, with the prepositional phrase attached to the verb phrase (high attachment, see figure 2c) has the following probability:

(5) $$P(t_2) = 1.0 \times 0.2 \times 0.3 \times 0.7 \times 1.0 \times 1.0 \times 0.6 \times 1.0 \times 0.6 \times 1.0 \times 0.5 \times 1.0 \times 0.5 = 0.00378$$

Under the assumption that the probability of a parse is a measure of processing effort, we predict that t_2 (high attachment) is easier to process than t_1, as it has a higher probability.

In applying PCFGs to the problem of human sentence processing, an important additional property must be taken into account: incrementality. That is, people face a local ambiguity as soon as they hear the fragment *John hit the man with ...* and must decide which of the two possible structures is to be preferred. This entails that the parser is able to compute **prefix** probabilities for sentence initial substrings, as the basis for comparing alternative (partial) parses. Existing models provide a range of techniques for computing and comparing such parse probabilities incrementally (Brants and Crocker 2000, Hale 2001, Jurafsky 1996). For the example in figure 2, however, the preference for t_2 would be predicted even before the final NP is processed, since the probability of that NP is the same for both structures.

Note that the move from CFGs to PCFGs also raises a number of other computational problems, such as the problem of efficiently computing the most probable parse for a given input sentence. Existing parsing schemes can be adapted to PCFGs, including shift-reduce parsing (Briscoe and Carroll 1993) and left-corner parsing (Stolcke 1995). These approaches all use the basic Viterbi algorithm (Viterbi 1967) for efficiently computing the best parse generated by a PCFG for a given sentence.

INSERT FIGURES 2a, 2b, 2c HERE

3.2 Probabilistic models of human behavior

Jurafsky (1996) suggests using Bayes' rule to combine structural probabilities generated by a probabilistic context free grammar with other probabilistic information. The model therefore integrates multiple sources of experience into a single, mathematically founded framework. As an example consider again the fragment in (1). When a speaker reads or hears *know*, he or she has the choice between two syntactic readings, involving either an S complement or an NP complement.

Jurafsky's model computes the probabilities of these two readings based on two sources of information: the overall structural probability of the S reading and the NP reading, and the lexical probability of the verb *know* occurring with an S or an NP frame. The structural probability of a reading is independent of the particular verb involved; the frame probability, however, varies with the verb. This predicts that in some cases lexical probabilities can override structural probabilities.

Jurafsky's model is able to account for a range of parsing preferences reported in the psycholinguistic literature. However, it might be criticized for its limited coverage, i.e., for the fact that it uses only a small lexicon and grammar, manually designed to account for a handful of example sentences. In the computational linguistics literature, on the other hand, **broad coverage** parsers are available that compute a syntactic structure for arbitrary corpus sentences with an accuracy of about 90% (Charniak 2000). Psycholinguistic models should aim for similar coverage, which is clearly part of human linguistic performance.

This issue has been addressed by Corley and Crocker's (2000) broad coverage model of lexical category disambiguation. Their approach uses a bigram model to incrementally compute the probability that a string of words $w_0 \ldots w_n$ has the part of speech sequence $t_0 \ldots t_n$ as follows:

(6) $$P(t_0 \ldots t_n, w_0 \ldots w_n) \approx \prod_{i=1}^{n} P(w_i|t_i)\, P(t_i|t_{i-1})$$

Here, $P(w_i|t_i)$ is the conditional probability of word w_i given the part of speech t_i, and $P(t_i|t_{i-1})$ is the probability of t_i given the previous part of speech t_{i-1}. This model capitalizes on the insight that many syntactic ambiguities have a lexical basis, as in (7):

(7) The warehouse prices/makes___ .

These fragments are ambiguous between a reading in which *prices* or *makes* is the main verb or part of a compound noun. After being trained on a large corpus, the model predicts the most likely part of speech for *prices*, correctly accounting for the fact that people understand *prices* as a noun, but *makes* as verb (Crocker and Corley 2002, Frazier and Rayner 1987, MacDonald 1993). Not only does the model account for a range of disambiguation preferences rooted in lexical category ambiguity, it also explains why, in general, people are highly accurate in resolving such ambiguities.

More recent work on broad coverage parsing models has extended this approach to full syntactic processing based on PCFGs (Crocker and Brants 2000). This research demonstrates that when such models are trained on large corpora, they are nolt only able to account for human disambiguation behavior, but are also able to maintain high overall accuracy under strict memory and incremental processing restrictions (Brants and Crocker 2000).

Finally, it is important to stress that the kind of probabilistic models we outline here emphasizes lexical and syntactic information in estimating the probability of a parse structure. To the extent that a PCFG is lexicalized, with the head of each phrase being projected upwards to phrasal nodes (Collins 1999), some semantic information may also be implicitly represented in the form of word co-occurrences (e.g. head-head co-occurrences). In addition to being incomplete models of interpretation, such lexical dependency probabilities are poor at modeling the likelihood of plausible but improbable structures. Probabilistic parsers in their current form are therefore only appropriate for modeling syntactic processing preferences. Probabilistic models of human semantic interpretation and plausibility remain a largely unexplored area of research.

3.3 Towards quantitative models of performance

So far, probabilistic models of sentence processing have only been used to account for qualitative data about human sentence processing (e.g. to predict whether a garden path occurs). By quantifying the likelihood of competing structural alternatives, however, such models in principle offer hope for more quantitative accounts of gradient behavioral data. (e.g. to predict the strength of a garden path). In general terms, this would entail that the probability assigned to a syntactic structure is to be interpreted as a measure of the degree of processing difficulty triggered by this structure. Gradient processing difficulty in human sentence comprehension can be determined experimentally, for example by recording reading times in self-paced reading studies or eyetracking experiments. An evaluation of a probabilistic model should therefore be conducted by correlating the probability predicted by the model for a given structure with reading times (and other indices of processing difficulty).

This new way of evaluating processing models raises a number of questions. Most importantly, an explicit **linking hypothesis** is required, specifying which quantity computed by the model would be expected to correlate with human processing data. One possible measure of processing difficulty would be the probability ratio of alternative analyses (Jurafsky 1996). That is, in addition to predicting the highest probability parse to be the easiest, we might expect the cost of switching to a less preferred parse to be correlated with the probability ratio of the preferred parse with respect to the alternative.

Hale (2003) suggest an alternative, proposing that the word by word processing complexity is dominated by the amount of **information** the word contributes concerning the syntactic structure, as measured by **entropy reduction**. Hale's model is thus in stark contrast with the previous probabilistic parsing accounts, in that he does not assume that switching from a preferred parse to an alternative is the primary determinant of processing cost. To date, Hale's model has been evaluated on rather different kinds of structures than the probabilistic parsers discussed above. Reconciliation of the probabilistic disambiguation versus entropy reduction approaches - and their ability to qualitatively model reading time data - remains an interesting area for future research.

3.4 Evidence against likelihood in sentence processing

Experience-based models often assume some frequency-based ambiguity resolution mechanism: prefer the interpretation which has the highest likelihood of being correct, namely the higher relative frequency. One well-studied ambiguity is prepositional phrase attachment:

(8) John hit the man [$_{PP}$ with the book].

Numerous on-line experimental studies have shown an overall preference for high attachment, i.e., for the association of the PP with the verb (e.g. as the instrument of *hit*) (Ferreira and Clifton 1986, Rayner et al. 1983). Corpus analyses, however, reveal that low attachment (e.g. interpreting the PP as a modifier of *the man*) is about twice as frequent as attachment to the verb (Hindle and Rooth 1993). Such evidence presents a challenge for accounts relying on exclusively on structural frequencies, but may be accounted for by lexical preferences for specific verbs (Taraban and McClelland 1988). Another problem for structural tuning comes from three-site relative clause attachments analogous to that in figure 1, but containing an additional NP attachment site:

(9) [$_{high}$ The friend] of [$_{mid}$the servant] of [$_{low}$the actress] [$_{RC}$who was on the balcony] died.

While corpus analysis suggest a preference for *low* > *middle* > *high* attachment (though such structures are rather rare), experimental evidence suggests an initial preference for *low*>*high* >*middle* (with middle being in fact very difficult) (Gibson et al. 1996a,b). A related study investigating noun phrase conjunction ambiguities (instead of relative clause) for such three site configurations revealed a similar asymmetry between corpus frequency and human preferences (Gibson and Schütze 1999).

Finally, there is recent evidence against lexical verb frame preferences:

(10) The athlete realized $[_S\ [_{NP}$ her shoes/goals] were out of reach].

Reading times studies have shown an initial preference for interpreting *her goals* as a direct object (Pickering et al. 2000), even when the verb is more likely to be followed by a sentence complement (see also Sturt et al. 2001, for evidence against the use of such frame preferences in reanalysis). These findings might be taken as positive support for the tuning hypothesis, since object complements are more frequent than sentential complements overall (i.e., independent of the verb). Pickering et al. (2000), building on previous theoretical work (Chater et al. 1998), suggest that the parser may in fact still be using an experience-based metric, but not one which maximizes likelihood alone.

4 Probabilistic models of gradient grammaticality

As argued in detail in the previous section, probabilistic grammars can be used to construct plausible models of human language processing, based on the observation that the disambiguation decisions of the human parser are guided by experience. This raises the question whether experience-based models can also be developed for other forms of linguistic behavior, such as gradient grammaticality judgments. This issue will be discussed in this section.

4.1 Probabilities vs. degrees of grammaticality

We might want to conjecture that probabilistic models such as PCFGs can be adapted so as to account for gradient grammaticality, with probabilities being reinterpreted as degrees of grammaticality. The underlying assumption of such an approach is that language experience (approximated by the frequencies in a balanced corpus) not only determines disambiguation behavior, but also determines (or at least influences) the way speakers make grammaticality judgments. The simplest model would be one where the probability of a syntactic structure (as estimated from a corpus) is directly correlated with its degree of grammaticality. This means that a speaker, when required to make a grammaticality judgment for a given structure, will draw on his or her experience with this structure to make this judgment. Manning (2003) outlines a probabilistic model of gradient grammaticality that comes close to this view. (However, he also acknowledges that such a model would have to take the context of an utterance into account, so as to factor out linguistically irrelevant factors, including world knowledge.)

Other authors take a more skeptical view of the relationship between probability and grammaticality. Keller (2000b), for instance, argues that the degree of grammaticality of a structure and its probability of occurrence in a corpus are two distinct concepts, and it seems unlikely they can both be modeled in the same probabilistic framework. A related point of view is put forward by Abney (1996), who states that '[w]e must also distinguish degrees of grammaticality, and indeed, global goodness, from the probability of producing a sentence. Measures of goodness and probability are mathematically similar enhancements to algebraic grammars, but goodness alone does not determine probability. For example, for an infinite language, probability must ultimately decrease with length, though arbitrary long sentences may be perfectly good' (Abney 1996, 14). He also gives a number of examples for sentences that have very improbable, but perfectly grammatical readings. A similar point is made by Culy (1998), who argues that the statistical distribution of a construction does not bear on the question of whether it is grammatical or not.

Riezler (1996) agrees that probabilities and degrees of grammaticality are to be treated as separate concepts. He makes this point by arguing that, if one takes the notion of degree of grammaticality seriously for probabilistic grammars, there is no sensible application to the central problem of ambiguity resolution any more. A probabilistic grammar model cannot be trained so that the numeric value is assigned to a structure can function both as a well-formedness score (degree of grammaticality) and as a probability to be used for ambiguity resolution.

Keller and Asudeh (2002) present a similar argument in the context of optimality theory (OT). They point out that if an OT grammar were to model both corpus frequencies and degrees of grammaticality, then this would entail that the grammar incorporates both performance constraints (accounting for frequency effects) and competence constraints (accounting for grammaticality effects). This is highly undesirable in an OT setting, as it allows the crosslinguistic re-ranking of performance and competence constraints. Hence such a combined competence/performance grammar predicts that crosslinguistic differences can be caused by performance factors (e.g. memory limitations). Clearly, this is a counterintuitive consequence.

A further objection to a PCFG approach to gradient grammaticality is that, in assigning probabilities to gradient structures requires the grammar to contain rules used in 'ungrammatical' structures. It might not be plausible to assume that such rules are part of the mental grammar of a speaker. However, any realistic grammar of naturally occurring language (i.e., a grammar that covers a wide range of constructions, genres, domains, and modalities) has to contain a large number of low-frequency rules anyway, simply in order to achieve broad coverage and robustness. We can therefore assume that these rules are also being used to generate structures with a low degree of grammaticality.

4.2 Probabilistic grammars and gradient acceptability data

The previous section reviewed a number of arguments regarding the relationship between probabilities (derived from corpora) and degrees of grammaticality. However, none of the authors cited offers any experimental results (or corpus data) to support their position; the discussion remains purely conceptual. A number of empirical studies have recently become available to shed light on the relationship between probability and grammaticality.

Keller (2003) studies the probability/grammaticality distinction based on a set of gradient acceptability judgments for word order variation in German. The data underlying this study were gathered by Keller (2000a), who used an experimental design that crossed the factors verb order (initial or final), complement order (subject first or object first), pronominalization, and context (null context, all focus, subject focus, and object focus context). Eight lexicalizations of each of the orders were judged by a total of 51 native speakers using a magnitude estimation paradigm (Bard et al. 1996). The results show that all of the experimental factors have a significant effect on judged acceptability, with the effects of complement order and pronominalization modulated by context. A related experiment is reported by Keller (2000b), who uses ditransitive verbs (i.e., complement orders including an indirect object) instead of transitive ones.

Keller (2003) conducts a modeling study using the materials of Keller (2000a) and Keller (2000b), based on the syntactically annotated Negra corpus (Skut et al. 1997). He trains a probabilistic context-free grammar on Negra and demonstrates that the sentence probabilities predicted by this model correlate significantly with acceptability scores measured experimentally.

Keller (2003) also shows that the correlation is higher if a more sophisticated lexicalized grammar model (Carroll and Rooth 1998) is used.

This result is not incompatible with the claim that there is a divergence between the degree of acceptability of a sentence and its probability of occurrence, as discussed in the previous section. The highest correlation Keller (2003) reports is .64, which corresponds to 40% of the variance accounted for. However, this is achieved on a data set (experiment 1) which contains a contrast between verb final (fully grammatical) and verb initial (fully ungrammatical) sentences; it is not surprising that a PCFG trained on a corpus of fully grammatical structures (but not on ungrammatical ones) can make this distinction and thus achieves a fairly high correlation. On a corpus of only verb final structures that show relatively small differences in acceptability (experiment 2), a much lower (though still significant) correlation of .23 is achieved. This means that the PCFG only models 5% of the variance. In other words, Keller's (2003) results indicate that the degree of grammaticality of a sentence is largely determined by factors other than its probability of occurrence (at least as modeled by a PCFG).

A related result is reported by Kempen and Harbusch (2004), who again deal with word order variation in German. They compare 24 word orders obtained by scrambling the arguments of ditransitive verbs (all possible argument permutations, with zero or one of the arguments pronominalized). Frequencies were obtained for these 24 orders from two written corpora and one spoken corpus and compared against gradient grammaticality judgments from Keller's (2000b) study. The results are surprising in that they show that there is much less word order variation than expected; just four orders account for the vast majority of corpus instances. Furthermore, Kempen and Harbusch (2004) demonstrate what they term the **frequency-grammaticality gap**: all the word orders that occur in the corpus are judged as highly grammatical, but some word orders that never occur in the corpus nevertheless receive grammaticality judgments in the medium range. This result is consistent with Keller's (2003) finding: it confirms that there is only an imperfect match between the frequency of a structure and its degree of grammaticality (as judged by a native speaker). Kempen and Harbusch (2004) explain the frequency-grammaticality gap in terms of sentence production: they postulate a **canonical rule** that governs word order during sentence production. The judgment patterns can then be explained with the additional assumption that the participants in a grammaticality judgment task estimate how plausible a given word order is as the outcome of incremental sentence production (governed by the canonical rule).

Featherston (2004) presents another set of data that sheds light on the relationship between corpus frequency and grammaticality. The linguistic phenomenon he investigates is object co-reference for pronouns and reflexives in German (comparing a total of 16 co-reference structures, e.g. *ihn*$_i$ *ihm*$_i$ 'him.ACC him.DAT', *ihn*$_i$ *sich*$_i$ 'him.ACC REFL.DAT'). In a corpus study, Featherston (2004) finds that only one of these 16 co-reference structures is reasonably frequent; all other structures occur once or zero times in the corpus. Experimentally obtained grammaticality data show that the most frequent structure is also the one with the highest degree of grammaticality. However, there is a large number of structures that also receive high (or medium) grammaticality judgments, even though they are completely absent in the corpus. This result is fully compatible with the frequency-grammaticality gap diagnosed by Kempen and Harbusch (2004). Like them, Featherston (2004) provides an explanation in terms of sentence production, but one that assumes a two-stage architecture. The first stage involves the cumulative application of linguistic constraints, the second stage involves the competitive selection of a surface string. Grammaticality judgments are made based on the output of the first stage (hence constraints violation are cumulative, and there are multiple output forms with a similar degree of grammaticality). Corpus data, on the other hand, are produced as the output of the second stage (hence there is no cumulativity, and only a small number of optimal output forms can occur).

5 Conclusion

There is clear evidence for the role of lexical frequency effects in human sentence processing, particularly in determining lexical category and verb frame preferences. Since many syntactic ambiguities are ultimately lexically based, direct evidence for purely structural frequency effects, as predicted by the **tuning hypothesis**, remains scarce (Jurafsky 2002).

Probabilistic accounts offer natural explanations for lexical and structural frequency effects, and a means for integrating the two using lexicalized techniques that exists in computational linguistics (e.g. Carroll and Rooth 1998, Charniak 2000, Collins 1999). Probabilistic models also offer good scalability and a transparent representation of symbolic structures and their likelihood. Furthermore, they provide an inherently gradient characterization of sentence likelihood, and the relative likelihood of alternative interpretations, promising the possibility of developing truly quantitative accounts of experimental data.

More generally, however, experience-based models not only offer an account of specific empirical facts, but can more generally be viewed as **rational** (Anderson 1990). That is, their behavior typically resolves ambiguity in a manner that has worked well before, maximizing the likelihood of correctly understanding ambiguous utterances. This is consistent with the suggestion that human linguistic performance is indeed highly adapted to its environment and the task rapidly of correctly understanding language (Chater et al. 1998, Crocker, to appear). It is important to note however, that such adaptation based on linguistic experience does not necessitate mechanisms which are strictly based on frequency-based estimations of likelihood

(Pickering et al. 2000). Furthermore, different kinds and grains of frequencies may interact or be combined in complex ways (McRae et al. 1998).

It must be remembered, however, that experience is not the sole determinant of ambiguity resolution behavior (Gibson and Pearlmutter 1998). Not only are people clearly sensitive to immediate linguistic and visual context (Tanenhaus et al. 1995), some parsing behaviors are almost certainly determined by alternative processing considerations, such as working memory limitations (Gibson 1998). Any complete account of gradience in sentence processing must explain how frequency of experience, linguistic and non-linguistic knowledge, and cognitive limitations are manifest in the mechanisms of the human sentence processor.

An even greater challenge to the experience-based view is presented by gradient grammaticality judgments. A series of studies is now available that compares corpus frequencies and gradient judgments for a number of linguistic phenomena (Featherston 2004, Keller 2003, Kempen and Harbusch 2004). These studies indicate that there is no straightforward relationship between the frequency of a structure and its degree of grammaticality, which indicates that not only experience, but also a range of processing mechanisms (most likely pertaining to sentence production) have to be invoked in order to obtain a plausible account of gradient grammaticality data.

References

Abney, S. (1996). Statistical methods and linguistics. In J. Klavans, & P. Resnik (eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, (pp. 1-26). Cambridge, MA: MIT Press.

Altmann, G. T. M. (1998). Ambiguity in sentence processing. *Trends in Cognitive Sciences*, 2, 146-152.

Anderson, J. R. (1990). *The Adaptive Character of Thought. Hillsdale*, NJ: Lawrence Erlbaum Associates.

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72, 32-68.

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3, 1-8.

Brants, T., & Crocker, M. W. (2000). Probabilistic parsing and psychological plausibility. In *Proceedings of the 18th International Conference on Computational Linguistics*, Saarbruecken/Luxembourg/Nancy.

Briscoe, T., & Carroll, J. (1993). Generalised probabilistic LR parsing for unification-based grammars. *Computational Linguistics*, 19, 25-60.

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

Burnard, L. (1995). *Users Guide for the British National Corpus*. British National Corpus Consortium, Oxford University Computing Service.

Carroll, G., & Rooth, M. (1998). Valence induction with a head-lexicalized PCFG. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (pp. 36-45), Granada.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, (pp. 132-139), Seattle, WA.

Chater, N., Crocker, M., & Pickering, M. (1998). The rational analysis of inquiry: The case for parsing. In N. Chater, & M. Oaksford (eds.), *Rational Models of Cognition*, (pp. 441-468). Oxford: Oxford University Press.

Chomsky, N. (1955). *The logical structure of linguistic theory*. Manuscript, Harvard University and MIT. Published as Chomsky (1975).

Chomsky, N. (1964). Degrees of grammaticalness. In J. A. Fodor, & J. J. Katz (eds.), *The Structure of Language: Readings in the Philosophy of Language*, (pp. 384-389). Englewood Cliffs, NJ: Prentice-Hall.

Chomsky, N. (1975). *The Logical Structure of Linguistic Theory*. New York: Plenum Press.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Christiansen, M. H., & Chater, N. (1999). Connectionist natural language processing: The state of the art. *Cognitive Science*, 23, 417-437.

Christiansen, M. H., & Chater, N. (2001). Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences*, 5, 82-88.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Connine, C. M., Ferreira, F., Jones, C., Clifton, C., & Frazier, L. (1984). Verb frame preferences: Descriptive norms. *Journal of Psycholinguistic Research*, 13, 307-319.

Corley, S., & Crocker, M. (2000). The modular statistical hypothesis: Exploring lexical category ambiguity. In M. Crocker, M. Pickering, & C. Clifton (eds.), *Architectures and Mechanisms for Language Processing*. Cambridge: Cambridge University Press.

Crocker, M. (1999). Mechanisms for sentence processing. In S. Garrod, & M. Pickering (eds.), *Language Processing*. Psychology Press, London.

Crocker, M. (to appear). Rational models of comprehension: Addressing the performance paradox. In A. Cutler (ed.), *Twenty-First Century Psycholinguistics: Four Cornerstones*. Hillsdale: Lawrence Erlbaum.

Crocker, M., & Corley, S. (2002). Modular architectures and statistical mechanisms: The case from lexical category disambiguation. In P. Merlo, & S. Stevenson (eds.), *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*, (pp. 157-180). Amsterdam: John Bejamins.

Crocker, M. W. (1996). *Computational Psycholinguistics: An Interdisciplinary Approach to the Study of Language*. Dordrecht: Kluwer.

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

Cuetos, F., & Mitchell, D. C. (1988). Cross-linguistic differences in parsing: Restrictions on the late closure strategy in Spanish. *Cognition*, 30, 73-105.

Cuetos, F., Mitchell, D. C., & Corley, M. M. B. (1996). Parsing in different languages. In M. Carreiras, J. Garcia-Albea, & N. Sabastian-Galles (eds.), *Language Processing in Spanish*, (pp. 145-189). Mahwah, NJ: Lawrence Erlbaum Associates.

Culy, C. (1998). Statistical distribution and the grammatical/ungrammatical distinction. *Grammars*, 1, 1-19.

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, 27, 429-446.

Elman, J. L. (1991). Distributed reprentations, simple recurrent networks and grammatical structure. *Machine Learning*, 9, 195-225.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48, 71-99.

Featherston, S. (2004). The decathlon model of empirical syntax. In S. Kepser, & M. Reis (eds.), *Linguistic Evidence*. Berlin: Mouton de Gruyter.

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

Fodor, J. D., & Frazier, L. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, 6, 291-325.

Ford, M., Bresnan, J., & Kaplan, R. M. (1982). A competence-based theory of syntactic closure. In J. Bresnan (ed.), *The Mental Representation of Grammatical Relations*, (pp. 727-796). Cambridge, MA: MIT Press.

Francis, N., Kucera, H., & Mackie, A. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Frazier, L., & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language*, 26, 505-526.

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

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68, 1-76.

Gibson, E., Pearlmutter, N., Canseco-Gonzalez, E., & Hickok, G. (1996a). Crosslinguistic attachment preferences: Evidence from English and Spanish. *Cognition*, 59, 23-59.

Gibson, E., & Pearlmutter, N. J. (1998). Constraints on sentence comprehension. *Trends in Cognitive Sciences*, 2, 262-268.

Gibson, E., & Schuetze, C. T. (1999). Disambiguation preferences in noun phrase conjunction do not mirror corpus frequency. *Journal of Memory and Language*, 40, 263-279.

Gibson, E., Schuetze, C. T., & Salomon, A. (1996b). The relationship between the frequency and the processing complexity of linguistic structure. *Journal of Psycholinguistic Research*, 25, 59-92.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics*, 28, 267-283.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA.

Hale, J. (2003). The information conveyed by words. *Journal of Psycholinguistic Research*, 32, 101-122.

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

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

Jurafsky, D. (2002). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (eds.), *Probabilistic Linguistics*. MIT Press.

Keller, F. (2000a). Evaluating competition-based models of word order. In L. R. Gleitman, & A. K. Joshi (eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, (pp. 747-752), Mahwah, NJ. Lawrence Erlbaum Associates.

Keller, F. (2000b). *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. Ph.D. thesis, University of Edinburgh.

Keller, F. (2003). A probabilistic parser as a model of global processing difficulty. In R. Alterman, & D. Kirsh (eds*.), Proceedings of the 25th Annual Conference of the Cognitive Science Society*, (pp. 646-651), Boston.

Keller, F., & Asudeh, A. (2002). Probabilistic learning algorithms and Optimality Theory. *Linguistic Inquiry*, 33, 225-244.

Kempen, G., & Harbusch, K. (2004). Why grammaticality judgments allow more word order freedom than speaking and writing: A corpus study into argument linearization in the midfield of German subordinate clauses. In S. Kepser, & M. Reis (eds.), *Linguistic Evidence*. Berlin: Mouton de Gruyter.

Lapata, M., Keller, F., & Schulte im Walde, S. (2001). Verb frame frequency as a predictor of verb bias. *Journal of Psycholinguistic Research*, 30, 419-435.

MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language*, 32, 692-715.

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

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101, 676-703.

Manning, C. D. (2003). Probabilistic syntax. In R. Bod, J. Hay, & S. Jannedy (eds*.), Probabilistic Linguistic*s, (pp. 289-341). MIT Press.

Manning, C. D., & Schuetze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marcus, M. P. (1980). *A Theory of Syntactic Recognition for Natural Language*. Cambridge, MA: MIT Press.

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

Mitchell, D. C., Cuetos, F., Corley, M. M. B., & Brysbaert, M. (1996). Exposure-based models of human parsing: Evidence for the use of coarse-grained (nonlexical) statistical records. *Journal of Psycholinguistic Research*, 24, 469-488.

Narayanan, S., & Jurafsky, D. (1998). Bayesian models of human sentence processing. In M. A. Gernsbacher, & S. J. Derry (eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, Mahwah, NJ. Lawrence Erlbaum Associates.

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

Pritchett, B. L. (1992). *Grammatical Competence and Parsing Performance*. Chicago: University of Chicago Press.

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

Riezler, S. (1996). Quantitative constraint logic programming for weighted grammar applications. In *Proceedings of the 1st Conference on Logical Aspects of Computational Linguistics*. Berlin: Springer.

Roland, D., & Jurafsky, D. (1998). How verb subcategorization frequencies are affected by corpus choice. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, (pp. 1122-1128), Montreal.

Roland, D., & Jurafsky, D. (2002). Verb sense and verb subcategorization probabilities. In P. Merlo, & S. Stevenson (eds.), *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*, (pp. 325-346). Amsterdam: John Bejamins.

Skut, W., Krenn, B., Brants, T., & Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, DC.

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

Sturt, P., Pickering, M. J., & Crocker, M. W. (1999). Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, 40, 136-150.

Sturt, P., Pickering, M. J., Scheepers, C., & Crocker, M. W. (2001). The preservation of structure in language comprehension: Is reanalysis the last resort? *Journal of Memory and Language*, 45, 283-307.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632-1634.

Tanenhaus, M. K., Spivey-Knowlton, M. J., & Hanna, J. E. (2000). Modelling discourse context effects: A multiple constraints approach. In M. Crocker, M. Pickering, & C. Clifton (eds.), *Architectures and Mechanisms for Language Processing*, (pp. 90-118). Cambridge: Cambridge University Press.

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

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

Trueswell, J. C., & Tanenhaus, M. K. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In C. Clifton, L. Frazier, & K. Rayner (eds.), *Perspectives on Sentence Processing*, (pp. 155-179). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Processing*, 13, 260-269.

Footnotes

[1] The authors would like to thank the volume editors, the anonymous reviewers, and also Marshall Mayberry for their helpful comments. The authors gratefully acknowledge the support of the German Research Foundation (DFG) through SFB 378 Project ‘Alpha’ awarded to the first author, and an Emmy Noether fellowship awarded to the second author.

Figures

Figure 1:

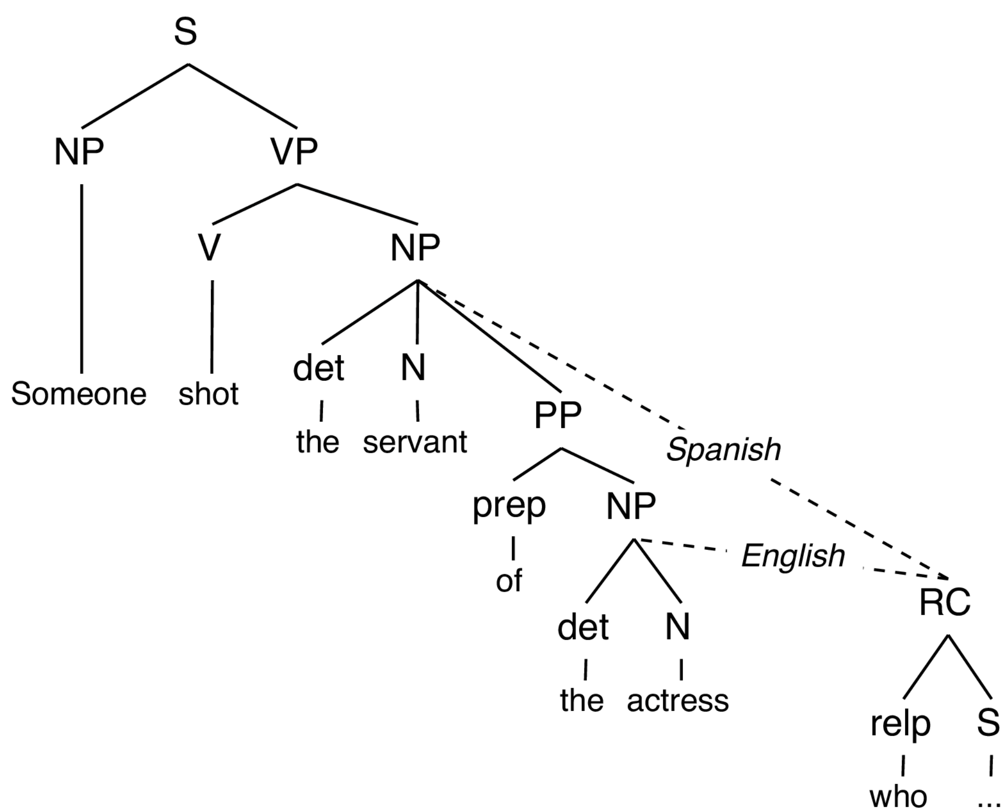

Figure 2a:

S --> NP VP	1.0	NP --> Det NP	0.6	V --> hit	1.0
PP --> P NP	1.0	NP --> NP PP	0.2	N --> man	0.5
VP --> V NP	0.7	NP --> John	0.2	N --> book	0.5
VP --> VP PP	0.3	P --> with	1.0	Det --> the	1.0

Figure 2b:

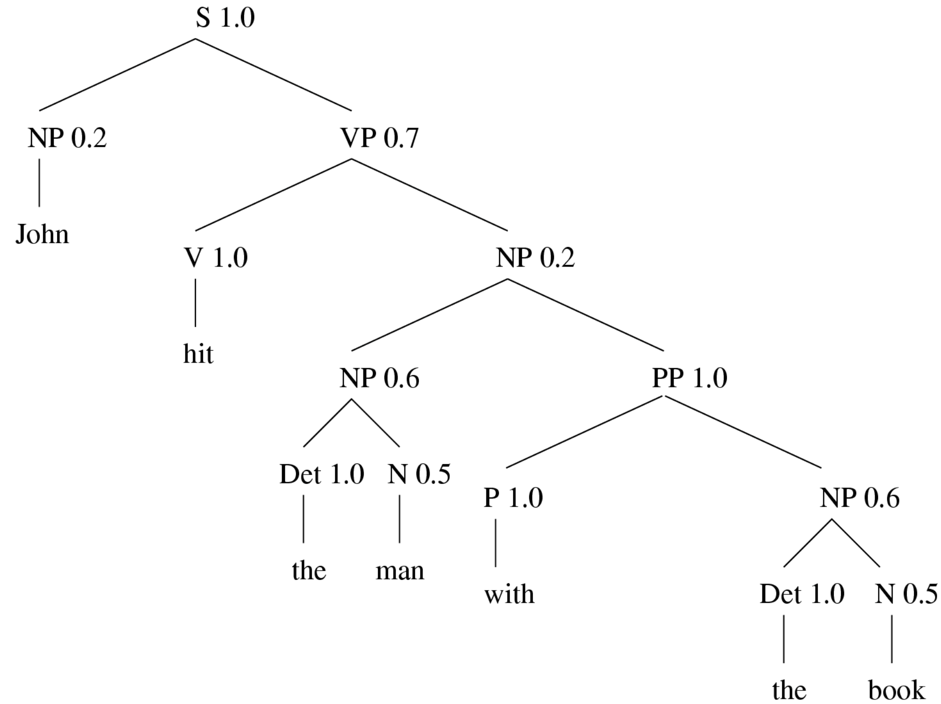

Figure 2c:

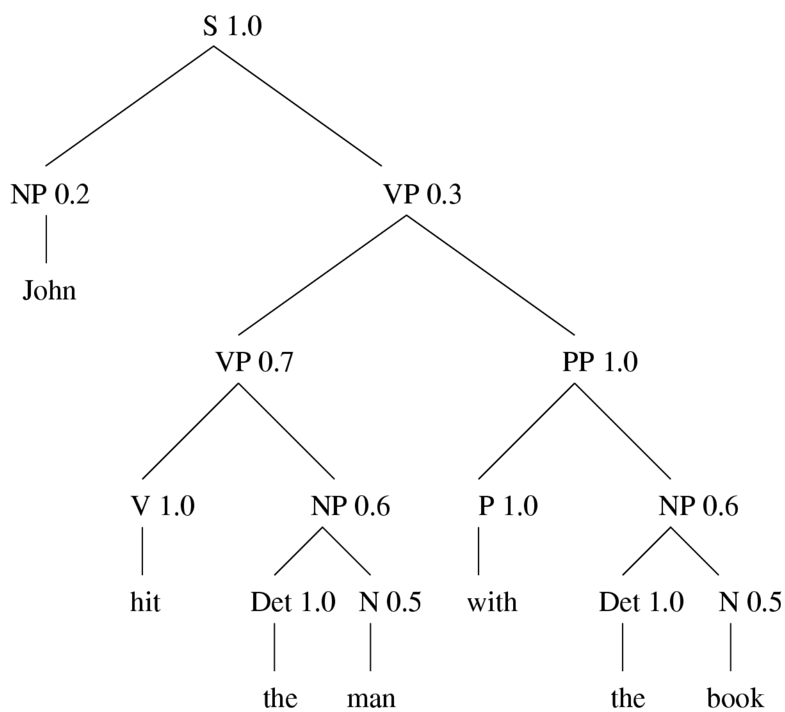

Figure Captions

Figure 1: Evidence from relative clause (RC) attachment ambiguity has been taken to support an experience-based treatment of structural disambiguation. Such constructions are interesting because they do not hinge on lexical preferences. When reading sentences containing the ambiguity depicted above, English subjects demonstrate a preference for low-attachment (where **the actress** will be further described by the RC *who* . . .), while Spanish subjects, presented with equivalent Spanish sentences, prefer high-attachment (where the RC concerns *the servant*) (Cuetos and Mitchell 1988). The **tuning hypothesis** was proposed to account for these findings (Brysbaert and Mitchell 1996, Mitchell et al. 1996), claiming that initial attachment preferences should be resolved according to the more frequent structural configuration. Later experiments further tested the hypothesis, examining subjects' preferences before and after a period of two weeks in which exposure to high or low examples was increased. The findings confirmed that even this brief period of variation in 'experience' influenced the attachment preferences as predicted (Cuetos et al. 1996).

Figure 2: An example for the parse trees generated by a probabilistic context free grammar (PCFG). (a) The rules of a simple PCFG with associated rule application probabilities. (b) and (c) The two parse trees generated by the PCFG in (a) for the sentence *John hit the man with the book*.