To appear in 'Probabilistic Linguistics', edited by Rens Bod, Jennifer Hay, and Stefanie Jannedy, MIT Press, 2002

Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production

Dan Jurafsky

Probability is not really about numbers; it is about the structure of reasoning
Glenn Shafer, cited in Pearl (1988)

1 Introduction

It must certainly be accounted a paradox that probabilistic modeling is simultaneously one of the oldest and one of the newest areas in psycholinguistics. Much research in linguistics and psycholinguistics in the 1950s was statistical and probabilistic. But this research disappeared throughout the 60's, 70's, and 80's. In a highly unscientific survey (conducted by myself) of six college textbooks and handbooks in psycholinguistics published in the last 10 years, not a single one of them mentions the word 'probability' in the index.

This omission is astonishing when we consider that the input to language comprehension is noisy, ambiguous, and unsegmented. In order to deal with these problems, computational models of speech processing, by contrast, have had to rely on probabilistic models for over 30 years. Computational techniques for processing of text, an input medium which is much less noisy than speech, rely just as heavily on probability theory. Just to pick an arbitrary indicator, 77% of the papers in the year 2000 annual conference of the Association for Computational Linguistics relied on probabilistic models of language processing or learning.

Probability theory is certainly the best normative model for solving problems of decision-making under uncertainty. But perhaps it is a good normative model, but a bad descriptive one. Despite the fact that probability theory was originally invented as a cognitive model of human reasoning under uncertainty, perhaps people do not use probabilistic reasoning in cognitive tasks like language production and comprehension. Perhaps human language processing is simply a non-optimal, non-rational process?

In the last decade or so, there is an emerging consensus that human cognition is in fact rational, and relies on probabilistic processing. The seminal work of Anderson (1990) gave Bayesian underpinnings to cognitive models of memory, categorization, and causation. Probabilistic models have cropped up in many areas of cognition; one area in which there are a number of recent probabilistic models is categorization (Rehder 1999; Glymour and Cheng 1998; Tenenbaum 2000; Tenenbaum and Griffiths 2001b; Tenenbaum and Griffiths 2001a),

Probabilistic models are also now finally being applied in psycholinguistics, drawing from early Bayesian-esque precursers in perception such as the Luce (1959) choice rule and the work of Massaro. What does it mean to claim that human language processing is probabilistic? This claim has implications for language comprehension, production and learning.

Three claims have been made for language comprehension. Consider the task of accessing linguistic structure from the mental lexicon or grammar. Perhaps more probable structures are accessed more quickly, or with less effort. Or perhaps they can merely be accessed with less evidence than less probable structures. Another role for probability in comprehension is in disambiguation. Ambiguity is ubiquitous in language comprehension; speech input is ambiguously segmented, words are syntactically and semantically ambiguous, sentences are syntactically ambiguous, utterances have ambiguous illocutionary force, and so on. Probability is one of the factors that play a role in disambiguation; the more probable an interpretation, the more likely it is to be chosen. Probability may also play a key role in comprehension in explaining processing difficulty. Recent models of what makes certain sentences difficult to process are based, at least in part, on certain interpretations having particularly low probabilities, or on sudden switches of probabilistic preference between alternate interpretations. In this chapter I will summarize models of all three of these roles for probability in comprehension: access, disambiguation, and processing difficulty.

The claim that human language processing is probabilistic also has implications for production. Probability may play a role in accessing structures from the mental lexicon or grammar. High-probability structures may be accessed faster, or more easily, or simply with more confidence. Disambiguation, itself a phenomenon of comprehension, has a correlate in production: choice. Given multiple possible structures to choose to say, probability may play a role in selecting among these. I will give an overview of the experimental and modeling literature on probabilistic production, although there is significantly less to summarize here than there is for comprehension.

A third role for probability is in learning. Many models of how linguistic structure is empirically induced rely on probabilistic and information-theoretic models. This chapter will not focus on learning; the interested reader should turn to the literature; relevant papers include Brent and Cartwright (1996); Saffran *et al.* (1996c); Saffran *et al.* (1996a); Saffran *et al.* (1996b); Saffran (2001) and Tenenbaum and Xu (2000).

What probabilistic modeling offers psycholinguistics is a model of the structure of evidential reasoning: a principled and well-understood algorithm for weighing and combining evidence to chose interpretations in comprehension, and to chose certain outcomes in production. Throughout the chapter, I will make reference to Bayesian reasoning, and in particular to Bayes' rule (equation (9) in Chapter 1). Bayes' rule gives us a way to break down complex probabilities into ones that are easier to operationalize and compute. Suppose we are trying to compute the probability of some interpretation i given some evidence e. Bayes' rule states that this probability can be broken down as follows:

$$P(i|e) = \frac{P(e|i)P(i)}{P(e)} \tag{1}$$

This says that we can compute the probability of an interpretation i given evidence e by instead asking how likely the interpretation i is a priori, and how likely the evidence e would be to occur if we knew the interpretation was correct. Both of these are often easier to compute than $P(i|e)$.

Probabilistic modeling has been applied to many areas of psycholinguistics; phonological processing, morphological processing, lexical processing, syntactic processing, discourse processing. I will focus in this chapter on lexical and syntactic processing. The reader should be warned that this survey will include some extra gory details on my own work and work to which I have contributed. As for other areas of processing, Baayen (this volume) covers processing with respect to morphology and Pierrehumbert (this volume) touches on processing issues in phonology. I have left out probabilistic work on dialogue and discourse processing because some of it is covered in Jurafsky (2001).

2 Summary of Evidence for Probabilistic Knowledge

This section will summarize evidence from psycholinguistic experimentation that bears on the use of frequency-based or probabilistic knowledge in human language processing. The focus will be on lexical and syntactic processing, both in comprehension and production.

What kinds of experiments provide evidence for probabilistic modeling? First, I will summarize many experiments which show that frequencies of various linguistic structures play a role in processing; frequencies of words, of word pairs, of lexical categories of words, of subcategorization relations, etc. Why should finding evidence for frequencies support probabilistic models?

One reason is that relative frequency can play the role of prior probability in the computation of conditional probability. Recall from Bayes' rule that the probability $P(i|e)$ of some structure or interpretation i given some evidence e can be computed as follows:

$$P(i|e) = \frac{P(e|i)P(i)}{P(e)} \tag{2}$$

This means that the conditional probability of an interpretation or structure i is directly related to the prior probability of i. Since the relative frequency of i is an easy way to come up with an estimate of the prior probability of i, the Bayesian model predicts that we should find frequency effects for various kinds of structures.

But many complex structures are too rare for their probability to be computed by counting the number of times they have occurred. Language is creative, after all, and many large structures (like sentences) may only occur once. In these cases we would not expect to see evidence for the frequency of these unique events.

In just these cases, however, probabilistic modeling gives us tools to estimate the prior probability of these structures by making various independence assumptions, allowing us to estimate the probability of one large complex object from the counts of many smaller objects. This, for example, is the goal of probabilistic grammar formalisms like DOP and stochastic context-free grammars. Since these models compute larger structures and their probabilities by combining smaller structure and their probabilities, probabilistic modeling suggests that the frequency of these various kinds of smaller more primitive structures should play a role in processing. This once again predicts that we should find effects of various frequencies in processing. We will explore some of these assumptions in more detail in Section 3.

In addition to evidence for frequencies, I will also summarize evidence for various kinds of conditional probabilities. In general, I will define the probabilities of interest as the experiments are introduced.

2.1 Lexical Frequency

As we will see, word frequency plays a robust effect in lexical comprehension and production. But before summarizing these results, it's important to know how lexical frequency is measured.

Most studies since 1970 have relied on word frequency statistics calculated from the Brown corpus of American English, a 1 million word collection of samples from 500 written texts from different genres (newspaper, novels, non-fiction, academic, etc.), which was assembled at Brown University in 1963–64. Kučera and Francis (1967) reports the frequency for each wordform in the corpus, while Francis and Kučera (1982) uses a lemmatized and part-of-speech tagged version of the Brown corpus to report frequencies for lemmas (e.g. reporting combined as well as distinct frequencies for *go*, *goes*, *going*, *went*, and *gone*, and distinct frequencies for e.g. *table* the verb and *table* the noun).

From the very beginning of the field, it was clear that deriving frequencies from corpora in this way was not unproblematic. It might seem astonishing that the wide variety of frequency effects reported in the literature are based on using this one corpus. Indeed, the use of corpora like the Brown corpus to derive frequencies, either as a control factor, or as an explicit part of a probabilistic model, is problematic in a number of ways. First, consider that a corpus is an instance of language production, but the frequencies derived from corpora are often used to model or control experiments in comprehension. While comprehension and production frequencies are presumably highly correlated, there is no reason to expect them to be identical. Second, the Brown corpus is a genre-stratified corpus. It contains equal amounts of material from newspapers, fiction, academic prose, etc. But presumably a corpus designed for psychological modeling of frequency would want to model the frequency with which an individual hearer or speaker is exposed to (or uses) linguistic input. This would require a much larger focus on spoken language, on news broadcasts, and on magazines. Third, the Brown corpus dates from 1961; most subjects in psycholinguistics experiments run in 2001 are college undergraduates and weren't even born in 1961; the frequencies that would be appropriate to model their language capacity may differ widely from Brown frequencies.

I see these problems as introducing a very strong bias *against* finding any effects of corpus frequencies in experimental materials. Nonetheless, as we will see, strong and robust effects of corpus frequencies have been found. The first reason for this is that studies have shown for over 50 years that frequencies from different corpora are very highly correlated (Howes and Solomon 1951). But a more important reason is that most studies report only very broad-grained frequencies, often using just three bins; high frequency, low frequency, and other). Finally, studies are beginning to use larger and more recent corpora and databases such as the CELEX lexical database (based on a corpus of 18 million words) (Baayen *et al.* 1995) and the British National Corpus (which has roughly 10 million tokens of tagged spoken English and 90 million tokens of written English). These corpora are large enough to allow for the direct use of unbinned frequencies in recent work such as De Jong *et al.* (2001), Allegre and Gordon (1999), and Baayen *et al.* (in press).

Lexical Frequency in Comprehension

One of the earliest and most robust effects in psycholinguistics is the word frequency effect. Word frequency plays a role in both the auditory and visual modalities, and in both comprehension and production.

The earliest work seems to have been by Howes and Solomon (1951), who used a tachistoscope to display a word for iteratively longer and longer durations. They showed that the log frequency of a word (as computed from corpora of over four million words) correlated highly with the mean time to recognize the word; more frequent words were recognized with shorter presentations. Later, the *naming* paradigm, in which subjects read a word out loud, was used to show that high-frequency words are named more rapidly than low-frequency words (Forster and Chambers 1973). The lexical decision paradigm, in which subjects decide if a string of letters presented visually is a word or not, has also been used. Lexical decisions about high-frequency words are made faster than decisions about low-frequency words (Rubenstein *et al*. 1970; Whaley 1978; Balota and Chumbley 1984) Again, these results are robust and have been widely replicated. Frequency also plays a role in more on-line reading measures such as fixation duration and gaze duration.

Similarly robust results have been found for auditory word recognition. Howes (1957) first found results with speech that were similar to his earlier visual results. He presented subjects with high and low-frequency words which were immersed in noise. Subjects were better at identifying high-frequency words than low ones. In an extension of this experiment, Savin (1963) further found that when subjects made recognition errors, they responded with words that were higher in frequency than the words that were presented. Grosjean (1980) used the gating paradigm, in which subjects hear iteratively more and more of the waveform of a spoken word, to show that high-frequency words were recognized earlier (i.e. given less of the speech waveform) than low-frequency words. Tyler (1984) showed the same result for Dutch.

In conclusion, the evidence shows that in both the visual and auditory domains, high-frequency words are accessed more quickly, more easily, and with less input signal than low-frequency words.

Lexical Frequency in Production

The effects of lexical frequency on production have been measured via a number of tests, including *latency* (the time to start producing a word), *duration* (the time from word onset to word offset), *phonological reduction* (number of deleted or reduced phonemes), rate of speech errors, and others.

The earliest studies were on duration; indeed lexical frequency effects in production on duration have been remarked upon for over a hundred years. Schuchardt (1885) noticed, for example, that more frequent words tended to be shorter. Fidelholz (1975) and Hooper (1976) showed for example that frequent words such as *forget* are more likely to have lexically reduced vowels (e.g. [fər]) than less frequent words such as *forgo* (e.g. [fɔr]) (Table 1).

Reduced Vowel [fər]		Full Vowel [fɔr]	
Word	Count per Million	Word	Count per Million
forget	148	forfend	<1
forgive	40	forgo	<1

Table 1: Lexically reduced vowels in high-frequency words (after Fidelholz (1975)).

While these early studies showing an effect of frequency on a word's phonological make-up are suggestive, they do not confirm that the effect of frequency on lexical production is on-line and productive. It could be that frequent words have reduced vowels and fewer phonemes due to some diachronic fact staticly reflected in the lexicon that is only related to on-line production in a complex and indirect way.

To show that frequency is playing an active and on-line role in language production, it is necessary to examine the effect of frequency on some dynamic process. One such process is phonological variation; thus a number of studies have examined whether frequency dynamically affects variation in production. Bybee (2000) examined word-final /t/ and /d/ in a corpus of spoken Chicano English. She first excluded the extremely high-frequency words *just*, *went*, and *and*, and then classified the remaining 2000 word tokens into two bins, high-frequency (defined as greater than 35 per million in the Brown corpus) and low-frequency (less than 35 per million). She showed that final /t/ and /d/ deletion rates in high frequency words (54.5%) were greater than deletion rates in low frequency words (34.3%). Hay (2000) shows that for complex words the ratio of the frequency of the derived word and the frequency of its base is an important predictor.

Gregory *et al.* (1999) and Jurafsky *et al.* (2001) provided further evidence that these frequency effects on reduction are on-line, by controlling for a wide variety of contextual factors such as rate of speech, and also by investigating the effect of frequency on a word's duration, in addition to its phonological reduction. They examined the duration and the percentage of final-consonant deletion of words in a 38,000-word phonetically-transcribed subcorpus from the Switchboard corpus of American English telephone conversations (Godfrey *et al.* 1992; Greenberg *et al.* 1996). They used multiple regression to control for factors like segmental context, rate of speech, number of phones, word predictability, etc.

They first confirmed Bybee's results by analyzing 2042 word tokens whose full pronunciation ended in /t/ or /d/. After controlling for the contextual factors mentioned above, they found that these final obstruents are more likely to be deleted in more frequent words. High frequency words (at the 95th percentile) were 2.0 times more likely to have deleted final /t/ or /d/ than the lowest frequency words (at the 5th percentile).

Gregory *et al.* (1999) and Jurafsky *et al.* (2001) also investigated the effects of frequency on word duration, using 1412 monosyllabic word tokens ending in /t/ or /d/. They found a strong effect of word frequency on duration. Overall, high frequency words (at the 95th percentile of frequency) were 18% shorter than low frequency words (at the 5th percentile).

Taken together, these results suggest that frequency plays an on-line role in lexical production. Duration studies, however, may not be completely convincing. It is possible, for example, that high-frequency words are simply stored with multiple phonological lexemes (Jurafsky *et al.* 2001 (to appear)), or are stored with very detailed phonetic information about the length of each phone in each word (Pierrehumbert 2001).

The most unambiguous evidence for frequency effects in production, then, must come from latency. Oldfield and Wingfield (1965), for example, showed an on-line effect of word frequency on latency of picture naming times. They presented subjects with pictures and found that pictures with high frequency names were named faster than pictures with low frequency names. Wingfield (1968) showed that this effect must be caused by word frequency rather than the frequency of pictured objects, by showing that the effect was not replicated when subjects were asked to recognize but not verbalize picture names. These results were also replicated for Dutch by Jescheniak and Levelt (1994).

In conclusion, more frequent words are accessed more quickly (shorter latency) and are articulated more quickly (shorter duration).

2.2 Frequency of Lexical Semantic Form and Lexical Category

Words are ambiguous in many ways. A word can have multiple senses (*bank* can refer to a location alongside a river or a financial institution), multiple lexical categories (*table* can be a noun or a verb), and multiple morphological categories (*searched* can be a participle or a preterite). These different categories of an ambiguous word vary in frequency. For example the word *table* is more likely to be a noun than a verb. In this section I summarize experiments showing that the frequency of these various categories for an ambiguous word play a role in processing.

A number of experiments have shown that the frequency of a particular sense of an ambiguous word plays a role in comprehension. Simpson and Burgess (1985), for example, studied lexical access in the visual domain. Subjects were first presented with an ambiguous prime word (homograph) that had a more frequent sense and a less frequent sense. Subjects then performed lexical decision on targets that were associated with either the more frequent or the less frequent meaning of the homograph prime. They found that the more frequent meaning of the homograph caused faster response latencies to related associates, suggesting that the more frequent meaning is retrieved more quickly. This result is robust, and has been replicated with many paradigms, including eye fixation times in reading and cross-modal priming. Evidence for the use of word sense frequency in comprehension has also been reported cross-linguistically, for example in Chinese (Li and Yip 1996; Ahrens 1998).

The frequency of the syntactic category of an ambiguous word plays a role in comprehension as well. One class of studies has come from the literature on sentence processing and human parsing. Gibson (1991) and Jurafsky (1992,1996) suggest that lexical category frequencies might be playing a role in the processing difficulty of some garden path sentences. For example (3) and (4) are known to be difficult to process. Gibson suggests that the difficulty of (3) is due to the *man* being much more likely to be a verb than a noun, while Jurafsky suggests that the difficulty of (4) is due to the lexical category preferences of *complex* (more likely to be an adjective than a noun), and *house* (more likely to be a noun than a verb).

(3) The old man the boats. (From Milne (1982))

(4) The complex houses married and single students and their families. (From Jurafsky (1992,1996))

The third category of lexical frequencies is morphological category frequencies. Words such as *searched*, *scratched*, *proposed*, and *selected* are ambiguous between a participle and a preterite (simple past) reading. For some of these words, the participle reading has a higher frequency. For example the percentage of participle readings of *selected* (in the Brown corpus) is 89%, compared to a 11% percentage for the simple past reading. By contrast the preferences are reversed for *searched*, which has a 78% simple past readings, compared to a 22% participle readings. Burgess and Hollbach (1988) suggested that these lexical category probabilities might play a role in disambiguation.

Trueswell (1996) investigated this hypothesis by embedding these verbs in sentences which have a local ambiguity. Each sentence has an initial word sequence like *The room searched* which is syntactically ambiguous between a relative clause reading (compatible with the participle form) and a main-verb reading (compatible with the simple past). Trueswell found that verbs with a frequency-based preference for the simple past form caused readers to prefer the main clause interpretation (as measured by longer reading time for a sentence which required the other interpretation such as (5)):

(5) The room searched by the police contained the missing weapon.

This suggests that the frequency with which the different morphological categories of a verb occur plays a role in whether one syntactic parse is preferred or not.

In summary, the frequencies of the semantic, syntactic, or morphological categories associated with an ambiguous word play an important role in comprehension. More frequent categories are accessed more quickly and are preferred in disambiguation.

Rather surprisingly, given this robust effect of the frequency of lexical semantic/syntactic category in comprehension, there may not be any such effect in production. Instead, some studies have suggested that frequency effects in lexical production are confined to the level of the wordform or lexeme, rather than the semantic/syntactically defined lemma level.

Both Dell (1990) and Jescheniak and Levelt (1994), for example, studied whether word frequency effects in production took place at the level of the semantic *lemma* or the phonological *wordform*. Finding an effect of frequency for the semantic/syntactic lemma would be the correlate in lexical production of finding an effect of semantic sense or syntactic category in comprehension. Dell (1990) used experimentally elicited speech errors to study word frequency effects. Previous work had shown that low frequency words are more susceptible to phonological speech errors than high frequency words. Dell showed that some low frequency words are not susceptible to phonological speech errors; specifically, low frequency words (such as *wee*) with a high frequency homophone (such as *we*). In other words, a low-frequency word which shares a lexeme with a high-frequency word exhibits some of the frequency properties of the high-frequency word. One way to model this result is to store frequency effects only at the lexeme level; the words *we* and *wee* would then share a single frequency node.

Jescheniak and Levelt (1994) used a novel translation task to study word frequency effects. Once again, they looked at homophones, in which two distinct lemmas share one lexeme. If frequency effects are localized at the lemma level, they would expect that accessing a low-frequency lemma in production would have slower latency than a high-frequency lemma. If frequency effects are localized at the lexeme level, they would expect that low-frequency and high-frequency lemmas of the same homophone should have identical latencies. Testing this task is not possible with standard paradigms like picture naming, since it is unlikely that both the high-frequency and low-frequency senses of a word are picturable (e.g., neither *we* nor *wee* are obviously picturable). They therefore used a novel translation latency task: bilingual Dutch subjects were ask to produce the Dutch translation for a visually presented English word, and the translation latency was recorded. For example, subjects saw the English word *bunch*, whose Dutch translation is *bos*. The Dutch word *bos* has another sense, *forest*. If frequencies are stored at the lexeme level, latencies to low-frequency words like *bunch*/*bos* should act like high-frequency words. This is what Jescheniak and Levelt (1994) found. Latency to homophones patterned like latency to high-frequency words, and not like latency to low-frequency words.

There is one important caveat to the results from Dell (1990) and Jescheniak and Levelt (1994): they crucially rely on the assumption that lexical production is modular. If lexical production is completely interactionist, frequencies could be stored at the lemma level but activation could spread from the lemma, down to the lexeme, and back up to both lemmas, allowing a low-frequency lemma to act like a high-frequency one. In fact, this is exactly Dell's (1990) proposal, and he built a non-modular computational model to show that the idea is possible. The evidence for a lack of a lemma effect, then, rests only on the evidence for a modular (non-interactionist) model of lexical production.

In order to help resolve this dilemma, Jurafsky *et al.* (2001 (to appear)) proposed a different, corpus-based methodology, to study frequency effects in production. They examined the production of ambiguous words like *to* (which can be an infinitive marker (*we had* **to** *do it*) and a preposition (*I would have gone* **to** *the store*), or *that* (which can be (at least) a complementizer, a pronoun, or a determiner). Again using the 38,000-word phonetically-transcribed subcorpus from the Switchboard corpus of American English telephone conversations, they measured the duration of the function words. They then used multiple regression to control for known factors affecting duration, including rate of speech, segmental context, contextual predictability and so on, and then test for an effect of lemma frequency on duration. They found that the different surface pronunciations and durations of these words could be completely accounted for by other factors such as pitch accent and contextual predictability. They found no evidence that lemma frequency affected lexical production.

Thus, although the frequencies of the semantic, syntactic, or morphological categories associated with an ambiguous word play a role in comprehension, preliminary studies suggest that they may not play a similar role in production.

2.3 Neighboring word-to-word probabilities

The previous sections focused on frequency effects for single words. We now turn to evidence that frequency plays a role for more complex and structured relationships between words and syntactic structure. The simplest such relationship is the well-known first-order Markov relationship; the probability of a word given its neighbors. A number of studies show that the probabilistic relationships between neighboring words play a role in both comprehension and production.

Some of these studies looked at raw frequency, while others looked at various probabilistic measures. Researchers have investigated both the conditional probability of a word given the previous word $P(w_i|w_{i-1})$ and the joint probability of two words together $P(w_{i-1}w_i)$. The *joint probability* of the two words is generally estimated from the relative frequency of the two words together in a corpus, normalized by the total number N of word-pair tokens in the corpus (which is one more than the total number of words in the corpus):

$$P(w_{i-1}w_i) = \frac{Count(w_{i-1}w_i)}{N} \tag{6}$$

Some experiments use this normalized joint probability, while others just use the raw joint frequency.

Another common metric is the first-order Markov relation; the *conditional probability of a word given the previous word* (sometimes called the *transitional probability* (Saffran *et al.* 1996c; Bush 1999)). The conditional probability of a particular target word w_i given a previous word w_{i-1} can be estimated from the counts of the number of times the two words occur together $Count(w_{i-1}w_i)$, divided by $Count(w_{i-1})$, the number of total times that the first word occurs:

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})} \tag{7}$$

MacDonald (1993) studied the effect of word-pair (joint) frequencies on comprehension. MacDonald was looking at the processing of a noun followed by a word that is ambiguous between a noun and verb, such as the pair *miracle cures*. MacDonald hypothesized that if the noun-noun pair was frequent (like *miracle cures*), it would bias the interpretation toward the noun reading of the second word. She predicted no such bias for infrequent noun-noun pairs (like *shrine cures*). She confirmed this hypothesis by looking at reading time just after the ambiguous word in sentences that were otherwise biased toward a verb reading.

For example, she found that subjects spent longer reading the word *people* in (9) than (8), since the frequent noun-noun phrase (9) biases the reader toward the noun reading of *cure*, and the word *people* is only compatible with the verb reading.

(8) The doctor refused to believe that the *shrine cures* people of many fatal diseases...

(9) The doctor refused to believe that the *miracle cures* people of many fatal diseases...

This study has been extended recently. McDonald *et al.* (2001) showed in a eye-tracking study that bigram probability $P(w_i|w_{i-1})$ was a good predictor of gaze duration on word w_i. Bod (2000) and (2001) showed in a recognition task that this extends to larger structures than bigrams. Bod found that frequent three-word (subject-verb-object) sentences are more easily and faster recognized than infrequent three-word sentences.

The psycholinguistic role of word-to-word frequencies or probabilities has also been extensively studied in production. The production studies have generally focused on the effects of frequency or probability given neighboring words on the phonetic form of a word. The main result is that words in high-frequency word-pairs or high-probability word pairs are phonetically reduced in some way.

Krug (1998), for example, showed that cliticization was more common in more frequent word pairs. Bybee and Scheibman (1999) and Bush (1999) showed that palatalization of word-boundary coronals is more likely between word sequences with high conditional probabilities, as shown in Table 2.

High $P(w_i|w_{i-1})$	Low $P(w_i|w_{i-1}$
More Palatalized	Less Palatalized
did you	at you
told you	but you
would you	good you

Table 2: Examples of word-boundary coronals which are more or less likely to palatalize.

Gregory *et al.* (1999), Jurafsky *et al.* (2001), and Bell *et al.* (2001) studied the effect of a number of different kinds of probability on reduction. As mentioned above, they used a phonetically transcribed portion of the Switchboard telephone corpus of American English. They used multiple regression to control for other factors affecting reduction, and looked at more measures of predictability and more measures of reduction. In particular, they looked at both the joint probability and conditional probability of target words with both previous and following words. Confirming earlier studies, they found that words which have a higher probability given neighboring words are reduced. In particular, Gregory *et al.* (1999) and Jurafsky *et al.* (2001) found that high-probability content words are durationally shorter and more likely to have final /t/ or /d/ deleted. Jurafsky *et al.* (2001), and Bell *et al.* (2001) found that high probability function words have more vowel reduction, more coda deletion, and are shorter. All of their studies found more reduction in high-probability words no matter how probability was defined (conditional probability given the previous word or given the next word; joint probability with the previous word or with the next word).

Pan and Hirschberg (2000) show that conditional bigram probability correlates highly with the location of pitch accent. Pitch accent is more likely to occur on low-probability words. Gregory (2001) extends this result by showing that conditional probability from previous and following words is a significant predictor of pitch accent even after controlling for other contextual factors like position in the intonation phrase, part of speech, number of syllables, and so on.

In summary, the probability of a word given the previous or following word plays a role in comprehension and production. Words which have a high joint or conditional probability given preceding or following words have shorter durations in production. In comprehension, if a word pair has a high frequency, any ambiguous words in that pair are likely to be disambiguated consistently with the category of word pair.

2.4 Syntactic Subcategorization Frequencies

I turn now to a kind of probability that has received quite a bit of attention. This is the frequency of the different subcategorization frames of a verb. For example, the verbs *remember* and *suspect* are both subcategorized for either a direct object noun phrase or a sentential complement, as in (10)–(13):

(10) The doctor remembered [$_{NP}$ the idea].

(11) The doctor remembered [$_{S}$ that the idea had already been proposed].

(12) The doctor suspected [$_{NP}$ the idea].

(13) The doctor suspected [$_{NP}$ that the idea would turn out not to work].

While both verbs allow both subcategorization frames, they do so with different frequencies. *Remembered* is more frequently used with a noun phrase complement, while *suspected* is more frequently used with a sentential complement. These frequencies can be computed either from a parsed or transitivity-coded corpora (Merlo 1994; Roland and Jurafsky 1998) or by asking subjects to write sentences with the verbs (Connine *et al*. 1984; Garnsey *et al*. 1997).

Since these frequencies are contingent on the verb, they are the maximum likelihood estimate of the conditional probability of the subcategorization frame given the verb $P(\text{frame}|\text{verb})$. These conditional probabilities have been shown to play a role in disambiguation. For example, the noun phrase *the idea* is ambiguous in the following prefix:

(14) The doctor suspected the idea. . .

The idea could function as a direct object noun phrase complement of *suspected*, or it could be the syntactic subject of an embedded sentential complement. A wide variety of experiments shows that the verb's "bias" (its probabilistic preference for a subcategorization frame) plays a role in which of these two continuations subjects expect.

This idea of subcategorization bias was first suggested in a slightly different context by Fodor (1978), who predicted that a verb's preference for being transitive or intransitive could affect whether the human parser hypothesizes gaps afterwards. Ford *et al*. (1982) proposed a generalization of Fodor's hypothesis: that each verb has strengths for different subcategorization frames for each verb, that these strengths are based on some combination of frequency and contextual factors, and that these strength-based expectations are used throughout parsing. They tested this idea by asking subjects in an off-line experiment to perform a forced choice between two interpretations of an ambiguous utterance, showing that some set of subcategorization strengths could be used to predict the interpretation selected by the subjects.

Ford *et al.* (1982) did not actually test whether these preferences were related to frequency in any way. Although Jurafsky (1996) later confirmed that some of their preferences corresponded to Brown corpus frequencies, this was not clear at the time, and furthermore the fact that their experiment was off-line left open the possibility that semantic plausibility or some other factor rather than subcategorization frequency was playing the causal role. Clifton, Jr. *et al.* (1984) tested the model more directly by using the frequency norms collected by Connine *et al.* (1984) to show that a frequency-based interpretation of transitivity preference predicted quicker understanding in filler-gap sentences, and Tanenhaus *et al.* (1985) showed that anomalous fronted direct-objects showed extra reading time at transitive-bias verbs, but not intransitive-bias verbs.

Trueswell *et al.* (1993) extended these results to show that these frequency-based subcategorization preferences play an on-line role in the disambiguation of various syntactic ambiguities. One experiment was based on cross-modal naming. Subjects heard a sentence prefix ending in either an S-bias verb (*The old man suspected...*) or an NP-bias verb (*The old man remembered...*). They then had to read out loud ('name') the word *him*. (Cross-modal naming is so-called because the stimulus is auditory while the target is orthographic.) Previous research had shown that naming latencies are longer when the word being read is an ungrammatical or unexpected continuation. Trueswell *et al.* (1993) showed that naming latency to *him* was longer after S-bias verbs (*The old man suspected...him*) than after NP-bias verbs (*The old man remembered...him*). This suggests that subjects preferred the more frequent frame of the verb and were surprised when this preference was overturned, causing longer naming latencies. Trueswell *et al.* (1993) also confirmed these results with an eye-tracking study which focused on the difference in reading times between sentences with and without the complementizer *that*. Controlled first-pass reading times at the disambiguating verb phrase were longer for NP-bias verbs but not for S-bias verbs, indicating that subjects attached the post-verbal noun-phrase as a direct object for NP-bias verbs but not for S-bias verbs.

MacDonald (1994) showed that the effect of subcategorization frame frequency also played a role in disambiguation of a different kind of ambiguity: main-clause/relative-clause (MC/RR) ambiguities. These ambiguities have been the object of much study since Bever (1970) first pointed out the difficulty of the garden path sentence:

(15) The horse raced past the barn fell.

Until the word 'fell', this sentence is ambiguous between a reading in which 'race' is a main verb, and one in which it is a part of a reduced relative clause modifying 'the horse'. The difficulty of the sentence is caused by the fact that readers incorrectly select the main verb sense, and then are confused upon reaching 'fell'.

MacDonald (1994) suggested that the subcategorization frequencies proposed by earlier researchers could play a role in explaining processing difficulties in main-verb/reduced-relative ambiguities. She used transitive-bias verbs like *push* and intransitive-bias verbs like *move*, in sentences like the following:

(16) The rancher could see that the nervous cattle *pushed* into the crowded pen were afraid of the cowboys.

(17) The rancher could see that the nervous cattle *moved* into the crowded pen were afraid of the cowboys.

MacDonald showed that the corrected reading times in the the disambiguation region *were afraid* were longer for intransitive-bias verbs like *move* than transitive-bias verbs like *push*.

Jennings *et al*. (1997) extended the study of Trueswell *et al*. (1993) on the effect of verb subcategorization bias on disambiguation, using a similar cross-modal naming paradigm. One of the goals of the Jennings *et al*. (1997) study was to clear up some potential problems with the Trueswell *et al*. (1993) materials. But perhaps the most significant result of the Jennings *et al*. (1997) study addressed an important issue that none of the previous research on the role of frequency in syntactic disambiguation had addressed. Previous studies had generally binned their verb-bias frequency into two bins; high transitive-bias and low transitive-bias, or high S-bias versus low S-bias. All previous results on syntactic disambiguation, then, were compatible with a model in which the representation of subcategorization preferences was as a ranked or ordered list, with no link to an actual frequency or probability. Jennings *et al*. (1997) showed a correlation between the strength of the bias of a verb and reading time at the target word. The stronger the bias of the verb for one subcategorization frame over the other, the larger the advantage they found in naming latency for the preferred over the non-preferred continuation.

Despite the many studies of subcategorization frequencies in comprehension, there are no equivalent studies in production. Of course, the frequencies used to model the comprehension studies are derived from production data. But there have been no clear tests in production that verb-argument probability is playing an active on-line role here, as opposed, say, to merely being correlated with semantics or world knowledge. There is at least one suggestive study, by Stallings *et al*. (1998), who note a similarity between sentential complement-taking verbs and heavy-NP shift; in both cases, the verb and the complement can be separated by other material. For example, sentential complements can be separated from the verb by adverbial expressions (*She said the other day that…*). By contrast, direct objects cannot be separated in this way from their verbs. They then show in a production experiment that verbs which can take either sentential complements or noun phrase direct objects are more likely to undergo Heavy-NP shift than verbs which only take noun phrase direct objects. They also show that verbs which frequently undergo Heavy-NP shift are slower to respond to when placed in a non-shifted context. They suggest that each verb is stored with a 'shifting disposition'; a frequency-based preference for whether it expects to appear contiguous with its arguments or not.

In summary, the conditional probability of a subcategorization frame given a verb plays a role in disambiguation in comprehension. The higher the conditional probability of the frame, the more it will be preferred in disambiguation. In production, the evidence is less conclusive, and awaits further study.

2.5 Conditional and lexicalized syntactic frequencies

The subcategorization bias of a verb is a kind of conditional probability; the probability of seeing an NP or an S given the verb. A number of experiments have found evidence that sentence comprehension makes use of another kind of conditional probability: the probability of a word or the probability of a lexical category conditioned on previous context or on particular syntactic structure.

Juliano and Tanenhaus (1993) studied the role of the frequency of the different lexical categories of *that*. *That* can be a determiner, a complementizer, a pronoun, or an intensifier. Overall, the pronoun reading of *that* has a higher frequency than any of the other uses. But Juliano and Tanenhaus (1993) noticed that the frequencies of these different categories are dependent on the syntactic context, as shown with percentages in Table 3.

	Determiner	Complementizer
Start of Sentence	35%	11%
After Verb	6%	93%

Table 3: Brown Corpus part-of-speech percentages for *that* (from Juliano and Tanenhaus 1993).

They conducted a self-paced reading study using sentences like those in (18–21). In (19) and (21), *that* must be a complementizer, while in (18) and (20), *that* must be a determiner. The word *diplomat/s* provides the disambiguating information (the plural is only compatible with the complementizer reading).

(18) The lawyer insisted *that* experienced **diplomat would** be very helpful

(19) The lawyer insisted *that* experienced **diplomats would** be very helpful

(20) *That* experienced **diplomat would** be very helpful to the lawyer.

(21) *That* experienced **diplomats would** be very helpful made the lawyer confidence.

If subjects make use of the conditional probability of the part of speech given the context, they would treat initial *that* as a determiner, and post-verbal *that* as a complementizer. This would predict increased reading time for the sentence-initial complementizer reading (21) and for the post-verbal determiner reading (20). Juliano and Tanenhaus (1993) found just such an interaction; reading times for *would* were longer in (21) and (20) and shorter in (19) and (18). Notice that the simple unconditioned use of the different lexical-category frequencies for *that* would not predict this interaction.

A second piece of evidence for the use of probabilities conditioned on previous structure comes from the Trueswell *et al.* (1993) experiment discussed above. Recall that Trueswell *et al.* (1993) showed that cross-modal naming latency to *he* and *him* was longer after hearing S-bias verbs (*The old man suspected...he*) than after hearing NP-bias verbs (*The old man remembered...he*). I had introduced this result as evidence for the use of verb subcategorization frequencies in comprehension. But in a separate analysis, Trueswell *et al.* (1993) also showed that the longer latency to S-bias verbs was not uniform for all S-bias verbs. It has been often noted that the complementizer *that* is optional after some S-bias verbs. Trueswell *et al.* (1993) measured the frequency with which each S-bias verb occurred with an explicit *that*, to compute the 'that-preference' for each verb. They found that this that-preference correlated with the increased reading time; the more an S-bias verb expected to be followed by *that*, the longer the latency on naming *he*. Once again, this suggests that subjects are computing the probability of the *that* complementizer conditioned on previous structure (in this case the verb).

A third study supporting the idea of probabilities conditioned on previous structure came from the MacDonald (1993) paper described above. Recall that MacDonald was looking at the processing of a noun followed by a word that is ambiguous between a noun and verb, such as the pair *miracle cures*. MacDonald looked at the frequency with which the first word occurs as the head of a noun-phrase, versus the frequency with which it occurs as the modifier of a noun-phrase. For example, the noun *warehouse* appeared in the Brown Corpus more often as a modifier, while *corporation* occurred more often as a head. If subjects make use of this probability, they should treat more-frequent-heads as complete noun phrases, parsing the following word (*cures*) as a verb; nouns which are more likely to be modifiers might cause the following word to be treated as a noun. MacDonald found that the frequency with which a word occurred as a head versus modifier in the Brown corpus did predict reading time difficulty on the word following these bigrams.

In summary, the conditional probability of a word (like *that*), or a lexical category (like *Det* or *Comp*) given previous words or structure plays a role in disambiguation. Words or categories with higher conditional probabilities are preferred.

2.6 Constructional Frequencies

The frequency effects that I have described so far are all lexical in some way. Indeed the vast majority of frequency effects that have been studied involve lexical structure. A small number of studies have looked for frequency effects for larger (supralexical) structures, but the results are relatively inconclusive.

d'Arcais (1993), for example, studied the effect of idiom frequency on word-by-word reading of Dutch idioms. Some of the idioms had syntactic errors (such as agreement errors) inserted in them. d'Arcais (1993) found that subjects were able to find these errors more quickly in frequently used idioms than in less frequently occurring idioms.

A number of researchers have suggested that one of the factors contributing to the difficulty of the main-verb/reduced relative ambiguity is the relative rarity of reduced relative clauses. Tabossi *et al.* (1994), in a norming study for an experiment on 32 verbs, checked 772 sentences from the Brown Corpus containing *-ed* forms of the 32 verbs. They found that the verb occurred as part of a simple main clause in 37 percent of the sentences, a relative clause in 9 percent, and a reduced relative in 8 percent.

Jurafsky (1996), McRae *et al.* (1998), and Narayanan and Jurafsky (1998) among others showed that (various) models which include the corpus-based frequency of the main-clause (MC) versus reduced-relative (RR) construction are able to model certain reading time effects in MC/RR sentences such as (15), repeated here as (22). Jurafsky (1996), for example, showed that the Stochastic Context-Free Grammar (SCFG) probability for the main clause parse was significantly lower than the SCFG probability for the reduced relative parse because of two factors: first, the reduced relative construction includes one more SCFG rule, and second, this SCFG rule introducing the reduced relative structure has a very low probability.

(22) The horse raced past the barn fell.

A series of studies by Cuetos, Mitchell, and colleagues (Mitchell 1994; Cuetos *et al.* 1996) has focused on a model of disambiguation called *tuning*. Tuning models claim that people tabulate every ambiguity they encounter, together with the disambiguation decision. Future disambiguation decisions are based on choosing the most likely previously-chosen disambiguation for the ambiguity. Tuning models thus claim that syntactically ambiguous sentences are resolved to whichever choice has been made more often in the past. As a simplifying assumption, Mitchell and colleagues assume that the frequency of this choice is isomorphic to the total frequency of the structures in the language.

I discuss tuning models in this section because while tuning models could hypothetically apply to any kind of disambiguation, all previous research has focused on the frequency of two specific complex syntactic constructions. In particular, Cuetos *et al.* (1996) looked at ambiguities like those in (23), where a relative clause *who was on the balcony* could attach to either the first of two noun phrases (*the servant*) or the second noun phrase (*the actress*) (not counting the subject *someone*).

(23) Someone shot [$_{NP1}$ the **servant**] of [$_{NP2}$ the **actress**] *who was on the balcony.*

Figure 1 shows a simplified schematic of the two parse trees whose frequency is being computed.

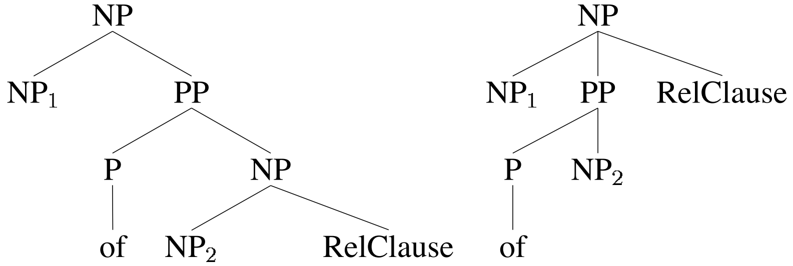

Figure 1: The Tuning Hypothesis predicts that frequencies are stored for these parses for (23), with attachment to NP$_2$ (on left) and NP$_1$ (on right).

Cuetos *et al.* (1996) found cross-linguistic differences in disambiguation preference between English and many other languages, including Spanish and Dutch. Supporting the tuning hypothesis, English speakers preferred to attach relative clauses to NP$_2$, and the NP$_2$ attachment construction shown on the left of Figure 1 was more common in a corpus (Cuetos *et al.* 1996). Also supporting the hypothesis, Spanish speakers preferred to attach relative clauses to NP$_1$, and the NP$_1$ attachment shown on the right of Figure 1 was more frequent in a Spanish corpus (Cuetos *et al.* 1996). But more recent studies have cast doubt on the link between the frequency of the two constructions and the disambiguation decisions. Studies on 3-site relative clause ambiguities in English have not found a link between corpus frequency of these (very) complex constructions and disambiguation preference (Gibson *et al.* 1996). A recent study on Dutch shows that Dutch speakers preferred to attach relative clauses to NP$_1$, but that the NP$_2$ attachment construction shown on the left of Figure 1 was more common in a corpus (Mitchell and Brysbaert 1998).

Very recent studies, such as Desmet *et al.* (2001), however, have shown that human preferences do match corpus preferences when the animacy of the NP$_1$ is held constant. Since corpora are used to estimate frequencies in most probabilistic models, this is an important result; I will return to this issue in Section 4.3. But since this control factor concerned the semantics of the noun phrases, it means that a purely structure-frequency account of the Tuning hypothesis cannot be maintained.

More recently, Bod (2000) and (2001) showed that frequent three-word (subject-verb-object) sentences (e.g. *I like it*) are more easily and faster recognized than infrequent three-word sentences (e.g. *I keep it*), even after controlling for plausibility, word frequency, word complexity, and syntactic structure. These results suggest that frequent sentences or at least some structural aspects of these frequent sentences are stored in memory.

While studies of the complex structures hypothesized by the Tuning hypothesis do not show strong evidence for frequency effects in comprehension, there have been some studies on frequency effects for simpler syntactic structure in production. Bates and Devescovi (1989) performed a cross-linguistic series of production studies which attempted to control for semantic and pragmatic factors. They found that relative clauses, which are generally more frequent in Italian than in English, occurred more frequently in their production study even after these controls. They suggest that the frequency of the relative clause construction in Italian may play a role in its being selected in production.

In conclusion, while some studies seem to suggest that the frequency of larger non-lexical syntactic structures plays a role in disambiguation, the evidence is still preliminary and not very robust. None of the studies that found an effect of non-lexical syntactic or idiom structure did so after carefully controlling for lexical frequencies and two-word or three-word bigram frequencies. The results of Bod (2001) clearly point to storage of three-word chunks, but it's not necessary that it is higher-level structure that is playing a causal role. But of course the frequency of complex constructions is much lower than lexical frequencies, and so we expect frequency effects from larger constructions to be harder to find. This remains an important area of future research.

2.7 Summary of Psycholinguistic Results on Frequency and Probability

Frequency plays a key role in both comprehension and production, but solid evidence exists only for frequency related in some way to lexical items, or the relationship between lexical items and syntactic structure.

High-frequency words are recognized more quickly, with less sensory input, and with less interference by neighbors than low-frequency words. High-frequency words are produced with shorter latencies and shorter durations than low-frequency words. Low frequency words are more subject to phonological speech errors.

The frequencies of various lexical categories related to a word play a role in language processing. For words which are morphologically, syntactically or semantically ambiguous, the more frequent part of speech, morphological category, or sense is accessed more quickly and is preferred in disambiguation. But this effect of lexical semantic/syntactic category does not seem to extend to production.

The frequency of multiple-word structures plays a role in both comprehension and production. Frequent word pairs or idioms are faster to access and/or preferred in disambiguation. Frequent word pairs or words which have a high Markov bigram probability given neighboring words are shorter in duration and phonologically more reduced.

Various kinds of conditional probabilities play a role in comprehension and production. For verbs which have more than one possible syntactic subcategorization, the more frequent subcategorization frame is preferred in disambiguation. The probability of a verb appearing non-contiguous with its complement plays a role in production. For words with more than one potential part of speech, the part of speech with higher conditional probability given the preceding part of the sentence is preferred.

Finally, a frequency effect for other, larger syntactic structures, while not disproved, certainly remains to be shown.

Of course many other kinds of knowledge play a role in probabilistic evidence-combination; I have only focused on knowledge for which a frequency effect has been found. But another important kind of knowledge is the relationship between lexical and thematic knowledge. For example, animate nouns are more likely to be agents, while inanimate nouns are more likely to be patients. Or the word *cop*, for example, is more likely to be the agent of the verb *arrested* than is the noun *crook*. Many studies have shown that this kind of thematic role information plays a role in comprehension (Trueswell *et al*. 1994; Garnsey *et al*. 1997; McRae *et al*. 1998).

3 Probabilistic Architectures and Models

The preceding section showed that frequencies of linguistic structure, especially linguistic structure related to lexical items, plays a role in language processing. In this section I turn to probabilistic architectures for modeling these frequency effects. Practically all of these models address the process of comprehension, most of them focusing on syntactic comprehension. I will discuss a few preliminary directions toward probabilistic models of production.

3.1 Constraint-Based Models

A large class of experimental and modeling work in sentence comprehension belongs to the framework generally called *constraint-based* or sometimes *constraint-based lexicalist* (MacDonald *et al*. 1994; McRae *et al*. 1998; Spivey-Knowlton *et al*. 1993; Spivey-Knowlton and Sedivy 1995; Seidenberg and MacDonald 1999; Trueswell and Tanenhaus 1994; Trueswell *et al*. 1994; Kim *et al*. 2001). Specific models differ in various ways, but constraint-based models as a class focus on the interactions of a large number of probabilistic constraints to compute parallel competing interpretations.

The focus of much work in the constraint-based framework has been on experiments showing that certain frequency-based constraints play a role in sentence processing, either via regression or full factorial analysis of reading time data. I have summarized above a number of these experimental results on the roles of verb bias, collocation frequencies, etc., in sentence comprehension. The constraint-based framework includes some computational models in addition to experimental results. In general these are neural-network models which take as input various frequency-based and contextual features, and combine these features via activation to settle on a particular interpretation (Burgess and Lund 1994; Kim *et al*. 2001; Spivey-Knowlton 1996; Pearlmutter *et al*. 1994; Tabor *et al*. 1997).

I have chosen one of these models to describe, the competition-integration model of Spivey-Knowlton (1996), because it has been most completely implemented, because it, more than other such models, is clearly intended to be probabilistic, and because it has been widely tested against experimental results from a number of reading-time studies (McRae *et al*. 1998; Tanenhaus *et al*. 2000; Spivey and Tanenhaus 1998). The input to the Spivey model is a set of probabilistic features like the bias for main clauses versus reduced relatives, the verb's preference for participle versus preterite, the contextual support for a particular interpretation, and so on. Some input features are derived from frequencies; others come from rating studies. All features are then normalized to estimate a probabilistic input feature varying between 0 and 1. The model uses a neural network to combine these constraints to support alternative interpretations in parallel. Each syntactic alternative is represented by a pre-built localist node in a network; thus the network models only the disambiguation process itself rather than the generation or construction of syntactic alternatives. The alternatives compete until one passes an activation threshold.

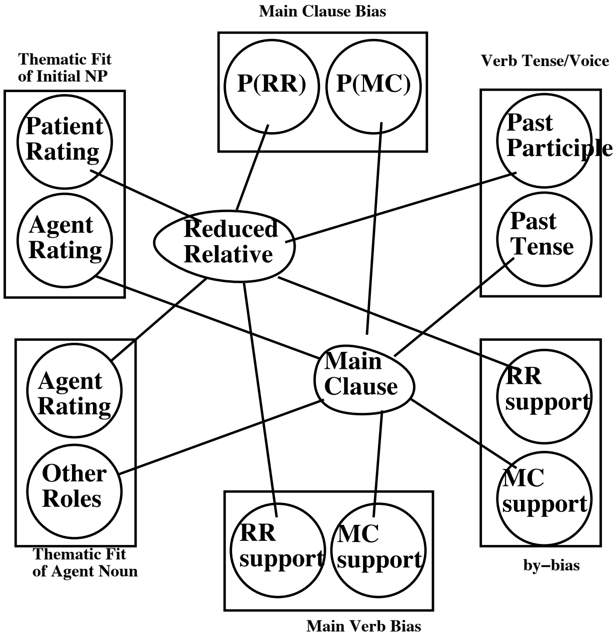

Figure 2: A schematic of the competition model, from McRae *et al.* (1998).

Each interpretation receives activation from the constraints which is then fed back to the constraint nodes within each cycle of competition. The algorithm first normalizes each pair of constraints. Let $C_{i,a}$ be the activation of the ith constraint node connected to the ath interpretation node. $C'_{i,a}$ will be the normalized activation; the activation of each constraint thus ranges from 0 to 1.

$$C'_{i,a} = \frac{C_{i,a}}{\sum_a C_{i,a}} \tag{24}$$

The activation I_a from the constraints to interpretation a is a weighted sum of the activations of the constraints, where w_i is the weight on constraint i:

$$I_a = \sum_i w_i \times C'_{i,a} \tag{25}$$

Finally, the interpretations send positive feedback to the constraints:

$$C_{i,a} = C'_{i,a} + I_a \times w_i \times C'_{i,a} \tag{26}$$

These three steps are iterated until one interpretation reaches criterion. Reading time is modeled as a linear function of the number of cycles it takes an interpretation to reach criterion.

This model accounts for reading time data in a number of experiments on disambiguation of main-verb/reduced-relative ambiguities (McRae *et al.* 1998; Tanenhaus *et al.* 2000; Spivey and Tanenhaus 1998). Let's look at the McRae *et al.* (1998) study, which included two experiments.

The first was a sentence completion experiment. For each verb in their study, McRae *et al.* (1998) had subjects complete four sentence fragments like the following:

The crook arrested
The crook arrested by
The crook arrested by the
The crook arrested by the detective

For each of these fragments, they measured the proportion of reduced relative clause completions. McRae *et al.* (1998) then showed that combining a number of probabilistic factors via the competition-integration model correctly predicted the completion preferences for main clause versus reduced relatives.

McRae *et al.* (1998) also showed that the thematic fit of a subject with the verb plays a role in reading time. Consider the difference between good agents for *arrested* like *cop* (*The cop arrested...*) and good patients for *arrested* like *crook*. Figure 3 shows that controlled human reading time for good agents like *cop* gets longer after reading the *by*-phrase, (requiring cop to be a patient) while controlled reading time for good patients like *crook* gets easier.[1] McRae *et al.* (1998) again show that the competition-integration model predicts this this reading time difference.

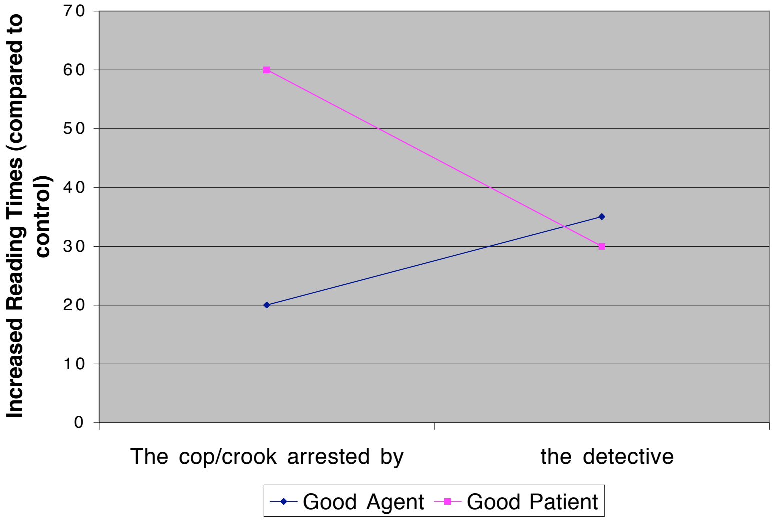

Figure 3: Self-paced reading times (from Figure 6 of McRae *et al.* (1998))

3.2 Rational and Utility-Based Probabilistic Models

3.2.1 The Competition Model

There are remarkable similarities between some of the earliest probabilistic models of sentence processing and some very recent models. They all attempt to combine the ideas of probability with utility, cost or other ideas of rationality in processing.

[1] Controlled reading times are computing by subtracting reading times on reduced relative clauses from those for unreduced relative clauses).

The *competition* model (MacWhinney *et al.* 1984; MacWhinney and Bates 1989) may have been the first probabilistic model of sentence processing. The goal of the model is to map from the 'formal' level (surface forms, syntactic constructions, prosodic forms, etc) to the 'functional' level (meanings, intentions). Since input is ambiguous and noisy, the model assumes that the sentence processor relies in a probabilistic manner on various surface cues for building the correct functional structures. The model focuses on how these cues probabilistically combine to suggest different interpretations, and on how these probabilities differ from language to language. Consider the problem of assigning agent and patient roles to noun phrases in an input sentence. An English-speaking comprehender relies heavily on word order cues in making this mapping while a German speaker relies more heavily on morphological (case) cues.

The competition model formalizes the notion of cues via *cue validity*, which is generally defined in the competition model as a combination of *cue availability* and *cue reliability*. Consider the task of identifying an interpretation i given a cue c. Cue availability is defined in Bates and MacWhinney (1989) as the ratio of the cases in which the cue is available over the total number of cases in a domain. Probabilistically, we can think of this as an estimate of the prior probability of a cue. Cue reliability is defined in Bates and MacWhinney (1989) as the ratio of cases in which a cue leads to the correct conclusion over the number of cases in which it is available. Probabilistically, this relative frequency is the maximum likelihood estimate of $P(i|c)$. If cue availability and cue reliability are combined by multiplication, as suggested by McDonald (1986), then cue validity $v(c, i)$ of cue c for interpretation i works out to be the joint probability of cue and interpretation:

$$v(c,i) = \text{availability}(c) \times \text{reliability}(c) = P(c) \times P(i|c) = P(c,i) \tag{27}$$

The competition framework views cue validity as an objectively correct value for the usefulness of a cue in a language, derived from corpora and studies of multiple speakers. *Cue strength* is the subjective property of a single human language user, the probability that the human attaches to a given piece of information relative to some goal or meaning. This use of joint probability as a measure of cue strength seems to be equivalent to the cue strength used in memory models like SAM (Search of Associative Memory) (Raaijmakers and Shiffrin 1981; Gillund and Shiffrin 1984).

How do cues combine to support an interpretation? McDonald and MacWhinney (1989) formalize cue combination by assuming that the contribution of each cue toward an interpretation is independent, and that cue strengths vary between 0 and 1. Given these assumptions, they give the following equation for cue combination, where A and B are interpretations, and C is the set of all cues $c_1 \ldots c_n$.

$$P(A|C) = \frac{\prod_i P(A|c_i)}{\prod_i P(A|c_i) + \prod_i P(B|c_i)} \tag{28}$$

The numerator in (28) factors the probability $P(A|C)$ into separate terms $P(A|c_i)$ for each of the individual cues c_i, while the denominator acts as a normalizing term. Multiplying the factors of the individual cues c_i implies that they are independent, as we saw in equation (5) in Chapter 1. Assuming that the contribution of each cue is independent is a simplifying assumption that resembles the 'naive Bayes' independence assumption often used in the categorization literature.

The competition model also considers another kind of validity in addition to cue validity: *conflict validity*. Conflict validity is based on how useful a cue is in a competition situation. It is defined (Bates and MacWhinney 1989) as the ratio between the number of competition situations in which a cue leads to a correct interpretation divided by the number of competition situations in which that cue participates. Thus the absolute frequency or validity of a cue isn't as important as how useful the cue is in disambiguation situations. Conflict validity is thus related to the idea of discriminative training in machine learning, and to the Tuning hypothesis of Mitchell (1994) and Cuetos *et al.* (1996) which proposed that ambiguities are resolved to whichever interpretation was chosen more frequently in the past.

Finally, the competition model also considers factors related to the cost of a cue. For example, certain cues may be difficult to perceive ('perceivability cost'), or may use up short-term memory to hold around until they are integrated ('assignability costs').

3.2.2 Rational Models

Anderson (1990) proposed a rational model for human cognitive processing. The rational framework claims that human cognitive processing makes optimal use of limited resources to solve cognitive problems. The optimal solution to many problems of decision given noisy data and limited resources is known to be probabilistic. Anderson thus applies a probabilistic formulation of his rational model to the task of modeling human memory and categorization. In the course of doing this, he shows how his model explains some results in lexical access.

Anderson assumes that a rational system for retrieving memory would retrieve memory structures serially ordered by their probabilities of being needed p, and would consider the gain G associated with retrieving the correct target, and the cost C of retrieving the item. Such a memory should stop retrieving items when

$$pG < C \tag{29}$$

Anderson shows that by making certain strong independence assumptions it is possible to produce a relatively straightforward equation for $P(A|H_A\&Q)$, the probability that memory trace A is needed, conditional on some history H_A of it being relevant in the past, and context Q. Let i range over elements that comprise the context Q. Anderson gives the following equation:

$$P(A|H_A\&Q) = P(A|H_A) * \prod_{i \in Q} \frac{P(i|A)}{P(i)} \tag{30}$$

Equation (30) says that we can estimate the posterior probability that A is needed from two terms; a term representing A's past history (how frequent it was and how often it was needed) and a term representing the ratio of the conditional probabilities of the cues given that the structure is relevant versus not relevant. Anderson proposes that an increase in need probability $P(A|H_A\&Q)$ maps monotonically into higher recall probability and faster latency (reaction time). He shows that his model predicts a number of basic results in recall rates and latencies, including some results in lexical processing. For example his model predicts the result that low-frequency words are better-recognized than high-frequency words (Kintsch 1970). This sketch of equation (30) and the model has been necessarily but unfortunately brief; the interested reader should turn to Anderson (1990).

Chater *et al.* (1998) apply Anderson's rational model to sentence processing. They first propose that the goal of the human parser is to maximize the probability of obtaining the globally correct parse. They assume, extending Anderson's serial model of memory, that as each word is input, the parser considers all possible parses in series. But they suggest that ordering these parses just by their probabilities may not be optimal. A parse which seems (locally) to be optimal may turn out to be the wrong parse. A serial parser would garden-path at this point, then have to backtrack and reanalyze a sentence. They therefore suggest that an optimal parser would need to include the cost of this backtracking into its algorithm for choosing a parse to follow at points of ambiguity. In particular, they suggest that it is important to balance the probability of a hypothesis, how long it would take to *settle* on the hypothesis (i.e. follow it and build the parse tree) and how long it would take to *escape* from the hypothesis (test and reject it). Given this assumption, they show that a serial parser should first consider the hypothesis H_i with the highest value of the following function f of H_i:

$$f(H_i) = P(H_i) \times P(\text{settle}H_i) \times \frac{1}{1 - P(\text{escape}H_i)} \tag{31}$$

Chater *et al.*'s (1998) proposal that the function f, rather than unaugmented probability p, is the utility function maximized by the human parser, is an intriguing claim about sentence processing that remains to be tested.

3.3 Markov Models of lexical category preference

The previous sections motivated the use of probabilities as a tool for ranking interpretations or actions taken by the human comprehension mechanism. We turn now to the details of some probabilistic models. The next three sections describe probabilistic models of lexical category and syntactic disambiguation. Each section describes iteratively more sophisticated probabilistic models. We begin in this section with Hidden Markov Models, turn in the next section to Stochastic Context-Free Grammars, and conclude with Bayesian Belief Networks. All of these are instances of what are often called *graphical models* (Jordan 1999).

Corley and Crocker (1996) and Corley and Crocker (2000) focus on the problem of lexical category disambiguation as part of human sentence processing. They propose that human lexical category disambiguation can be modeled by a hidden Markov model (HMM) part-of-speech tagging algorithm (or a variant of the HMM algorithm known as the Church tagger (Church 1988)).

HMM taggers are used to compute the probability of a sequence of part-of-speech tags given a sequence of words. For example, given a sequence of words (32) a tagger would produce a series of tags (33):

(32) The miracle cures

(33) Det Noun Noun

HMM taggers rely on a very simple intuition: given a word, choose its most-likely tag in context. They operationalize 'most-likely' by using only two probabilistic sources of knowledge: the probability of a word given a lexical category tag $P(w_i|t_i)$ and the probability of one lexical category tag following another $P(t_i|t_{i-1})$.

For example, the word 'race' can be a verb or a noun. The noun is vastly more frequent. But verbs are more common after the infinitive marker *to*. Table 4 (taken from Jurafsky and Martin (2000), with probabilities from the combined Brown and Switchboard corpora) shows that an HMM tagger correctly chooses the verb part-of-speech in the context *to race*.

HMM taggers actually produce the most-likely sequence of tags $\hat{t_1^n}$ for an entire sentence or sequence of words of length n rather than just for a single word $\hat{t_i}$. We can use the function

Words	$P(t_i\|t_{i-1})P(w_i\|t_i)$	P
to/INF race/VERB	P(Verb|InfTo) × P(race|Verb)	.00001
to/INF race/NOUN	P(Noun|InfTo) × P(race|Noun)	.000007

Table 4: HMM tag probabilities for *race* (from Jurafsky and Martin (2000))

$\text{argmax}_x f(x)$, which returns the x which maximizes $f(x)$, to write the equation for what the tagger is maximizing:

$$\hat{t_1^n} = \underset{t_1^n}{\text{argmax}}\, P(t_1^n|w_1^n) \tag{34}$$

This equation can be rewritten by Bayes' rule (Chapter 1, Equation 9), as

$$\hat{t_1^n} = \underset{t_1^n}{\text{argmax}}\, \frac{P(w_1^n|t_1^n)P(t_1^n)}{P(w_1^n)} \tag{35}$$

Since an HMM tagger is choosing the most likely tag sequence for a fixed set of words w_1^n, we can drop the denominator term, producing:

$$\hat{t_1^n} = \underset{t_1^n}{\text{argmax}}\, P(w_1^n|t_1^n)P(t_1^n) \tag{36}$$

The HMM tagger makes two large simplifying assumptions; first, that the probability of a word depends only on its own tag, and not any neighboring tags, and second that the words are independent of each other. This results in the following equation by which a bigram tagger estimates the most probable tag sequence:

$$\hat{t_1^n} = \underset{t_1^n}{\text{argmax}}\, P(t_1^n|w_1^n) \approx \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1}) \tag{37}$$

Corley and Crocker (1996) and Corley and Crocker (2000) show that this probabilistic model accounts for a number of the psycholinguistic results discussed above. For example they model the Juliano and Tanenhaus (1993) results in which the subjects seem to treat sentence-initial *that* as a determiner, but post-verbal *that* as a complementizer. Table 5 shows that the HMM probabilities predict a determiner reading sentence-initially, but a complementizer reading after a verb:

Context	Part of Speech	$P(t_i\|t_{i-1})P(w_i\|t_i)$	P
Sentence Initial	Comp	P(Comp|#)P(that|Comp)	.0003
	Det	P(Det|#)P(that|Det)	**.0011**
Following Verb	**Comp**	P(Comp|Verb)P(that|Comp)	**.023**
	Det	P(Det|Verb)P(that|Det)	.00051

Table 5: Corley and Crocker's (1996) probability computation for HMM tagger on data from Juliano and Tanenhaus (1993).

Corley and Crocker (1996) and (2000) also show that their tagging model accounts for three other results on lexical category disambiguation.

3.4 Stochastic Context-free Grammars

Jurafsky (1996) proposed a probabilistic model for syntactic disambiguation. His probabilistic parser kept multiple interpretations of an ambiguous sentence, ranking each interpretation by its probability. The probability of an interpretation was computed by multiplying two probabilities: the stochastic context-free grammar (SCFG) 'prefix' probability of the currently-seen portion of the sentence, and the 'valence' (syntactic/semantic subcategorization) probability for each verb.

A stochastic context-free grammar, first proposed by Booth (1969), associates each rule in a context-free grammar with the conditional probability that the left-hand side expands to the right-hand side. For example, here is the probability of two of the expansions of the nonterminal NP, computed from the Brown corpus

[.42] NP → Det N
[.16] NP → Det Adj N

Jurafsky's model was on-line, using the left-corner probability algorithm of Jelinek and Lafferty (1991) and Stolcke (1995) to compute the SCFG probability for any initial substring (or 'prefix') of a sentence.

Subcategorization probabilities in the model were also computed from the Brown corpus. For example the verb *keep* has a probability of .81 of being bivalent (*keep something in the fridge*) and a probability of .19 of being monovalent (*keep something*).

While the model kept multiple interpretations, it was not fully parallel. Low probability parses were pruned via beam search. Beam search is an algorithm for searching for a solution in a problem space that only looks at the best few candidates at a time. The name derives from the metaphor of searching with a flashlight; only things that lie within the beam of the light are kept around. The use of beam search in the algorithm, rather than full parallel search, means that the model predicts extra reading time (the strong garden path effect) when the correct parse has been pruned away and the rest of the sentence is no longer interpretable without reanalysis.

Jurafsky (1996) showed that this model could account for a number of psycholinguistic results on parse preferences and on garden path sentences. For example, the corpus-based subcategorization and SCFG probabilities for *keep* and other verbs like *discuss* correctly modeled the preferences for these verbs in the off-line forced-choice experiment of Ford *et al.* (1982). The SCFG grammar also correctly models the misanalysis of garden path sentences like (38), by claiming that the correct parse (in which *house* is a verb) gets pruned:

(38) The complex houses married and single students and their families.

Finally, the combination of SCFG probability and subcategorization probability models the garden path effect for preferentially transitive verbs like *race* and the weaker garden path effect for preferential intransitive verbs like *find*.

(39) The horse raced past the barn fell.

(40) The bird found in the room died.

In summary, the Jurafsky (1996) parser has the advantages of a clean, well-defined probabilistic model, the ability to model the changes in probability word-by-word, a parallel processing architecture which could model both lexical and syntactic processing, accurate modeling of parse preference, and a probabilistic beam search architecture which explains difficult garden path sentences. The model has many disadvantages, however. First, it only makes very broad-grained reading-time predictions; it predicts extra reading time at difficult garden path sentences, because the correct parse falls out of the parser's beam width. But it does not make fine-grained reading-time predictions of any kind. In addition, although the description of the model claims that the interpreter can combine probabilistic information of any sort, the model as described only specifies SCFG and subcategorization probabilities. Finally, the model has only been tested on a handful of examples.

Crocker and Brants (2000) propose a wide-coverage probabilistic model of sentence processing which is similar to Jurafsky (1996) but which, unlike Jurafsky's, is designed to have wide-coverage and efficient scalability. Their *incremental cascaded Markov model* (ICMM) is based on the broad coverage statistical parsing techniques of Brants (1999). ICMM is a maximum-likelihood model, which combines stochastic context-free grammars with hidden Markov models, generalizing the HMM/SCFG hybrids of Moore *et al.* (1995). The original non-incremental version of the model constructs a parse tree layer by layer, first at the preterminal (lexical category) nodes of the parse tree, then the next higher layer in the tree, and so on. In the incremental version of the model, information is propagated up the different layers of the model after reading each word. Each Markov model layer consists of a series of nodes corresponding to phrasal (syntactic) categories like NP or ADVP, with transitions corresponding to trigram probabilities of these categories. The output probabilities of each layer are structures whose probabilities are assigned by a stochastic context-free grammar. Figure 4 shows a part of the first Markov model layer for one sentence. Each Markov model layer acts as a probabilistic filter, in that only the highest probability non-terminal sequences are passed up from each layer to the next higher layer. The trigram transition probabilities and SCFG output probabilities are trained on a treebank.

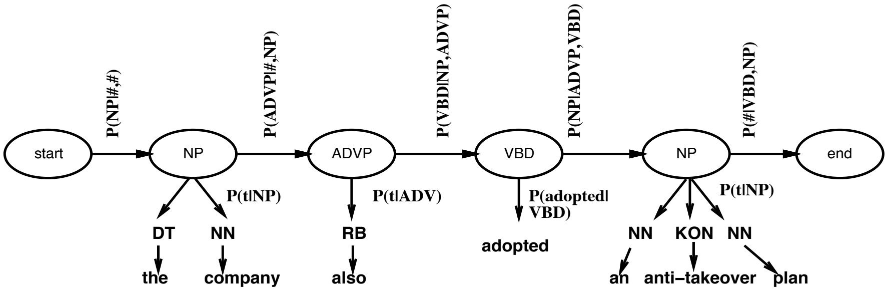

Figure 4: Part of the first layer Markov model for one sentence, from Crocker and Brants (2000). The letter t indicates the subtrees generated by the SCFG. Thus for example $P(t|NP)$ is the conditional probability of the subtree $NN \rightarrow$ *company* given the NP.

The Crocker and Brants (2000) model accounts for a number of experimental results on human parsing. For example, because the ICMM is a generalization of the Corley and Crocker (1996) model, it handles the same lexical category effects that I described in the previous section, including Juliano and Tanenhaus's (1993) conditional probability effect of *that*.

The ICMM also models a recent result of Pickering *et al.* (2000), who were looking at disambiguation of the role of noun phrases like *his goals* in NP/S ambiguities like the following:

(41) The athlete realized $[_{NP}$ his goals] at the Olympics.

(42) The athlete realized $[_{S}[_{NP}$ his goals] were out of reach].

Realize is an S-bias verb. Nonetheless, Pickering *et al.* (2000) showed that readers must be considering the NP interpretation of *his goals*. They did this by creating pairs of sentences with sentential complements. In one of the sentences (43), the noun phrase *potential* was a plausible direct object for *realize*. In the other sentence (44), the noun phrase *her exercises* was not a plausible direct object.

(43) The young athlete realized her potential one day might make her a word-class sprinter.

(44) The young athlete realized her exercises one day might make her a word-class sprinter.

Pickering *et al.* (2000) showed that reading time was delayed on the phrase *might make her* after the implausible direct object *her exercises* but not after the plausible direct object *her potential*. In order to be influenced by the plausibility of the direct object, the human parser must be building the direct object interpretation, despite the S-bias of the verb *realized*.

Crocker and Brants (2000) use the structure of the SCFG to model this result; sentential complements involve an extra SCFG rule than direct objects (the rule VP → S). The probability of the sentential complement will thus be lower than it would be otherwise; since probabilities are less than 1, multiplying by an additional rule lowers the probability of a parse. Thus their model predicts that the probability of the direct object reading of (42) is actually higher than the probability of the sentential complement reading.

Like Jurafsky (1996) and Crocker and Brants (2000), Hale (2001) proposes to model human sentence processing via a probabilistic parser based on SCFG probabilities. But Hale's model offers an important new contribution: much more fine-grained predictions about parsing difficulty and hence reading time. Hale proposes that the cognitive effort to integrate the next word into a parse is related to how surprising or unexpected that word is. The 'surprisal' of a word is an alternate term in information theory for the information value of the word (Attneave 1959), which can be computed by the negative log of its probability. Thus Hale's proposal is that reading times at a word are a function of the amount of information in the word. A word which is surprising or informative (has a large negative log probability and hence a large positive information content) will cause extended reading times, and hence a garden path sentence:

$$h(w_i) = -\log P(w_i) \tag{45}$$

How should the probability $P(w_i)$ be estimated? This is of course a cognitive modeling question; the appropriate probability is whatever people can be shown to use. Hale proposes to use a simple syntax-based probability metric: the conditional SCFG probability of the word given the parse tree of the preceding prefix.

The conditional probability of a word given the previous structure can be computed from a SCFG by using the *prefix* probability. Recall from the earlier discussion that the prefix probability is the probability of an initial substring of a sentence given the grammar. Unlike the probability of an entire sentence, computing the probability of a prefix is somewhat complex, since it involves summing over the probability of all possible recursive structures before the parser knows exactly how many recursions will be seen. Jelinek and Lafferty (1991) shows how this prefix probability can be computed, and Stolcke (1995) show how this computation can be integrated into a probabilistic Earley parser. If we use α_i to mean the prefix probability of words $w_0 w_1 \ldots w_i$, then the conditional probability of a new word given all the previous words is

$$P(w_i | w_1, w_2 \cdots w_{i-1}) = \frac{P(w_1 \ldots w_i)}{P(w_1 \ldots w_{i-1})} = \frac{\alpha_i}{\alpha_{i-1}} \qquad (46)$$

Hale's proposal is that reading times at a word will be proportional to the information value assigned by this probability, or:

$$h(w_i) = -\log \frac{\alpha_i}{\alpha_{i-1}} \qquad (47)$$

Hale actually gives a different, but equally valid way of thinking about this equation. He proposes that the cognitive effort for parsing any sequence of words is proportional to the total probability of all the structural analyses which are incompatible with that sequence. That is, cognitive effort, and particularly the garden path effect, occurs wherever the parser disconfirms potential parses that together comprise a large probability mass. The simplest way to measure the amount of probability mass that is disconfirmed is to look at the amount of probability mass in the prefix leading up to the previous word that is no longer in the prefix leading up to the current word, which is the difference between α_i and α_{i-1}.

Hale shows that his model predicts the large increases in reading time corresponding to two well-known cases of the garden path effect. First, he shows that the surprise at the word *fell* is very high in the garden path sentence (48) by hand-building a mini context-free grammar for the rules in the sentence, and setting the probabilities from a sample of the Penn Treebank. Figure 5 shows the prediction of extra surprise, hence reading time at *fell*.

(48) The horse raced past the barn fell.

Hale's model predicts a large increase in reading time at *fell* because the probability of *fell* is extremely low. In this sense, Jurafsky's (1996) pruning-based model is just a special case of Hale's; Jurafsky's model predicts extra reading time because the probability of *fell* is zero; the potential parse which could have incorporated *fell* was pruned away. Hale's model is thus able to make more fine-grained reading-time predictions than Jurafsky's.

These more fine-grained predictions can be seen in Hale's probabilistic explanation for the phenomenon known as *subject-object relative asymmetry*. Many researchers had noticed that object relative clauses (49) are more difficult to parse than subject relative clauses (50); see Gibson (1998) for a summary of previous research (and a non-probabilistic model).

(49) The man who you saw saw me.

(50) The man who saw you saw me.

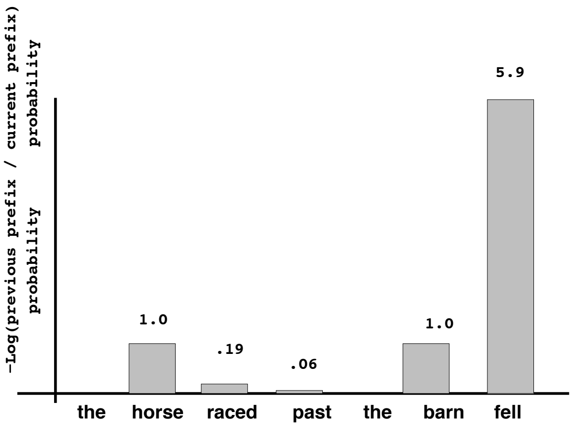

Figure 5: Hale's predictions of reading time based on surprise values computed from simple SCFG.

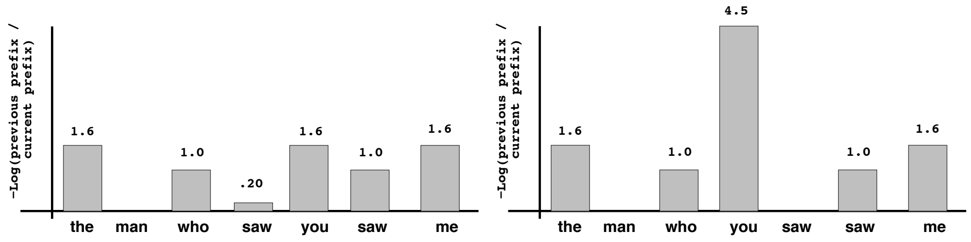

Figure 6: Hale's predictions for reading time based on surprise values computed from simple SCFG for (a) object relatives and (b) subject relatives.

Figure 6 shows the reading-time predictions of Hale's model; note that the object relative has a much higher maximum (and mean) surprisal compared to the subject relative.

In summary, probabilistic models of human parsing based on Markov models and stochastic context-free grammars use the SCFG or HMM probability to predict which parse of an ambiguous sentence a human will prefer. These models also make some predictions about time course. Jurafsky (1996) and Crocker and Brants (2000) use the beam search paradigm to prune low-probability interpretations, hence predicting longer reading time when the next word is compatible only with a parse that has already been pruned. Hale (2001) offers more fine-grained predictions about reading time, predicting an increase in reading time for surprising or unexpected words, as measured by parse probability.

3.5 Bayesian Belief Networks

I began this chapter by defining the goal of probabilistic modeling in language processing as solving the problem of choosing among possible alternatives in comprehension and production, given only incomplete and noisy evidence. The models that I have summarized so far go a long way toward this task. The Competition and Constraint-Satisfaction models both focus on the use of multiple probabilistic or frequency-based cues. The Markov and SCFG models in the previous section extend this use of probabilistic cues to show how some of these cues can be combined in a probabilistically correct manner; in particular, they focus on how we can use independence assumptions, like the assumptions of SCFGs, to combine syntactic probabilities in a structured and motivated way.

In this section I introduce a more general framework for combining probabilistic knowledge: the Bayesian belief network. Bayesian belief networks are data-structures that represent probability distributions over a collection of random variables. A network consists of a directed acyclic graph (DAG), in which nodes represent random variables (unknown quantities), and the edges between nodes represent causal influences between the variables. The strengths of these influences are quantified by conditional probabilities; thus for each variable node A which can take values $a_1 \dots a_n$, with parents $B_1, \dots B_n$, there is an attached conditional probability table $p(A = a_1|B_1 = b_x, \dots, B_n = b_z)$, $p(A = a_2|B_1 = b_x, \dots, B_n = b_z)$, and so on. The table expresses the probabilities with which the variable A can take on its different values, given the values of the parent variables. The structure of the network reflects conditional independence relations between variables, which allow a decomposition of the joint distribution into a product of conditional distributions. The Bayes net thus allows us to break down the computation of the joint probability of all the evidence into many simpler computations.

Recall that the advantage of a Bayesian approach to language processing is that it gives a model of what probability to assign to a particular belief, and how these beliefs should be updated in the light of new evidence. Bayesian belief-networks are thus on-line models; for example if we are estimating the probabilities of multiple possible interpretations of an ambiguous utterance, the network will allow us to compute the posterior probability of each interpretation as each piece of evidence arrives. In addition, the use of Bayes nets as a probabilistic estimator allows us to incorporate any kind of evidence; syntactic, semantic, discourse. This will allow us to capture the syntactic probabilities captured by graphical models like HMMs and SCFGs, while augmenting them with other probabilities, all in an on-line manner.

Jurafsky (1996) suggested that access and disambiguation of linguistic knowledge followed an evidential Bayesian model, he only gave the briefest sketch of what the model should look like. Narayanan and Jurafsky (1998) and Narayanan and Jurafsky (2001) followed up on this proposal by implementing a Bayesian model of syntactic parsing and disambiguation.

In this model, each interpretation of an ambiguous input is assigned a probability by combining multiple probabilistic sources of evidence, such as SCFG probabilities, syntactic and thematic subcategorization probabilities, and other contextual probabilities using a Bayesian belief network.

For example after seeing the first few words of an MC/RR ambiguous sentence (*The horse raced*), the Bayesian model assigns probabilities to both the main clause (MC) and reduced-relative (RR) interpretations using the belief net briefly sketched in Figure 7.

This particular belief net combines multiple sources of probabilistic evidence, such as the subcategorization probability of the verb *raced*, the probability that *horse* is the semantic *Theme* of a racing event, and the syntactic probability that a noun phrase will include a reduced relative clause, computed using SCFG probabilities.

This net is actually composed of two subnets, one computing the SCFG probabilities and one computing the lexical and thematic probabilities. The SCFG probabilities can be directly computed by this first subnet; the conditional independence assumptions in a stochastic context-free parse of a sentence can be translated into the conditional independence statements entailed by a Bayes net. Figure 8 illustrates the Belief network representations that correspond to the SCFG parses for the main clause and reduced relative interpretations of an ambiguous prefix like *The witness examined*.

Figure 9 gives the structure of the Bayes net that computes lexical and thematic support for

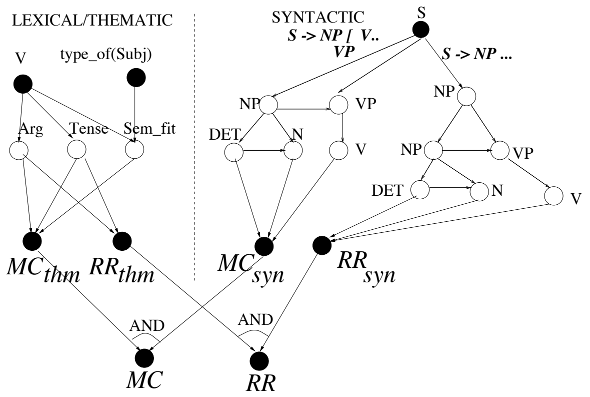

Figure 7: A belief net combining SCFG probabilities (*syn)* with subcategorization, thematic (*thm*), and other lexical probabilities to represent support for the main verb (MC) and reduced relative (RR) interpretations of a sample input. From Narayanan and Jurafsky (1998).

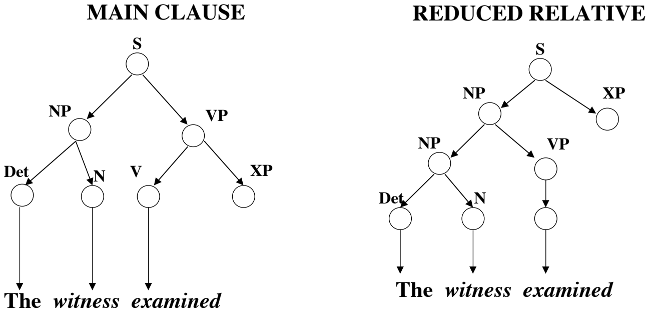

Figure 8: Pieces of belief networks corresponding to two SCFG parses for the prefix ‘The witness examined…’.

the two interpretations. The model requires conditional probability distributions specifying the preference of every verb for different argument structures, as well as its preference for different tenses. Narayanan and Jurafsky (2001) also compute probabilities from the semantic fit between head nouns (like *crook* or *cop*) and semantic roles (agent or patient) for a given predicate (like *arrested*) by normalizing the preferences given by McRae *et al.* (1998). Thus the probabilities include:

$P(\text{agent}|\text{subject=crook,verb=arrested})$,
$P(\text{patient}|\text{subject=crook,verb=arrested})$,
$P(\text{transitive}|\text{verb=arrested})$,
$P(\text{preterite}|\text{verb=arrested})$,

and so on. As shown in Figure 9, the MC and RR interpretations require the conjunction of specific values corresponding to tense, semantic fit and argument structure features. Note that only the RR interpretation requires the transitive argument structure.

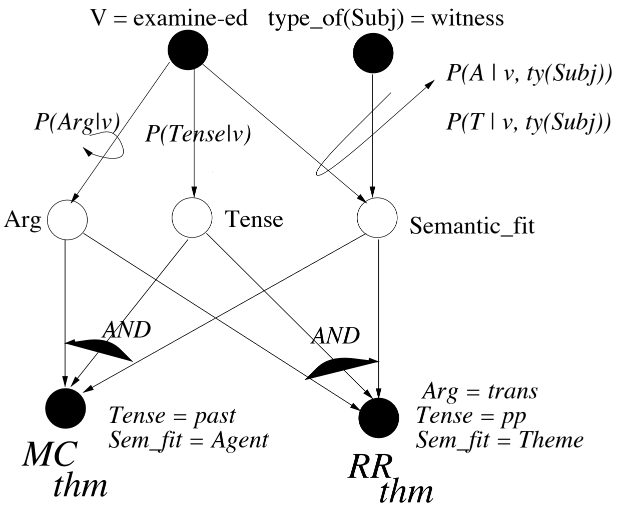

Figure 9: The belief net that represents lexical and thematic support for the two interpretations. A stands for *Agent*, and T for *Theme*.

In some cases, as with the SCFG, we have relatively complete models of the independence assumptions between probabilities. In other cases, for example between thematic and syntactic probabilities, we do not yet have a good idea what the exact causal relationship is between probabilities. The simplest thing to do in such cases is to assume the probabilities are independent and to multiply them. Figure 7 shows that Narayanan and Jurafsky (1998) make a somewhat weaker assumption by using the NOISY-AND model (Pearl 1988) in computing the conjunctive impact of the lexical/thematic and syntactic support to compute the probabilities for the MC and RR interpretations. A noisy-and model assumes that whatever *inhibits* a specific source (syntactic) from indicating support for an interpretation, is independent of mechanisms that inhibit other sources (lexical) from indicating support for the same interpretation. This is called the assumption of exception independence, and is used widely with respect to both disjunctive (NOISY-OR) and conjunctive sources. In the case of the RR and MC interpretations, as each piece of new evidence is introduced by reading new words, we compute the posterior support for the different interpretations using the following equation:

$$\begin{gathered} P(MC) = 1 - P(\neg MC) = 1 - P(\neg MC|Syn, Lex, Thm) = 1 - (P(\neg MC|Syn) \times P(\neg MC|lex, thm) \\ P(RR) = 1 - P(\neg RR) = 1 - P(\neg RR|Syn, Lex, Thm) = 1 - (P(\neg RR|Syn) \times P(\neg RR|lex, thm) \end{gathered} \tag{51}$$

Let's walk through the model word by word as it assigns probabilities to the different parses of the initial prefix of three sentences:

(52) The horse raced past...

(53) The horse carried past...

(54) The horse found in...

Previous research has found that (52) causes a severe garden path, while the other sentences do not (Pritchett 1988; Gibson 1991). The Narayanan and Jurafsky (1998) model will model this garden path effect via the beam search assumption of Jurafsky (1996); interpretations whose probability falls outside the beam-width of the best interpretation are pruned. Figure 10 shows the relevant posterior probabilities for the examples "The horse raced past the barn fell"and the replacement of *raced* by *carried* or *found* at different stages of the input, expressed in terms of the probability of the main-verb interpretation to the reduced-relative interpretation, or MC/RR ratio.

At the first point in the graph the network expresses the probability ratio after seeing the word *the horse*. The network is thus computing the following probabilities:

P(MC,S $\rightarrow$ NP $\ldots$, NP $\rightarrow$ Det N, Det $\rightarrow$ the, N $\rightarrow$ horse$|$the,horse)

P(RR,S $\rightarrow$ NP $\ldots$, NP $\rightarrow$ NP $\ldots$, NP $\rightarrow$ Det N, Det $\rightarrow$ the, N $\rightarrow$ horse$|$the,horse)

Next, the word *raced* appears, and we compute the new posterior given this new information:

P(MC,S $\rightarrow$ NP VP, NP $\rightarrow$ Det N, Det $\rightarrow$ the, N $\rightarrow$ horse, VP $\rightarrow$ V $\ldots$, V $\rightarrow$ raced, Vform,Agent $|$ Vform=preterite, subject="horse", verb=race)

P(RR,S $\rightarrow$ NP VP, NP $\rightarrow$ NP VP, NP $\rightarrow$ Det N, Det $\rightarrow$ the, N $\rightarrow$ horse, VP $\rightarrow$ V $\ldots$, V $\rightarrow$ raced, Vform,Agent $|$ Vform=participle, subject="horse",verb=race)

As shown in Figure 10 the Narayanan and Jurafsky (1998) model predicts that the MC/RR ratio exceeds the threshold immediately after the verb *raced* is accessed ($MC/RR \approx 387 \gg 5$) leading to the *pruning* of the RR interpretation. In the other cases, while the MC/RR ratio is temporarily rising, it never overshoots the threshold, allowing both the MC and the RR interpretations to be active throughout the ambiguous region.

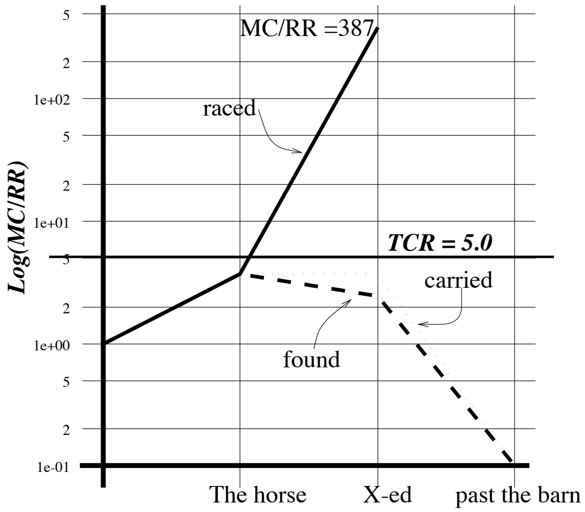

Figure 10: The MC/RR posterior probability ratio for *raced* falls above the threshold and the RR interpretation is pruned. For *found* and *carried*, both interpretations are active in the disambiguating region. From Narayanan and Jurafsky (1998).

Narayanan and Jurafsky (2001) tested the Narayanan and Jurafsky (1998) data on more data, by modeling both sentence completion probabilities and reading time data on 24 sentences from

McRae *et al*. (1998). They also included in the model new probabilities taken from McRae *et al*. (1998) which allow conditioning on the identity of the preposition. Finally, they extend the model's reading time predictions by predicting an increase in reading time whenever an input word causes the best interpretation to drop in probability enough to switch in rank with another interpretation.

The first experiment modeled by Narayanan and Jurafsky (2001) is the sentence completion experiment run by McRae *et al*. (1998) summarized above. Narayanan and Jurafsky (2001) showed that the same factors integrated by McRae *et al*. (1998) using the competition-integration model could instead be integrated by the Bayesian network shown in Figure 9 and Figure 8.

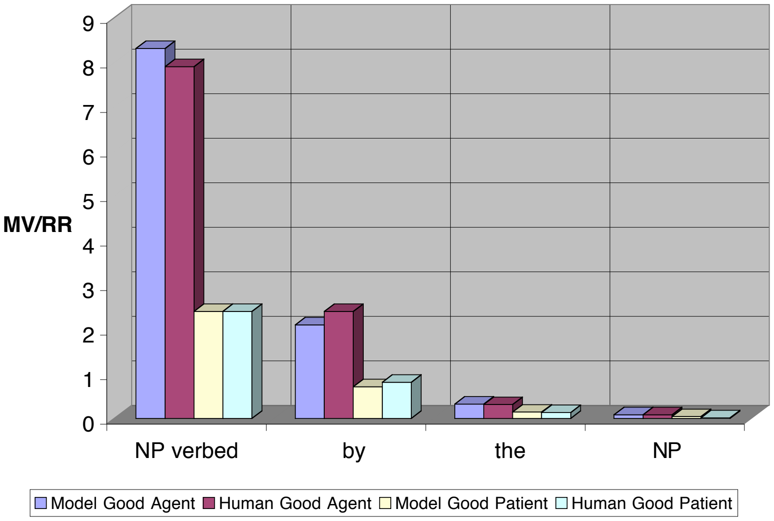

Figure 11: The MC/RR posterior probability ratio for sentence completion after each word, from the Bayesian model (Narayanan and Jurafsky 2001) and human completion data (McRae *et al*. 1998).

Figure 11 shows the human fragment completion preferences and the probabilities the model assigned to the RR or MC completions. The Bayesian model shows close correspondence to the human judgements about whether a specific ambiguous verb was used in the Main Clause (MC) or reduced relative (RR) constructions. As in McRae *et al*. (1998), the data shows that thematic fit clearly influenced the gated sentence completion task. The probabilistic account further captured the fact that at the *by* phrase, the posterior probability of producing an RR interpretation increased sharply; thematic fit and other factors influenced both the sharpness and the magnitude of the increase.

Narayanan and Jurafsky (2001) also model aspects of on-line reading experiments from McRae *et al*. (1998) discussed above. Recall that McRae *et al*. (1998) showed in Figure 3 that controlled human reading time for good agents like *cop* gets longer after reading the *by*-phrase, (requiring cop to be a patient) while controlled reading time for good patients like *crook* gets shorter. The Narayanan and Jurafsky (1998) model predicts this larger effect from the fact that the most probable interpretation for the Good Agent case *flips* from the MC to the RR interpretation in this region. No such flip occurs for the Good Patient (GP) case.

Figure 12(a) shows that the GP results already have an MC/RR ratio of less than one (the RR interpretation is superior) while a flip occurs for the GA sentences (from the initial state where

MC/RR >1 to the final state where MC/RR <1). Where Figure 12(a) showed MC/RR ratios for different initial NPs, Figure 12(b) focuses just on the Good Agent case, and breaks out the MC and RR probabilities into two separate lines, showing the crossing point where the flip occurs.

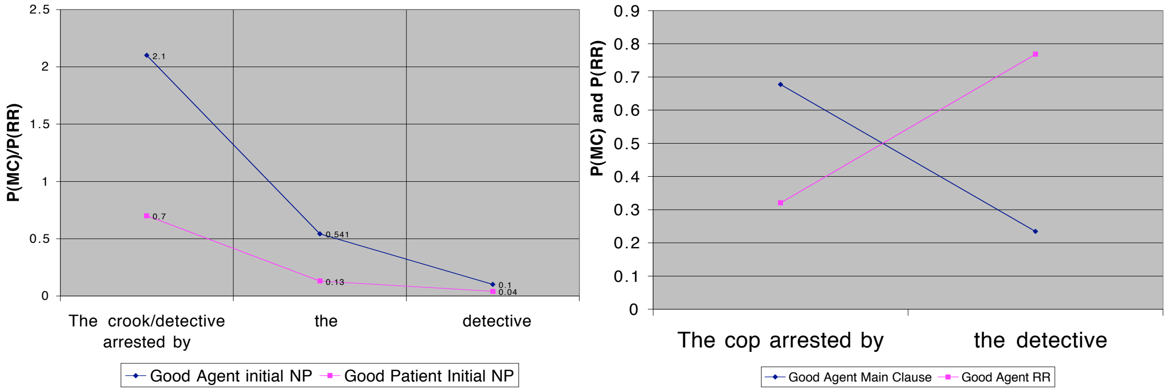

Figure 12: a) MC/RR for the ambiguous region showing a flip for the Good Agent (GA) case but no such flip for the Good Patent (GP). b) P(MC) and P(RR) just for the Good Agent cases.

3.6 Probabilistic Modeling of Production

I have now sketched the architectures of many probabilistic models of comprehension. In production, by contrast, there seems to be no worked-out probabilistic models. Perhaps the main cause of this, as Baayen (p.c.) has pointed out, is that getting probabilities for production studies is difficult, because it is so difficult to control the circumstances that will prompt a subject to coin a particular sentence. In any case, modern models of lexical production such as Levelt *et al.* (1999), Dell (1986), or Dell *et al.* (1997), and indeed most models, back to the seminal *logogen* model of Morton (1969), are based on some idea of activation which is tied in some way to frequency. In a logogen-style model a high-frequency word has a lower activation threshold, and hence is quicker to access. In other models, frequency plays its role instead via the weights on links that pass activation into a word node. In any case, Bates and Devescovi (1989) proposed that this sort of frequency-based activation model also be used to model syntactic frequency effects. In their model, given a semantic input, the production system allows various possible syntactic constructions and lexical realizations to compete for access. Frequency and function both play a role in ranking these competing realizations. Evidence for this role of syntactic frequency comes from the Bates and Devescovi (1989) study discussed in Section 2.6, which showed that in a controlled production study, relative clauses occurred far more often in Italian production than English production. They suggest that the frequency of relative clauses in Italian may play a direct role in their being chosen in production.

In addition, both Stallings *et al.* (1998) and Roland and Jurafsky (2001) suggest that various conditional probabilities relating to a verb's subcategorization frame play a role in production. The Stallings *et al.* (1998) experiment summarized in Section 2.4 suggested that verbs are stored with a 'shifting disposition'; a frequency-based preference for whether it expects to appear contiguous with its arguments or not. Stallings *et al.* (1998) suggest that this preference plays an on-line role in choosing an interpretation in production. Roland and Jurafsky (2001) suggest that a verb subcategorization probability is stored with the verb lemma and used in production to select among alternative subcategorizations.

3.7 Conclusion

I have sketched, at a surface level, a number of probabilistic models of comprehension. Most models focus on ambiguity, showing that human preference for one interpretation of an ambiguous input can be predicted by the probability of that interpretation. The Rational models extend this idea to investigate the role of cost and utility in disambiguation preferences. Finally, the most recent work has begun to explore a more fine-grained relationship between probability and processing time.

4 Potential Challenges to and Confusions About Probabilistic Models

In this section I discuss some commonly-encountered challenges to probabilistic models, and introduce some frequently-asked questions.

4.1 Surely you don't believe that people have little symbolic Bayesian equations in their heads?

No, probabilistic modeling of human language processing does not imply that little Bayesian equations are somehow symbolically dancing around in the head of a speaker. This misguided conception gives rise to a common objection to probabilistic models of language processing: that it seems hard to believe that people are "doing complex math in their heads".

Rather, many probabilistic modelers assume that probability theory is a good model of language processing at what Marr called the 'computational level'; it characterizes the input-output properties of the computations that the mind must somehow be doing. How this model is realized at lower levels (of 'implementation' or 'algorithm') is an interesting question which I unfortunately will not have much room for in this survey, although see some discussion in Baayen (this volume). The most common assumption, however, is that probability is realized either as an activation level of some mental structure or a distributed pattern of activation. Stored frequencies or probabilities can thus be encoded either as resting activation levels, or as weights on connections. It is well-known that the link between neural network or connectionist models and probabilistic ones is a close one; see for example McClelland (1998). Other realizations of probabilistic models are possible, however, such as the exemplar models of phonology of Pierrehumbert (2001) and others.

It is important to mention an alternative possible relation between probabilistic models and neural networks or other activation-based models, suggested by an Ariel Cohen. The alternative is that probabilistic models are simply wrong, and neural network or connectionist models are a better and more explanatory model of human language processing. Unfortunately, very little research has focused on discriminating between probabilistic models and connectionist models. Deciding whether probabilistic models are merely a higher-level description of connectionist models, or whether the two are mutually exclusive alternatives, remains a key problem for future research.

4.2 Are Probabilistic Models Always Non-Modular?

Frequency-based, constraint-based, and probabilistic models have often been opposed to models which exhibit Fodorian modularity, or models based on rich linguistic structure. While any individual model may make any particular confluence of claims, there is no *necessary* link between probability and anti-modularity. The probabilistic models I have discussed in this chapter are generally models of the probability *of* something; generally the probability of a certain linguistic structure, as computed by humans in the course of linguistic processing. This fact should make it clear that these probabilistic models are not meant to argue for 'using numbers instead of linguistic structure' or 'random number generators in people's heads' (to cite two more standard but misguided objections to probabilistic models). Some of the probabilistic models in this chapter are modular (for example Crocker and Brants 2000). Others are non-modular. Some involve 'emergent' structure, some involve explicit structure. Furthermore, the use of probability tells us nothing about whether many and varied probabilistic constraints are used immediately, as most probabilistic researchers believe, or after a short delay of a few milliseconds, as some psycholinguists have argued in offering the Garden Path and Construal models.

4.3 But Corpus Frequencies Don't Match Norming Study Frequencies

The probabilities in the models I have been describing are generally estimated from corpus frequencies. A number of researchers have recently noticed that these corpus frequencies don't always match the frequencies derived from various psychological experiments. This mismatch might suggest that frequency-based models are not psychologically plausible. I mentioned earlier the work of Gibson *et al*. (1996) in English and Mitchell and Brysbaert (1998) in Dutch which suggested that attachment preferences were different between corpora and production experiments. Desmet *et al*. (2001) showed that, at least in Dutch, this difference disappeared when the animacy of the NPs was controlled.

Other such mismatches have been reported, however. For example Merlo (1994) compared verb subcategorization frequencies computed from corpora with frequencies computed from psychological norming studies. In a kind of psychological norming studies called 'sentence production' studies, subjects are given a verb and asked to write a sentence using the verb. Transitivity biases are then computed from a collection of such sentences. Merlo found that transitivity preferences in a corpus of Wall Street Journal and DARPA Air Travel Information System sentences were different than transitivity preferences in norming studies such as Connine *et al*. (1984).

Roland and Jurafsky (2001) followed up on Merlo's research by looking at the causes of subcategorization differences between corpora such as the Brown Corpus and subcategorization norming studies such as Connine *et al*. (1984). Their analysis suggests that most of the differences between these verb subcategorization frequencies came from two factors. The first factor is word sense; different corpora tend to use different senses, and different senses tend to have different subcategorization biases. The second factor is discourse and genre effects; for example the single-sentence production tasks were much less likely to have passives, zero-anaphora, and other discourse-related phenomena than natural corpus sentences. Roland *et al*. (2000) extended this study, examining the subcategorization probabilities for 69 verbs. They found that after controlling for verb sense, the binned subcategorization probabilities (high transitive bias, medium transitive bias, low transitive bias) for each verb were relatively stable across corpora.

What are the implications for probabilistic models? Roland and Jurafsky (2001) and Roland (2001) proposed that the locus of verb subcategorization probabilities is the semantic lemma rather than the lexeme, and suggested that the frequency of a particular verb subcategorization in a corpus is a product of multiple factors. In particular, in lexical production, lexical subcategorization probabilities, which are stored at the semantic lemma level, might be combined with other probabilistic influences from discourse and genre to produce the observed subcategorization patterns in corpora.

Thus the mismatch between corpus frequencies and psychological norming studies is to be expected. These are essential two different kinds of production studies, with different constraints on the production process. A probabilistic model of production which is correctly conditioned on sense, genre, and other factors, would correctly model the different observed frequencies for these two kinds of corpora.

4.4 Maybe Frequency is Just an Epiphenomenon

Another common objection to probabilistic and other frequency-based models is that frequency is only an epiphenomenon of other structural factors. In this section I discuss one class of such claims, related to the processing of unaccusative and unergative verbs in English. Unaccusative and unergative verbs are both typically intransitive, but have been modeled as having a difference in underlying lexical-syntactic form, as suggested in the following table:

Unergatives	NP [$_{VP}$ V]	external argument, no internal argument
Unaccusatives	_ [$_{VP}$ V NP/CP]	internal argument, no external argument

Both unaccusative and unergative verbs generally alternate with a causative transitive form (for example unaccusative *melt* can be both intransitive and transitive). Kegl (1995) recently claims that unaccusatives (verbs like *bloom*, *melt*, *blush*, etc) are particularly hard for agrammatic aphasics to process. Her argument is based on a production study of an agrammatic aphasic subject. The aphasic subject's productions showed a significant absence of unaccusatives when compared with a matched control. Kegl's explanation was that unaccusatives are like passives in involving traces, and that other researchers claim that agrammatic aphasics have general difficulty with processing traces (Grodzinsky 2000).

An alternative explanation might be based on frequency; the comprehension difficulty of a verb might vary with the frequency-based subcategorization bias of the word, and unaccusative verbs could be more frequent in their intransitive than in their transitivized form. Gahl *et al.* proposed this hypothesis, and tested it on eight aphasic subjects using a plausibility in comprehension task, with both transitive and intransitive sentences. A given sentence thus either matched or didn't match the transitivity bias of the verb. They predicted that sentences should be easier if their structure matches verb bias, and predicted that there was no reason to expect unaccusatives to act like passives. They found that unaccusatives as a whole are much easier than passives, that unaccusatives as a whole are not harder than unergatives, and that in general sentences are easier when their syntactic structures match the subcategorization frequency bias of the verb. Thus processing of unaccusative was influenced by frequency bias, rather than by structural problems with traces.

The second proposal that structure rather than frequency causes processing difficulty comes from Stevenson and Merlo (1997), who noticed that the causativized form of unergative verbs (*race*, *sail*, *glide*), are more difficult to process than the causativized form of unaccusative verbs, as shown in Table 6.

Stevenson and Merlo (1997) were extending a proposal of Hale and Keyser (1993), in which verbs project their phrasal syntax in the lexicon. Stevenson and Merlo proposed that causativized (transitive) forms of unergatives are more complex than causativized (transitive) forms of unaccusative verbs, in terms of number of nodes and number of binding relations. This complexity, together with limitations on creating and binding empty nodes, caused the Stevenson (1994) parser to be unable to activate the structure needed to parse transitivized unergatives, hence explaining the garden path effect.

Causativized unergatives
The students advanced to the next grade had to study very hard.
The clipper sailed to Portugal carried a crew of eight.
The ship glided past the harbor guards was laden with treasure.
Causativized unaccusatives
The witch melted in the Wizard of Oz was played by a famous actress.
The oil poured across the road made driving treacherous.

Table 6: Causativized unergatives are more difficult than causativized unaccusatives, from Stevenson and Merlo (1997)

But an alternative explanation for the garden path effect relies on the subcategorization frequency biases discussed earlier. As Stevenson and Merlo and also Gahl (1999) show, unergative verbs like *race* have a huge bias toward appearing in the intransitive. Unaccusatives have a slight bias toward the causative/transitive, if anything, as Table 7 shows (see Gahl (1999) for further details of the comparison).

	Transitive		**Intransitive**	
Unergative	2869	13%	19194	**87%**
Unaccusative	17352	**54%**	14817	46%

Table 7: Transitivity counts for unergative versus unaccusative verbs, from Gahl (1999)

A frequency explanation for the difficulty of these garden path sentences also has the advantage of explaining gradient effects. Filip *et al.* (2002), for example, has shown that some unergatives are easier than others, which would be difficult to explain with a purely structural model.

The fact that frequency might be the psychological actor rather than structural factors in this instance does not mean that structural, semantic, or functional factors might not often be the causal force that is grammaticalized via frequency. Thus it is crucial to continue to investigate semantic or functional factors like those proposed by Merlo and Stevenson.

5 Conclusion

What is the state of knowledge about probabilistic modeling in 2002? We know that the frequency of many kinds of linguistic structure play a role in processing. The strongest evidence for this role, however, exists only for frequency related in some way to lexical items, or the relationship between lexical items and syntactic structure. The role of probabilities in non-lexical syntactic structure, while assumed in most probabilistic models, rests on very little psychological evidence. This is perhaps unsurprising, since the psychological evidence for constituency itself is so weak. Nonetheless, understanding the role of frequency of larger structures is an important unsolved problem.

As for models, it is clear that probabilistic models of linguistic processing are still in their infancy. Most models only include a very small number of probabilistic factors and make wildly unjustified assumptions about conditional independence. Furthermore, there is a terrible dearth of work exploring the crucial relationship between the neural network models, which focus on emergence, distributional evidence, and the details of input features, and Bayesian models, which focus on the mathematics of evidence combination and independence assumptions. Nonetheless, some conclusions are already possible. Probabilistic models do a good job of selecting the preferred interpretation of ambiguous input, and are starting to make headway in predicting the time-course of this disambiguation process.

Many unsolved problems remain; how exactly should prior probabilities be estimated from corpora? What exactly is the relationship between probability and reading or production time? We know that this relationship is logarithmic, but little about how or why.

The constraints of space and time have made this survey of probabilistic work in psycholinguistics unfortunately brief. I have given short shrift to the role of frequency in recall, to the role of phonological and orthographic neighborhood frequencies in processing, and, most distressing, to the vast connectionist literature which is so closely related to probabilistic modeling. Alas, those areas will have to await another survey.

Acknowledgements

Many thanks to the editors, Susanne Gahl, Alan Bell, and Ariel Cohen, and special thanks to Harald Baayen for help and detailed comments above and beyond the call of duty. An early version of this paper was given as a talk at the AMLAP 2001 conference in Saarbruecken; particular thanks to Matt Crocker, Don Mitchell, and Brian McElree for helpful comments. Of course all remaining errors are my own.

References

AHRENS, KATHLEEN V. 1998. Lexical ambiguity resolution: Languages, tasks, and timing. Syntax and semantics, volume 31: Sentence processing: A crosslinguistic perspective, ed. by D. Hillert. San Diego: Academic Press.

ALLEGRE, M., and P. GORDON. 1999. Frequency effects and the representational status of regular inflections. Journal of Memory and Language 40.41–61.

ANDERSON, JOHN R. 1990. The adaptive character of thought. Hillsdale, NJ: Lawrence Erlbaum.

ATTNEAVE, FRED. 1959. Applications of Information Theory to Psychology. New York: Holt, Rinehart and Winston.

BAAYEN, R. H., R. SCHREUDER, N. H. DE JONG, and A. KROTT. in press. Dutch inflection: the rules that prove the exception. Storage and computation in the language faculty, ed. by S. Nooteboom, F. Weerman, and F. Wijnen. Dordrecht: Kluwer Academic Publishers.

BAAYEN, R.H., R. PIEPENBROCK, and L. GULIKERS. 1995. The CELEX Lexical Database (Release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

BALOTA, D. A., and J. I. CHUMBLEY. 1984. Are lexical decisions a good measure of lexical access? the role of word frequency in the neglected decision stage. Journal of Experimental Psychology: Human Perception and Performance 10.340–357.

BATES, ELIZABETH, and ANTONELLA DEVESCOVI. 1989. Crosslinguistic studies of sentence production. The crosslinguistic study of sentence processing, ed. by Brian MacWhinney and Elizabeth Bates, 225–256. Cambridge: Cambridge University Press.

BATES, ELIZABETH, and BRIAN MACWHINNEY. 1989. Functionalism and the competition model. The crosslinguistic study of sentence processing, ed. by Brian MacWhinney and Elizabeth Bates, 3–76. Cambridge: Cambridge University Press.

BELL, ALAN, DANIEL JURAFSKY, ERIC FOSLER-LUSSIER, CYNTHIA GIRAND, MICHELLE GREGORY, and DANIEL GILDEA, 2001. Form variation of english function words in conversation. Submitted manuscript.

BEVER, THOMAS G. 1970. The cognitive basis for linguistic structures. Cognition and the development of language, ed. by John R. Hayes, 279–352. New York: Wiley.

BOD, RENS, 2000. The storage vs. computation of three-word sentences. Talk presented at AMLaP-2000.

——, 2001. Sentence memory: Storage vs. computation of frequent sentences. Talk presented at CUNY 2001.

BOOTH, TAYLOR L. 1969. Probabilistic representation of formal languages. IEEE Conference Record of the 1969 Tenth Annual Symposium on Switching and Automata Theory, 74–81.

BRANTS, THORSTEN. 1999. Cascaded markov models. Proceedings of the 9th Conference of the European Chapter of the ACL (EACL-99), Bergen, Norway. ACL.

BRENT, MICHAEL R., and TIMOTHY A. CARTWRIGHT. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. Cognition 61.93–125.

BURGESS, CURT, and S. C. HOLLBACH. 1988. A computational model of syntactic ambiguity as a lexical process. COGSCI-88, 263–269.

BURGESS, CURT, and KEVIN LUND. 1994. Multiple constraints in syntactic ambiguity resolution: A connectionist account of psycholinguistic data. COGSCI-94, Atlanta, GA.

BUSH, NATHAN, 1999. The predictive value of transitional probability for word-boundary palatalization in English. Master's thesis, University of New Mexico, Albuquerque, NM.

BYBEE, JOAN, and JOANNE SCHEIBMAN. 1999. The effect of usage on degrees of constituency: the reduction of *don't* in English. Linguistics 37.575–596.

BYBEE, JOAN L. 2000. The phonology of the lexicon: evidence from lexical diffusion. Usage-based models of language, ed. by Michael Barlow and Suzanne Kemmer, 65–85. Stanford: CSLI.

CHATER, NICK, MATTHEW J. CROCKER, and MARTIN J. PICKERING. 1998. The rational analysis of inquiry: the case of parsing. Rational models of cognition, ed. by Mike Oaksford and Nick Chater, 441–468. Oxford: Oxford University Press.

CHURCH, KENNETH W. 1988. A stochastic parts program and noun phrase parser for unrestricted text. Second Conference on Applied Natural Language Processing, 136–143. ACL.

CLIFTON, JR., CHARLES, LYN FRAZIER, and CYNTHIA CONNINE. 1984. Lexical expectations in sentence comprehension. Journal of Verbal Learning and Verbal Behavior 23.696–708.

CONNINE, CYNTHIA, FERNANDA FERREIRA, CHARLIE JONES, CHARLES CLIFTON, and LYN FRAZIER. 1984. Verb frame preference: Descriptive norms. Journal of Psycholinguistic Research 13.307–319.

CORLEY, STEFFAN, and MATTHEW W. CROCKER. 1996. Evidence for a tagging model of human lexical category disambiguation. COGSCI-96, 272–277.

——, and ——. 2000. The modular statistical hypothesis: Exploring lexical category ambiguity. Architectures and mechanims for language processing, ed. by Matthew W. Crocker, Martin Pickering, and Charles Clifton, 135–160.

CROCKER, MATTHEW W., and THORSTEN BRANTS. 2000. Wide-coverage probabilistic sentence processing. Journal of Psycholinguistic Research 29.647–669.

CUETOS, FERNANDO, DON C. MITCHELL, and MARTIN M. B. CORLEY. 1996. Parsing in different languages. Language processing in spanish, ed. by Manuel Carreiras, José E. García-Albea, and Núria Sebastián-Gallés, 156–187. Hillsdale, NJ: Lawrence Erlbaum.

D'ARCAIS, G. B. FLORES. 1993. The comprehension and semantic interpretation of idioms. Idioms: Processing, structure, and interpretation, ed. by Cristina Cacciari and Patrizia Tabossi, 79–98. New Jersey: Lawrence Erlbaum.

DE JONG, N. H., L. B. FELDMAN, R. SCHREUDER, M. PASTIZZO, and R. H. BAAYEN. 2001. The processing and representation of Dutch and English compounds: Peripheral morphological, and central orthographic effects. Brain and Language p. in press.

DELL, GARY. 1990. Effects of frequency and vocabulary type on phonological speech errors. Language and Cognitive Processes 5.313–349.

DELL, GARY S. 1986. A spreading activation theory of retrieval in sentence production. Psychological Review 93.283–321.

——, MYRNA F. SCHWARTS, NADINE MARTIN, ELEANOR SAFFRAN, and DEBORAH A. GAGNON. 1997. Lexical access in aphasic and nonaphasic speakers. Psychological Review 104.801–838.

DESMET, TIMOTHY, MARC BRYSBAERT, and CONSTANTIJN DE BAECKE. 2001. The correspondence between sentence production and corpus frequencies in modifier attachment. Quarterly Journal of Experimental Psychology . In Press.

FIDELHOLZ, JAMES. 1975. Word frequency and vowel reduction in English. Cls-75, 200–213. University of Chicago.

Filip, Hana, Michael K. Tanenhaus, Gregory N. Carlson, Paul D. Allopenna, and Joshua Blatt. 2002. Reduced relatives judged hard require constraint-based analyses. Sentence processing and the lexicon: formal, computational, and experimental perspectives, ed. by Paola Merlo and Suzanne Stevenson. Amsterdam: Benjamins.

Fodor, Janet Dean. 1978. Parsing strategies and constraints on transformations. Linguistic Inquiry 9.427–473.

Ford, Marilyn, Joan Bresnan, and Ronald M. Kaplan. 1982. A competence-based theory of syntactic closure. The mental representation of grammatical relations, ed. by Joan Bresnan, 727–796. Cambridge, MA: MIT Press.

Forster, K., and S. Chambers. 1973. Lexical access and naming time. Journal of Verbal Learning and Verbal Behavior 12.627–635.

Francis, W. Nelson, and Henry Kučera. 1982. Frequency Analysis of English Usage. Boston: Houghton Mifflin.

Gahl, Susanne. 1999. Unergative, unaccusative, (in)transitive and (in)frequent. Proceedings of the 34th annual meeting of the chicago linguistic society. Chicago: University of Chicago.

Garnsey, S. M, N. J. Pearlmutter, E. Myers, and M. A. Lotocky. 1997. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. Journal of Memory and Language 37.58–93.

Gibson, Edward, 1991. A Computational Theory of Human Linguistic Processing: Memory Limitations and Processing Breakdown. Pittsburgh, PA: Carnegie Mellon University dissertation.

——. 1998. Linguistic complexity: Locality of syntactic dependencies. Cognition 68.1–76.

——, Carson Schütze, and Ariel Salomon. 1996. The relationship between the frequency and the processing complexity of linguistic structure. Journal of Psycholinguistic Research 25.59–92.

Gillund, G., and R. M. Shiffrin. 1984. A retrieval model for both recognition and recall. Psychological Review 91.1–67.

Glymour, C., and P. W. Cheng. 1998. Causal mechanism and probability: A normative approach. Rational models of cognition, ed. by Mike Oaksford and Nick Chater, 296–313. Oxford: Oxford University Press.

Godfrey, J., E. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing (IEEE ICASSP-92), 517–520, San Francisco. IEEE.

GREENBERG, STEVEN, DAN ELLIS, and JOY HOLLENBACK. 1996. Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. Proceedings of the International Conference on Spoken Language Processing (ICSLP-96), S24–27, Philadelphia, PA.

GREGORY, MICHELLE L., 2001. Linguistic Informativeness and Speech Production: An Investigation of Contextual and Discourse-pragmatic Effects on Phonological Variation. Boulder, CO: University of Colorado, Boulder dissertation.

——, WILLIAM D. RAYMOND, ALAN BELL, ERIC FOSLER-LUSSIER, and DANIEL JURAFSKY. 1999. The effects of collocational strength and contextual predictability in lexical production. CLS-99. Chicago: University of Chicago.

GRODZINSKY, Y. 2000. The neurology of syntax: language use without brocas area. target article with 36 commentaries. Behavioral and Brain Sciences 23.47–117.

GROSJEAN, F. 1980. Spoken word recognition processes and the gating paradigm. Perception and Psychophysics 28.267–283.

HALE, JOHN. 2001. A probabilistic earley parser as a psycholinguistic model. Proceedings of NAACL-2001.

HALE, KEN, and SAMUEL J. KEYSER. 1993. On argument structure and the lexical expression of syntactic relations. The view from building 20: Essays in linguistics in honor of sylvain bromberger, ed. by Ken Hale and Samuel J. Keyser, 53–109. Cambridge, MA: MIT Press.

HAY, J. B., 2000. Causes and consequences of word structure. Northwestern University dissertation.

HOOPER, JOAN B. 1976. Word frequency in lexical diffusion and the source of morphophonological change. Current progress in historical linguistics, ed. by W. Christie, 96–105. Amsterdam: North Holland.

HOWES, DAVID. 1957. On the relation between the intelligibility and frequency of occurrence of English words. Journal of the Acoustical Society of America 29.296–305.

HOWES, DAVIS H., and RICHARD L. SOLOMON. 1951. Visual duration threshold as a function of word-probability. Journal of Experimental Psychology 41.401–410.

JELINEK, FREDERICK, and JOHN D. LAFFERTY. 1991. Computation of the probability of initial substring generation by stochastic context-free grammars. Computational Linguistics 17.315–323.

JENNINGS, F., B. RANDALL, and L. K. TYLER. 1997. Graded effects of verb subcategory preferences on parsing: Support for constraint-satisfaction models. Language and Cognitive Processes 12.485–504.

JESCHENIAK, J. D., and W. J. M. LEVELT. 1994. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. Journal of Experimental Psychology: Learning, Memory and Cognition 20.824–843.

JORDAN, MICHAEL I. (ed.) 1999. Learning in Graphical Models. Cambridge, MA: MIT Press.

JULIANO, CORNELL, and MICHAEL K. TANENHAUS. 1993. Contingent frequency effects in syntactic ambiguity resolution. COGSCI-93, 593–598, Boulder, CO.

JURAFSKY, DANIEL, 1992. An On-line Computational Model of Human Sentence Interpretation: A Theory of the Representation and Use of Linguistic Knowledge. Berkeley, CA: University of California dissertation. Available as University of California at Berkeley Computer Science Division Tech. rep. #92/676.

——. 1996. A probabilistic model of lexical and syntactic access and disambiguation. Cognitive Science 20.137–194.

——. 2001. Pragmatics and computational linguistics. Handbook of pragmatics, ed. by Laurence R. Horn and Gregory Ward. Blackwell.

——, ALAN BELL, and CYNTHIA GIRAND. 2001 (to appear). The role of the lemma in form variation. Papers in laboratory phonology 7, ed. by Natasha Warner and Carlos Gussenhoven.

——, ALAN BELL, MICHELLE GREGORY, and WILLIAM D. RAYMOND. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. Frequency and the emergence of linguistic structure, ed. by Joan Bybee and Paul Hopper. Amsterdam: Benjamins. To appear.

——, and JAMES H. MARTIN. 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall.

KEGL, JUDY. 1995. Levels of representation and units of access relevant to agrammatism. Brain & Language 151–200.

KIM, AL, BANGALORE SRINIVAS, and JOHN TRUESWELL. 2001. The convergence of lexicalist perspectives in psycholinguistics and computational lingusitics. Sentence processing and the lexicon: formal, computational, and experimental perspectives, ed. by Paola Merlo and Suzanne Stevenson. Amsterdam: Benjamins.

KINTSCH, WALTER. 1970. Models for free recall and recognition. Models of human memory, ed. by Donald A. Norman. New York: Academic Press.

KRUG, MANFRED. 1998. String frequency: A cognitive motivating factor in coalescence, language processing, and linguistic change. Journal of English Linguistics 26.286–320.

KUČERA, HENRY, and W. NELSON FRANCIS. 1967. Computational analysis of present-day American English. Providence, RI: Brown University Press.

LEVELT, W. J. M., A. ROELOFS, and A. S. MEYER. 1999. A theory of lexical access in speech production. Behavioral and Brain Science 22.1–75.

LI, PING, and MICHAEL C. YIP. 1996. Lexical ambiguity and context effects in spoken word recognition: Evidence from Chinese. COGSCI-96, 228–232.

LUCE, R. D. 1959. Individual Choice Behavior. New York: Wiley.

MACDONALD, MARYELLEN C. 1993. The interaction of lexical and syntactic ambiguity. Journal of Memory and Language 32.692–715.

—— 1994. Probabilistic constraints and syntactic ambiguity resolution. Language and Cognitive Processes 9.157–201.

——, NEAL J. PEARLMUTTER, and MARK S. SEIDENBERG. 1994. The lexical nature of syntactic ambiguity resolution. Psychological Review 101.676–703.

MACWHINNEY, B., E. BATES, and R. KLIEGL. 1984. Cue validity and sentence interpretation in English, German, and Italian. Journal of Verbal Learning and Verbal Behavior 23.127–150.

MACWHINNEY, BRIAN, and ELIZABETH BATES. 1989. The Crosslinguistic Study of Sentence Processing. Cambridge: Cambridge University Press.

MCCLELLAND, JAMES L. 1998. Connectionist models and Bayesian inference. Rational models of cognition, ed. by Mike Oaksford and Nick Chater, 21–53. Oxford: Oxford University Press.

MCDONALD, J. L. 1986. The development of sentence comprehension strategies in English and Dutch. Journal of Experimental Child Psychology 41.317–335.

MCDONALD, JANET, and BRIAN MACWHINNEY. 1989. Maximum likelihood models for sentence processing. The crosslinguistic study of sentence processing, ed. by Brian MacWhinney and Elizabeth Bates, 397–421. Cambridge: Cambridge University Press.

MCDONALD, SCOTT, RICHARD SHILLCOCK, and CHRIS BREW, 2001. Low-level predictive inference in reading: Using distributional statistics to predict eye movements. Poster presented at AMLaP-2001, Saarbruecken. September 20-22, 2001.

MCRAE, KEN, MICHAEL J. SPIVEY-KNOWLTON, and MICHAEL K. TANENHAUS. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. Journal of Memory and Language 38.283–312.

MERLO, PAOLA. 1994. A corpus-based analysis of verb continuation frequencies for syntact ic processing. Journal of Psycholinguistic Research 23.435–457.

MILNE, ROBERT WILLIAM. 1982. Predicting garden path sentences. Cognitive Science 6.349–374.

MITCHELL, DON C. 1994. Sentence parsing. Handbook of psycholinguistics, ed. by Morton Ann Gernsbacher, 375–409. San Diego, CA: Academic Press.

——, and MARC BRYSBAERT. 1998. Challenges to recent THeories of crosslinguistic variation in parsing: Evidence from Dutch. Syntax and semantics 31: Sentence processing: A crosslinguistic perspective, ed. by Dieter Hillert, 313–344. San Diego, CA: Academic Press.

MOORE, ROBERT, DOUGLAS APPELT, JOHN DOWDING, J. MARK GAWRON, and DOUGLAS MORAN. 1995. Combining linguistic and statistical knowledge sources in natural-language processing for ATIS. Proceedings of the January 1995 ARPA Spoken Language Systems Technology Workshop, 261–264, Austin, TX. Morgan Kaufmann.

MORTON, J. 1969. Interaction of information in word recognition. Psychological Review 76.165–178.

NARAYANAN, SRINI, and DANIEL JURAFSKY. 1998. Bayesian models of human sentence processing. COGSCI-98, 752–757, Madison, WI. Lawrence Erlbaum.

——, and ——, 2001. A Bayesian model predicts human parse preference and reading times in sentence processing. Submitted to NIPS-2001.

OLDFIELD, R. C., and A. WINGFIELD. 1965. Response latencies in naming objects. Quarterly Journal of Experimental Psychology 17.273–281.

PAN, SHIMEI, and JULIA HIRSCHBERG. 2000. Modeling local context for pitch accent prediction. Proceedings of the 38th Annual Conference of the Association for Computational Linguistics (ACL-00), 233–240, Hong Kong. ACL.

PEARL, JUDEA. 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, Ca.: Morgan Kaufman.

PEARLMUTTER, NEAL, K. DAUGHERTY, MARYELLEN MACDONALD, and MARK SEIDENBERG. 1994. Modeling the use of frequency and contextual biases in sentence processing. COGSCI-94, Atlanta, GA.

PICKERING, MARTIN J., MATTHEW J. TRAXLER, and MATTHEW W. CROCKER. 2000. Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. Journal of Memory and Language 43.447–475.

PIERREHUMBERT, JANET B. 2001. Exemplar dynamics: Word frequency, lenition and contrast. Frequency and the emergence of linguistic structure, ed. by Joan Bybee and Paul Hopper. Amsterdam: Benjamins. To appear.

PRITCHETT, BRADLEY. 1988. Garden path phenomena and the grammatical basis of language processing. Language 64.539–576.

RAAIJMAKERS, J. G. W., and R. M. SHIFFRIN. 1981. Search of associative memory. Psychological Review 88.93–134.

REHDER, BOB. 1999. A causal model theory of categorization. COGSCI-99, 595–600, Vancouver, British Columbia.

ROLAND, DOUG, and DANIEL JURAFSKY. 1998. How verb subcategorization frequencies are affected by corpus choice. COLING/ACL-98, 1122–1128, Montreal. ACL.

——, and ——. 2001. Verb sense and verb subcategorization probabilities. Sentence processing and the lexicon: formal, computational, and experimental perspectives, ed. by Paola Merlo and Suzanne Stevenson. Amsterdam: Benjamins.

ROLAND, DOUGLAS, 2001. Verb Sense and Verb Subcategorization Probabilities. University of Colorado, Boulder dissertation.

——, DANIEL JURAFSKY, LISE MENN, SUSANNE GAHL, ELIZABETH ELDER, and CHRIS RIDDOCH. 2000. Verb subcategorization frequency differences between business-news and balanced corpora: the role of verb sense. Proceedings of the Association for Computational Linguistics (ACL-2000) Workshop on Comparing Corpora, Hong Kong.

RUBENSTEIN, H., L. GARFIELD, and J. A. MILLIKAN. 1970. Homographic entries in the internal lexicon. Journal of Verbal Learning and Verbal Behavior 9.487–494.

SAFFRAN, J. R., E. L. NEWPORT, and R. N. ASLIN. 1996a. Statistical learning by 8-month old infants. Science 274.1926–1928.

——, ——, and ——. 1996b. Word segmentation: The role of distributional cues. Journal of Memory and Language 35.606–621.

SAFFRAN, JENNY R. 2001. The use of predictive dependencies in language learning. Journal of Memory and Language 44.493–515.

——, RICHARD N. ASLIN, and ELISSA L. NEWPORT. 1996c. Statistical cues in language acquisition: Word segmentation by infants. COGSCI-96, 376–380.

SAVIN, H. B. 1963. Word-frequency effect and errors in the perception of speech. Journal of the Acoustical Society of America 35.200–206.

SCHUCHARDT, HUGO. 1885. Über die Lautgesetze: Gegen die Junggrammatiker. Berlin: Robert Oppenheim. Excerpted with English translation in Theo Vennemann and Terence H. Wilbur, (Eds.), *Schuchardt, the Neogrammarians, and the Transformational Theory of Phonological Change*, Athenaum Verlag, Frankfurt, 1972.

SEIDENBERG, MARK S., and MARYELLEN C. MACDONALD. 1999. A probabilistic constraints approach to language acquisition and processing. Cognitive Science 23.569–588.

SIMPSON, GREG B., and CURT BURGESS. 1985. Activation and selection processes in the recognition of ambiguous words. Journal of Experimental Psychology: Human Perception and Performance 11.28–39.

SPIVEY, MICHAEL J., and MICHAEL K. TANENHAUS. 1998. Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. Journal of Experimental Psychology: Learning, Memory, and Cognition 24.1521–1543.

SPIVEY-KNOWLTON, M., and J. SEDIVY. 1995. Resolving attachment ambiguities with multiple constraints. Cognition 55.227–267.

SPIVEY-KNOWLTON, M., J. TRUESWELL, and M. K. TANENHAUS. 1993. Context effects in syntactic ambiguity resolution: Discourse and semantic influences in parsing reduced relative clauses. Canadian Journal of Experimental Psychology 47.276–309.

SPIVEY-KNOWLTON, MICHAEL J., 1996. Integration of visual and linguistic information: Human data and model simulations. University of Rochester dissertation.

STALLINGS, LUNNE M., MARYELLEN C. MACDONALD, and PADRAIG G. O'SEAGHDHA. 1998. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. Journal of Memory and Language 39.392–417.

STEVENSON, S., and P. MERLO. 1997. Lexical structure and parsing complexity. Language and Cognitive Processes 12.349–399.

STEVENSON, SUZANNE. 1994. Competition and recency in a hybrid network model of syntactic disambiguation. Journal of Psycholinguistic Research 23.295–322.

STOLCKE, ANDREAS. 1995. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. Computational Linguistics 21.165–202.

TABOR, WHITNEY, CORNELL JULIANO, and MICHAEL K. TANENHAUS. 1997. Parsing in a dynamical system. Language and Cognitive Processes 12.211–272.

TABOSSI, PATRIZIA, MICHAEL SPIVEY-KNOWLTON, KEN MCRAE, and MICHAEL K. TANENHAUS. 1994. Semantic effects on syntactic ambiguity resolution: Evidence for a constraint-based resolution process. Attention and performance XV, ed. by C. Umilta and M. Moscovitch, 589–615. Hillsdale, NJ: Lawrence Erlbaum.

TANENHAUS, MICHAEL K., MICHAEL J. SPIVEY-KNOWLTON, and JOY E. HANNA. 2000. Modeling thematic and discourse context effects with a multiple constraints approach: Implications for the architecture of the language comprehension system. Architectures and mechanims for language processing, ed. by Matthew W. Crocker, Martin Pickering, and Charles Clifton, 90–118.

TANENHAUS, MICHAEL K., LAURIE A. STOWE, and GREG CARLSON. 1985. The interaction of lexical expectation and pragmatics in parsing filler-gap constructions. COGSCI-85, 361–365, Irvine, CA.

TENENBAUM, JOSHUA B. 2000. Bayesian modeling of human concept learning. Advances in Neural Information Processing Systems 11.

——, and THOMAS L. GRIFFITHS. 2001a. Generalization, similarity, and bayesian inference. Behavioral and Brain Sciences 24.

——, and ——. 2001b. The rational basis of representativeness. COGSCI-01, Edinburgh.

——, and F. XU. 2000. Word learning ad bayesian inference. COGSCI-00.

TRUESWELL, JOHN C. 1996. The role of lexical frequency in syntactic ambiguity resolution. Journal of Memory and Language 35.566–585.

——, and MICHAEL K. TANENHAUS. 1994. Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. Perspectives on sentence processing, ed. by Charles Clifton, Jr., Lyn Frazier, and Keith Rayner, 155–179. Hillsdale, NJ: Lawrence Erlbaum.

——, ——, and SUSAN M. GARNSEY. 1994. Semantic influences on parsing: Use of of thematic role information in syntactic ambiguity resolution. Journal of Memory and Language 33.285–318.

——, ——, and CHRISTOPHER KELLO. 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.

TYLER, LORRAINE K. 1984. The structure of the initial cohort: Evidence from gating. Perception & Psychophysics 36.417–427.

WHALEY, C. P. 1978. Word–nonword classification time. Journal of Verbal Language and Verbal Behavior 17.143–154.

WINGFIELD, A. 1968. Effects of frequency on idenitifcaiton American Journal of Psychology 81.226–234.