Modular Architectures and Statistical Mechanisms: The Case from Lexical Category Disambiguation[1]

Matthew W. Crocker
Computational Linguistics
Saarland University
D-66041 Saarbrücken, Germany
crocker@coli.uni-sb.de

Steffan Corley
Sharp Laboratories of Europe
Oxford Science Park
Oxford, U.K.
steffan@sharp.co.uk

Abstract

This paper reviews the modular, statistical model of human lexical category disambiguation (SLCM) proposed by Corley and Crocker (2000). The SLCM is distinct lexical category disambiguation mechanism within the human sentence processor, which uses word-category frequencies and category bigram frequencies for the initial resolution of category (part-of-speech) ambiguities. The model has been shown to account for a range of existing experimental findings in relatively diverse constructions. This paper presents the results of two new experiments that directly confirm the predictions of the model. The first experiment demonstrates the dominant role of word-category frequency in resolving noun-verb ambiguities. The second experiment then presents evidence for the modularity of the mechanism, by demonstrating that immediately available syntactic context does not override the SLCMs initial decision.

[1] This paper presents entirely joint work, and the order of authors is arbitrary. Correspondence should be sent in the first instance to M. Crocker (crocker@coli.uni-sb.de). The authors would like to express particular thanks to Charles Clifton, Jr. and Martin Corley for their invaluable assistance. The authors gratefully acknowledge the support of an ESRC Research Fellowship (to Matthew Crocker, #H5242700394) and an ESRC Studentship (to Steffan Corley, #R00429334081), both of which were held at the Centre for Cognitive Science, University of Edinburgh.

Introduction

This paper reconsiders the nature of modular architectures in the light of recent empirical, theoretical and computational developments concerning the exploitation of statistical language processing mechanisms. We defend a simpler notion of modularity than that proposed by Fodor (1983). Given current conflicting theoretical arguments and empirical evidence for and against modularity, we argue for modularity strictly on computational and methodological grounds. We then apply this to a particular aspect of human language processing: the problem of lexical category disambiguation.

While previous work has often focused on the kinds of linguistic knowledge which are used in ambiguity resolution, we focus on the role of statistical, or frequency-based, knowledge. While such mechanisms are now a common element of non-modular, constraint-based models (see Tanenhaus *et al* (in press)), we argue that probabilistic mechanisms may be naturally associated with modular architectures. In particular, we suggest that a Statistical Lexical Category Module (SLCM) provides an extremely efficient and accurate solution to the sub-problem of lexical category disambiguation. Following a summary of the model and how it accounts for the range of relevant existing data, we review the results of two new experiments that test the predictions of both the statistical and modular aspects of the SLCM, and provide further support for our proposals.

Modularity, Constraints and Statistics

The issue of modularity continues to be a hotly debated topic within the sentence processing literature.[2] Parser-based models of human sentence processing led to the tacit emergence of syntactic modularity, which was then rationally defended by Fodor (1983). In particular, Fodor argued that cognitive faculties are divided into input processes, which are modular, and central processes, which are not. The divide between input and central processes is roughly coextensive with the divide between perception and cognition; in the case of language, Fodor located this divide between the subject matter of formal linguistics and that of pragmatics and discourse analysis.

Recently, their has been a shift in consensus towards more interactionist,

[2] See Crocker (1999) for a more complete introduction to the issues presented in this section.

non-modular positions. The term ‘constraint-based’ is often used to denote such an interactionist position. The constraint-based position is tacitly assumed to imply that all constraints can in principle apply immediately and simultaneously, across all levels of linguistic representation, and possibly even across perceptual faculties (Tanenhaus et al, 1995).

Modular and interactive positions are often associated with other computational properties. Spivey-Knowlton and Eberhard (1996) argue that modular positions tend to be symbolic, binary, unidirectional and serial. In contrast, interactive models tend to be distributed, probabilistic, bi-directional and parallel. Further, Spivey-Knowlton and Eberhard suggest that “when a model is specified in enough detail to be associated with a region in this space, that region’s projection onto the continuum of modularity indicates *the degree to which* a model is modular” (pp. 39 – 40, their italics).

Spivey-Knowlton and Eberhard’s position turns a historical accident into a definition. While existing models do pattern approximately along the lines they propose, we suggest that their characterisation inaccurately represents the underlying notion of modularity.[3] We propose a simplified definition of modularity that is independent of any commitment to orthogonal issues such as the symbolic-distributed, binary-probabilistic, unidirectional-bidirectional and serial-parallel nature of a particular theory. Rather our definition focuses purely on information-flow characteristics:

- A module can only process information stated in its own representational and informational vocabulary. For example, the syntactic processor can only make use of grammatical information.
- A module is independently predictive. That is, we do not need to know about any other component of the cognitive architecture to make predictions about the behaviour of a module (provided we know the module’s input).
- A module has low bandwidth in both feedforward and feedback connections. By this we mean that it passes a comparatively small amount of information (compared to its internal bandwidth) on to subsequent and prior modules.

[3] Of course their characterisation does define a particular computational position which one might dub ‘modular’, but the falsification of that position crucially does not falsify the general notion of modularity, only the particular position they define.

These three defining properties of a modular architecture overlap. If one module cannot understand the representational vocabulary of another, then information about its internal decision process is of no use; thus the cost of passing such information on would not be warranted. Similarly, a module cannot be independently predictive if its decisions depend on representations constructed by other modules that are not part of its input — independent prediction is therefore directly tied to low bandwidth feedback connections.

In sum, we propose a simple definition of modularity in which modules process a specific representation and satisfy the relevant constraints which are defined for that level of representation. Modules have high internal bandwidth and are connected to each other by relatively low bandwidth: the lower the bandwidth, the greater the modularity. This definition is independent of whether we choose to state our modules in more distributed or symbolic terms, as it should be.

Statistical Mechanisms

In the previous section, we noted Spivey-Knowlton and Eberhard's (1996) claim that modularity is normally associated with binary rather than probabilistic decision procedures. This claim derives largely from the association of constraint-based architectures with connectionist implementations (Tanenhaus *et al*, in press; MacDonald *et al*, 1994) which in turn have a natural tendency to exhibit frequency effects. We proposed a definition of modularity which is consistent with statistical mechanisms. In this section, we argue that modularity and statistical mechanisms are in fact natural collaborators.

The motivation for modularity is essentially one of computational compromise, based on the assumption that an unrestricted constraint-satisfaction procedure could neither operate in real-time (Fodor, 1983), nor could it acquire such a heterogeneous system of constraints in the first place (Norris, 1990). It is still reasonable to assume however, that modules will converge on highly effective processing mechanisms; that is, a mechanism which can accurately and rapidly arrive at the correct analysis of the input, based on the restricted knowledge available within the module. For purposes of disambiguation, the module should therefore use the best heuristics it can, again modulo any computational and informational limitations.

In the spirit of rational analysis (Anderson, 1991), one might therefore choose to reason about such a mechanism as an optimal process in probabilistic terms. This approach has been exploited both in the study of human sentence processing (Chater et al, 1999; Jurafsky, 1996) and in computational linguistics where statistical language models have been effectively applied to problems of speech recognition, part-of-speech tagging, and parsing (see Charniak (1993; 1997) for an overview). We propose a specific hypothesis, in which modules may make use of statistical mechanisms in their desire to perform as effectively as possible in the face of restricted knowledge. We define statistical modularity by introducing the 'Modular Statistical Hypothesis' (MSH):

> **The Modular Statistical Hypothesis**: The human sentence processor is composed of a number of modules, at least some of which use statistical mechanisms. Statistical results may be communicated between modules, but statistical processes are restricted to operating within, and not across, modules.

This hypothesis encompasses a range of possible models, including the coarse-grained architecture espoused by proponents of the Tuning Hypothesis (Mitchell *et al*, 1995; Mitchell & Brysbaert, to appear). However, it excludes interactive models such as those proposed by MacDonald et al. (1994), Tanenhaus *et al* (in press) and Jurafsky (1996) – despite their probabilistic nature – since the models that fall within the MSH are a necessarily subset of those that are modular.

In the case of a statistical module we assume that heuristic decision strategies are based on statistical knowledge accrued by the module, presumably on the basis of linguistic experience. Assuming that the module collates statistics itself, it must have access to some measure of the 'correctness' of its decision; this could be informed by whether or not reanalysis was requested by later processes. The most restrictive modular statistical model is therefore one in which modules are fully encapsulated and only offer a single analysis to higher levels of processing.

The statistical measures such a module depends on are thus architecturally limited. Such measures can not directly reflect information pertaining to higher levels of processing, as these are not available to the module. Assuming very low bandwidth feedforward connections, or shallow output, it is also impossible for the module to collate statistics concerning levels of representation that are the province of modules that precede it. A modular architecture therefore constrains the representations for which statistics may be accrued, and subsequently used to inform decision making processes; this contrasts with an interactive architecture, where there are no such constraints on the decision process.

It is worth noting that we have argued for the use of statistical mechanisms in modular architectures on primarily *rational* grounds. That is, such statistical mechanisms have been demonstrated to provide highly effective heuristic decisions in the absence of full knowledge, and their use is therefore highly strategic, not accidental. Indeed, it might even be argued that such mechanisms give good approximations of 'higher-level' knowledge. For example, simple word bigrams will model those words that co-occur frequently or infrequently. Since highly semantically plausible collocations are likely to be more frequent than less plausible ones, such statistics can appear to be modelling semantic knowledge, as well as just the distribution of word types.

In contrast, constraint-based, interactionist models motivate the existence of frequency effects as an essentially unavoidable consequence of the underlying connectionist architecture (see Seidenberg (1997) for general discussion), along with other factors such as neighbourhood effects. Interestingly, this may lead to some rather strong predictions. Since such mechanisms are highly sensitive to frequency, they would seem to preclude probabilistic mechanisms that do not select a "most-likely" analysis based on these prior frequencies. Pickering *et al* (2000), however, present evidence against likelihood-based accounts, and propose and alternative probabilistic model based on a rational analysis of the parsing problem (Chater *et al*, 1999).

Lexical Category Ambiguity

The debate concerning the architecture of the human language processor has typically focused on the syntax-semantics divide. Here, however, we consider the problem of lexical category ambiguity, and argue for the plausibility of a distinct lexical category disambiguation module. Lexical category ambiguity occurs when a word can be assigned more than one part of speech (noun, verb, adjective etc.). Consider, for example, the following sentence:

(1) He saw her duck.

There are two obvious, plausible readings for sentence 1. In one reading, ‘her’ is a possessive pronoun and ‘duck’ is a noun (cf. 2a); in the other reading, ‘her’ is a personal pronoun and ‘duck’ is a verb (cf. 2b).

(2) a) He saw her$_{\text{POSS}}$ apple.
b) He saw her$_{\text{PRON}}$ leave.

Lexical Category Ambiguity and Lexical Access

Lexical access is the stage of processing at which lexical entries for input words are retrieved. Evidence suggests that multiple meanings for a given word are activated even when semantic context biases in favour of a single meaning (Swinney, 1979; Seidenberg et al., 1982; but see Kawamoto (1993) for more thorough discussion). The evidence does not, however, support the determination of grammatical class during lexical access. Tanenhaus, Leiman and Seidenberg (1979) found that when subjects heard sentences such as those in (3), containing a locally ambiguous word in an unambiguous syntactic context, they were able to name a target word which was semantically related to either of the possible meanings of the ambiguous target (e.g. SLEEP or WHEEL) faster than they were able to name an unrelated target.

(3) a) John began to tire.
b) John lost the tire.

This suggests that words related to both meanings had been primed; both meanings must therefore have been accessed, despite the fact that only one was compatible with the syntactic context. Seidenberg, Tanenhaus, Leiman and Bienkowski (1982) replicated these results, and Tanenhaus and Donnenworth-Nolan (1984) demonstrated that they could not be attributed to the ambiguity (when spoken) of the word ‘to’ or to subjects inability to integrate syntactic information fast enough prior to hearing the ambiguous word.

Such evidence is consistent with a model in which lexical category disambiguation occurs after lexical access. The tacit assumption in much of the sentence processing literature has been that grammatical classes are determined during parsing (see Frazier (1978) and Pritchett (1992) as examples). If grammar terminals are words rather than lexical categories, then such a model requires no augmentation of the parsing mechanism. Alternatively, Frazier and Rayner (1987) proposed that lexical category disambiguation has a privileged status within the parser; different mechanisms are used to arbitrate such ambiguities from those concerned with structure building.

Finally, lexical categories may be determined after lexical access, but prior to syntactic analysis. That is, lexical category disambiguation may constitute a module in its own right.

The Privileged Status of Lexical Category Ambiguity

There are essentially three possible positions regarding the relationship between syntax and lexical category.

1. Lexical categories are syntactic: The terminals in the grammar are words and it is the job of the syntactic processes to determine the lexical category that dominates each word (Frazier, 1978; Pritchett, 1992).
2. Syntactic structures are in the lexicon: The bulk of linguistic competence is in the lexicon, including rich representations of the trees projected by lexical items. Parsing is reduced to connecting trees together (MacDonald et al, 1994; Kim and Trueswell, this volume).
3. Syntax and lexical category determination are distinct: Syntax and the lexicon have their own processes responsible for initial structure building and ambiguity resolution.

If we take the latter view of lexical category ambiguities, one possibility is that a pre-syntactic modular process makes lexical category decisions. These decisions would have to be made on the basis of a simple heuristic, without the benefit of syntactic constraints. In common with all modules, such a process will make incorrect decisions when potentially available information (such as syntactic constraints) could have permitted a correct decision. It does, however, offer an extremely low cost alternative to arbitration by syntactic and other knowledge. That is, disambiguation on the basis of full knowledge potentially entails the integration of constraints of various types, across various levels of representation. It may be the case that such processes cannot converge rapidly enough on the correct disambiguated form.

For this argument to be compelling, it must also be the case that lexical category ambiguities are frequent enough to warrant a distinct resolution process. This can be verified by determining the number of words that occur with more than one category in a large text corpus. DeRose (1988) has produced such an estimate from the Brown corpus; he found that 11.5% of word types and 40% of tokens occur with more than one lexical category. As the mean length of the sentences in the Brown corpus is 19.4 words, DeRose's figures suggest that there are 7.75 categorially ambiguous words in an average corpus sentence.

Our own investigations suggest the extent of the problem is even greater. Using the TreeBank version of the Brown corpus, we discovered 10.9% ambiguity by type, and a staggering 65.8% by token. To obtain these results, we used the coarsest definition of lexical category possible — just the first letter of the corpus tag (i.e. nouns were not tagged separately as singular, plural, etc.). Given the high frequency of lexical category ambiguity, a separate decision making process makes computational sense, if it can achieve sufficient accuracy. If category ambiguities are resolved prior to parsing, the time required by the parser is reduced (Charniak *et al*, 1996).

A Statistical Lexical Category Module

In this section we outline a specific proposal for a Statistical Lexical Category Module (SLCM). The function of the SLCM is to determine the best possible assignment of lexical part-of-speech categories for the words of an input utterance, as they are encountered. The model differs from other theories of sentence processing, in that lexical category disambiguation is postulated as a distinct modular process, which occurs prior to syntactic processing but following lexical access.

We argued earlier for a model of human sentence processing that is (at least partially) statistical on both rational and empirical grounds: such a model appears sensible and has characteristics which may explain some of the behaviour patterns of the HSPM. We therefore propose that the SLCM employs a statistically-based disambiguation mechanism, as such a mechanism can operate efficiently (in linear time) and achieve near optimal performance (most words disambiguated correctly, see next section), and we assume such a module would strive for such a rational behaviour.

What Statistics?

If we accept that the SLCM is statistical, a central question concerns what statistics condition its decisions. Limitations of the modular architecture we are proposing constrain the choice. The SLCM has no access to structural representations; structurally-based statistics could therefore not be expressed in its representational vocabulary. We will assume that the input to the module is extremely shallow — just a word and a set of candidate grammatical classes. In this case, the module also has no access to low level representations including morphs, phonemes and graphic symbols; the module may only make use of statistics collated over words or lexical categories, or combinations of the two.

It seems likely that the SLCM collates statistics concerning the frequency of co-occurrence of individual words and lexical categories. One possible model is therefore that the SLCM just picks the most frequent class for each word; for reasons that will become apparent, we will call this the 'unigram' approach. The SLCM may also gather statistical information concerning prior context. For example, decisions about the most probable lexical category for a word may also consider the previous word. Alternatively, such decisions may only consider the category assigned to the previous word, or a combination of both the prior word and its category may be used.

Probability Theory and the SLCM

The problem faced by the SLCM is to incrementally assign the most likely sequence of lexical categories to a given sequence of words as they are encountered. That is, as each word is input to the SCLM, it outputs the most likely category for it. Research in computational linguistics has concentrated on a (non-incremental) version of this problem for a number of years and a number of successful and accurate 'part-of-speech taggers' have been built (e.g. Weischedel et al, 1993; Brill, 1995). While a number of heuristic tagging algorithms have been proposed, the majority of modern taggers are statistically based, relying on distributional information about language (DeRose, 1988; Weischedel et al, 1993; Ratnarparkhi, 1996; see also Charniak, 1997 for discussion). It is this set of taggers that we suggest is most suitable for an initial model of statistical lexical category disambiguation. They provide a straightforward learning algorithm based on prior experience, are comparatively simple, employ a predictive and uniform decision strategy (i.e. don't make use of arbitrary or ad hoc rules), and can be naturally adapted to assign preferred lexical category tags incrementally.

The SLCM, as with part-of-speech taggers, is based on a Hidden Markov Model (HMM), and operates by probabilistically selecting the best sequence of category assignments for an input string of words.[4] Let us briefly consider the problem of tag assignment from the perspective of probability theory. The task of the SLCM is to find the best category sequence ($t_1 \ldots t_n$) for an input sequence of words ($w_1 \ldots w_n$). We assume that the 'best' such sequence is the one that is most likely, based on our prior experience. Therefore the SLCM must find the sequence ($t_1 \ldots t_n$) such that P($t_1 \ldots t_n$, $w_1 \ldots w_n$) is maximised. That is, we want to find the tag sequence that maximises the joint probability of the tag sequence and the word sequence.

One practical problem, however, is that determining such a probability directly is difficult, if we wish to do so on the basis of frequencies in a corpus (as in the case of taggers) or in our prior experience (as would be the case for the psychological model). The reason is that we may have seen very few (or quite often no) occurrences of a particular word-tag sequence, and thus probabilities will often be estimated as zero. It is therefore common practice to approximate this probability with another which can be estimated more reliably. Corley and Crocker (2000) argue that the SLCM approximates this probability using category bigrams, as follows:

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

The two terms in the right hand side of the equation are the two statistics that we hypothesise to dominate lexical category decisions in the SLCM. *$P(w_i|t_i)$* – the unigram or word-category probability – is the probability of a word given a

[4] See Corley and Crocker (in press) or Corley (1998) for a more thorough exposition of HMM taggers and the model being assumed here. See also Charniak (1993;1997), for more general and more formal discussion.

particular tag.[5] $P(t_i|t_{i-1})$ – the bigram or category co-occurrence probability – is the probability that two tags occur next to each other in a sentence. While the most accurate HMM taggers typically use trigrams (Brants, 1999), Corley and Crocker (2000) argue that the bigram model is sufficient to explain existing data and is simpler (requires fewer statistical parameters). It is therefore to be preferred as a cognitive model, until evidence warrants a more complex model.

Estimates for both of these terms are typically based on the frequencies obtained from a relatively small training corpus in which words appear with their correct tags. This equation can be applied incrementally. That is, after perceiving each word we may calculate a contingent probability for each tag path terminating at that word; an initial decision may be made as soon as the word is seen. Figure 1 depicts tagging of the phrase “that old man”.

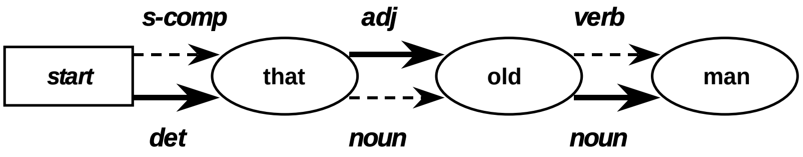

Figure 1: Tagging the sequence "that old man"

Each of the words has two possible lexical categories, meaning that there are eight tag paths. In the diagram, the most probable tag path is shown by the sequence of solid arcs. Other potential tags are represented by dotted arcs.

The tagger’s job is to find this preferred tag path. The probability of a sentence beginning with the start symbol is 1.0. When ‘that’ is encountered, the tagger must determine the likelihood of each reading for this word when it occurs sentence initially. This results in probabilities for two tag paths – start followed by a sentence complementiser and start followed by a determiner. The calculation of each of these paths is shown in Table 1.

[5] The use of P(w|t) makes the model appear top-down. See Corley (1998, pp. 85-87) for how this (apparently generative) statistical model is actually derived from an equation based on bottom-up recognition. See also Charniak (1997) for discussion.

Table 1: Tagging "that old man"; stage 1 - "that"

Path	Probability
1 scomp	*P("that"|scomp)P(scomp|start)*
2 det	*P("the"|scomp)P(det|start)*

While "that" occurs more frequently as a sentence complementiser than as a determiner in absolute terms, sentence complementisers are relatively uncommon at the beginning of a sentence. Therefore tag path 2 is likely to have a greater probability.

The next word, "old", is also category ambiguous as either an adjective or a noun. There are therefore four possible tag paths up until this point. Table 2 shows the calculations necessary to determine the probability of each of them.

Table 2: Tagging "that old man"; stage 2 - "old"

Path	Probability
1.1 scomp-adj	*P("old"|adj)P(adj|scomp)P(path1)*
1.2 scomp-noun	*P("old"|noun)P(noun|scomp)P(path1)*
2.1 det-adj	*P("old"|adj)P(adj|det)P(path2)*
2.2 det-noun	*P("old"|noun)P(noun|det)P(path2)*

In this case, "old" is far more frequently an adjective than a noun, and so this is the most likely reading. As an adjective following a determiner is more likely than one following a sentence complementiser, path 2.1 becomes far more probable than 1.1.

The process is identical when "man" is encountered. There are now eight tag paths to consider, shown in Table 3. As "man" occurs more frequently as a noun than a verb, and this reading is congruent with the preceding context, path 2.1.2 is preferred.

Table 3: Tagging "that old man"; stage 3 - "man"

Path	Probability
1.1.1 scomp-adj-verb	*P("man"|verb)P(verb|adj)P(path1.1)*
1.1.2 scomp-adj-noun	*P("man"|noun)P(noun|adj)P(path1.1)*
1.2.1 scomp-noun-verb	*P("man"|verb)P(verb|noun)P(path1.2)*
1.2.2 scomp-noun-noun	*P("man"|noun)P(noun|noun)P(path1.2)*
2.1.1 det-adj-verb	*P("man"|verb)P(verb|adj)P(path2.1)*
2.1.2 det-adj-noun	*P("man"|noun)P(noun|adj)P(path2.1)*
2.2.1 det-noun-verb	*P("man"|verb)P(verb|noun)P(path2.2)*
2.2.2 det-noun-noun	*P("man"|noun)P(noun|noun)P(path2.2)*

So far, we have assumed that it is necessary to keep track of every single tag path. This would make the algorithm extremely inefficient and psychologically implausible; as the length of the sentence grows, the number of possible tag paths increases exponentially. However, a large number of paths which will never be 'most probable' can rapidly be discarded, using a standard dynamic programming solution – the Viterbi (1967) algorithm (see Charniak (1993) for explanation). This algorithm is linear; this means that the amount of work required to determine a tag for each word is essentially constant, no matter how long the sentence is. Indeed, this property contributes directly to the psychological plausibility of this mechanism over more complex alternatives.

We have argued that taggers such as the SLCM are, in general, extremely accurate (approaching 97% – see Charniak (1997), Brants (1999)). However, they have distinctive breakdown and repair patterns. Corley and Crocker (2000) argue that these patterns are very similar to those displayed by people upon encountering sentences containing lexical category ambiguities. In particular, they show how the SLCM, when trained on a standard corpus of English, models the following experimental results:

'That' Ambiguity (Juliano & Tanenhaus, 1993): In this study, Juliano and Tanenhaus investigated the initial decisions of the HSPM when it encounters the categorially ambiguous word "that", in both sentence initial and post verbal contexts. In sentence initial position, "that" is more likely to be a determiner, while post-verbally, it is more likely to be a complementiser. Corley and Crocker provide a simulation demonstrating that the proposed bigram model accounts for the findings, while a simpler unigram model does not.

Noun-Verb Ambiguities (MacDonald, 1993): Following the study of Frazier and Rayner (1987), MacDonald investigated the processing of words that are ambiguous between noun and verb categories, e.g. as in "warehouse *fires*", to determine if semantic bias affected initial decisions. Corley and Crocker show how the SLCM can straightforwardly account for the findings. This is discussed in more detail in the next section.

Post-Ambiguity Constraints (MacDonald, 1994): Reanalysis may occur in the SLCM when the most probable tag sequence at a given point requires revising an adjacent, previous tag. Corley and Crocker (2000) demonstrate how such reanalysis in the SLCM can simulate the post-ambiguity constraints investigated by MacDonald, in which reduced relative clause constructions were rendered easier to process when the word following the ambiguous verb (simple past vs. participle) made the participle reading more likely.

New Evidence for the SLCM

The Modular Statistical Hypothesis posits the existence of identifiable subsystems within the human language processor, and argues for the use of statistical mechanisms within modules as optimal heuristic knowledge. For the task of lexical category disambiguation, we have presented a particular modular statistical mechanism. While our model accounts well for a range of relevant existing findings, as outlined in the previous section, many of those results were based on experiments designed to test rather different hypotheses, and as such provide imperfect and indirect support for the mechanism we have developed.

In this section we review two recent experimental results from Corley (1998) which directly test the central predictions of the theory. These predictions are:

- The **Statistical Lexical Category Hypothesis (SLCH)**: Initial lexical category decisions are made on the basis of frequency-based statistics.
- The **Modular Lexical Category Hypothesis (MLCH)**: Lexical category decisions are made by a pre-syntactic module.

Experiment 1 is concerned with the SLCH; it is designed to determine whether initial lexical category decisions are affected by the statistical bias of individual words. Experiment 2 more directly tests the MLCH; the experiment determines whether initial decisions are made on the basis of lexical statistical bias even in the face of strong syntactic evidence to the contrary.

Experiment 1: The Statistical Lexical Category Hypothesis

Words that are ambiguous between noun and verb readings are very common in English. Frazier and Rayner (1987) and MacDonald (1993) both employed this ambiguity in their experiments; their results were taken as support for the delay strategy and an interactive constraint-based view respectively. The SLCH simply asserts that the initial decisions of the HSPM will be strongly influenced by frequency-based statistics. For this ambiguity, all other things being equal, the HSPM will initially prefer a noun reading for a word that is frequency-biased towards a noun reading, and a verb reading for a verb-biased one.[6]

Previous studies of this ambiguity have not fully tested this hypothesis. For example, MacDonald's (1993) experimental items included only noun-biased words. In contrast, Corley (1998) produced a controlled set of experimental items in which both noun-biased and verb-biased conditions were represented. Example materials are shown below.

Experiment 1: Materials

(a) The woman said that the German *makes* the beer she likes best.
(b) The woman said that the German *makes* are cheaper than the rest.
(c) The foreman knows that the warehouse *prices* the beer very modestly.
(d) The foreman knows that the warehouse *prices* are cheaper than the others.

In (a) and (b), the ambiguous word ("makes") is biased towards a verb reading. In (a) the disambiguating region ("the beer") also favours this reading. In contrast, the disambiguating region in (b) favours a noun reading. (c) and (d) are analogous except that the ambiguous word is noun-biased.

The frequency bias of each of the ambiguous words used in this experiment was determined from the British National Corpus, chosen for both its size (100 million words) and its relatively balanced and British content. As this experiment is only designed to test whether statistical bias does have an effect, and not whether other constraints do not, only strongly biased items were used. The experimental items were further controlled to ensure that the possible noun compounds ("German makes", "warehouse prices") were plausible but infrequent and non-idiomatic. This control ensured that contextual bias effects (MacDonald, 1993) would not be expected to influence the outcome of the experiment.

[6] While the model we have presented uses $P(w_i|t_i)$ and $P(t_i|t_{i-1})$, the second measure has no effect in this experiment, where the ambiguous word always follows a noun. This is because P(noun|noun) and P(verb|noun) are approximately equal (as determined from the BNC and Brown corpora). This experiment therefore does not bear on the use of the bigram measure which was independently motivated in Corley and Crocker (in press).

If the SLCH is correct, reading times in the disambiguating region should reflect an interaction between bias and disambiguation. In other words, subjects' initial decisions should depend on the bias of the ambiguous words; we would therefore expect reading time increases reflecting reanalysis to occur only when the disambiguating region forces a reading at odds with the bias of the ambiguous word.

In contrast, a non-statistical model such as the Garden Path theory (Frazier, 1979) predicts the same initial decision in all four conditions. A main effect of disambiguation would be anticipated, but not one of bias, and no interaction between bias and disambiguation. Frazier and Rayner's (1987) delay strategy also does not predict a main effect of bias or an interaction; any main effect of disambiguation is compatible with, rather than predicted by, the strategy.

32 subjects took part in the experiment, which was performed as a self-paced reading study, using a moving window display (Just, Carpenter and Woolley, 1982). The resulting reading times were adjusted for word length using a procedure described in Ferreira and Clifton (1986).

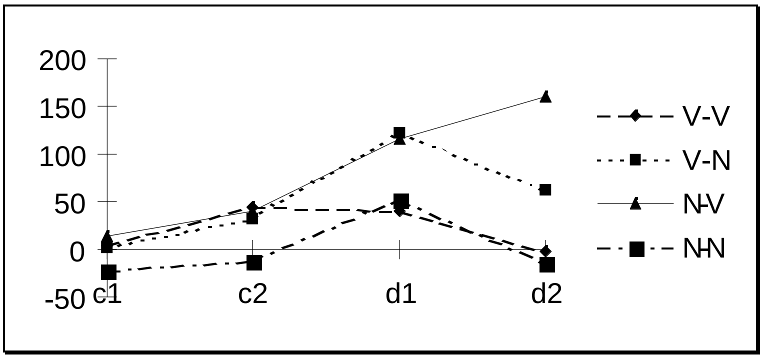

Figure 2: Experiment 1 length-adjusted reading times

Results and Discussion

Average length-adjusted reading times obtained for experiment 1 are shown in Figure 2. Here, c1 is the word preceding the ambiguous word, c2 is the ambiguous word and $d_1 \ldots d_n$ is the disambiguating region. V-V indicates that c2 is verb biased, and that the item is disambiguated as a verb, and so on.

The SLCH predicts effects at the start of the disambiguating region; the results for the first word of the disambiguating region are shown in Figure 3. These results show a highly significant interaction between bias and disambiguation (F_1 = 8.05, $p < .01$; F_2 = 27.99, $p < .001$). A planned comparison of means also revealed a highly significant difference in reading times between the verb disambiguation conditions (F_1 = 8.27, $p < .01$; F_2 = 10.86, $p < .01$) and a significant difference between the noun disambiguation conditions (F_1 = 4.72, $p < .05$; F_2 = 7.46, $p < .02$).

These results indicate that initial lexical category decisions are strongly influenced by the frequency-bias of the individual ambiguous words; the results are exactly as predicted by the SLCH and therefore provide very strong support for it. They are not compatible with any non-statistical model, including the delay strategy.

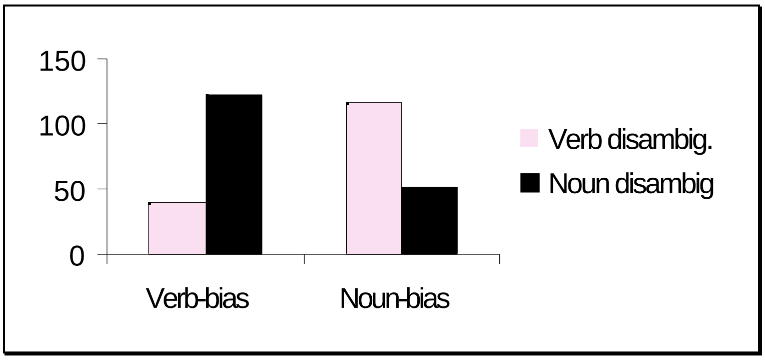

Figure 3: Experiment 1 results for the first word of the disambiguating region (d_1)

Experiment 2: The Modular Lexical Category Hypothesis

The SLCH posits that initial lexical category decisions are made on the basis of frequency-based preferences. It does not require that no other constraints influence these decisions; nor does it entail a modular architecture. If we presuppose the modular architecture argued for earlier, the SLCH still does not indicate the existence of a Statistical Lexical Category Module; lexical category decisions could be made by a statistical parser (e.g. Jurafsky, 1996).

The MLCH addresses the question of modularity, stating that a pre-syntactic module is responsible for lexical category decisions. Initial lexical category decisions should not be affected by syntax and ‘higher’ levels of processing. The MLCH therefore makes interesting predictions where syntactic constraints and frequency-based lexical category bias are in opposition. For example, in a syntactically unambiguous sentence containing words that display lexical category ambiguity, the MLCH asserts that reanalysis effects will be observed if the initial decision of the lexical category module is syntactically illicit.

Corley’s (1998) experiment 2 examined materials of this nature, again concerning the noun–verb ambiguity. Examples are given below.

Experiment 2: Materials

(a) The woman said that the German *makes* are cheaper than the rest.
(b) The woman said that the German *make* is cheaper than the rest.
(c) The foreman knows that the warehouse *prices* are cheaper than the others.
(d) The foreman knows that the warehouse *price* is cheaper than the others.

Example (a) is identical to (b) in experiment 1 – the ambiguous word is verb-biased, but the disambiguation favours a noun reading. In contrast, (b) is unambiguous; the plural verb “make” is not syntactically licit following the singular noun “German”; “make” must therefore be a noun. If (all) syntactic constraints affect initial lexical category decisions, we would expect this decision to favour the noun reading despite the verb bias of the lexically ambiguous word.

Examples (c) and (d) both contain noun-biased ambiguous words. In (c) the disambiguating material favours a noun reading. (d) is again unambiguous – the plural verb “price” cannot follow the singular noun “warehouse”; “price” must therefore be a noun.

Experiment 1 determined initial lexical category decisions in the absence of syntactic constraints. The MLCH asserts that these preferences should not be changed by the presence of syntactic constraints. We therefore predict that in (a) and (b), a verb reading will be initially preferred whereas in (c) and (d) a noun reading will be preferred.

As all materials are (eventually) only compatible with the noun reading, we would expect processing difficulty, realised as a reading time increase, to be evidenced downstream from the ambiguous word in the verb-bias conditions. (a) is identical to the materials in experiment 1, and we would therefore predict reading time increases at the disambiguating region. In (b), reading time increases may appear on the ambiguous word itself. This is because there is sufficient evidence for higher levels of processing to demand lexical category reanalysis as soon as the ambiguous word is read. We would therefore predict that reanalysis, reflected by reading time increases, would start on the ambiguous word in the verb-biased unambiguous condition. We do not predict any reading time increases on the noun-biased conditions.

In contrast, any model in which syntax affects initial lexical category decisions, including interactive constraint-based models, must predict no reanalysis effects on the unambiguous conditions. The delay strategy predicts decreased reading times for the ambiguous word and increased reading time for the disambiguating region in the ambiguous conditions compared to the unambiguous ones.

Results and Discussion

The method used was the same as that for experiment 1. Average length-adjusted reading times obtained for experiment 2 are shown in Figure 4. On the first word of the disambiguating region, a highly significant main effect of bias was observed (F_1 = 20.1, p < .001; F_2 = 18.68, p < .001), but there was no main effect of ambiguity (F_1 = 0.26, p > .6; F_2 = 0.16, p > .6). This suggests that initial decisions are based on word bias and ignore syntactic constraints. By the second word of the disambiguating region, recovery in the verb-bias unambiguous condition appears complete. In contrast, recovery in the verb-bias ambiguous condition lags into this word. This suggests that syntax does have a rapid effect on lexical category decisions, but only after the initial decision has been made.

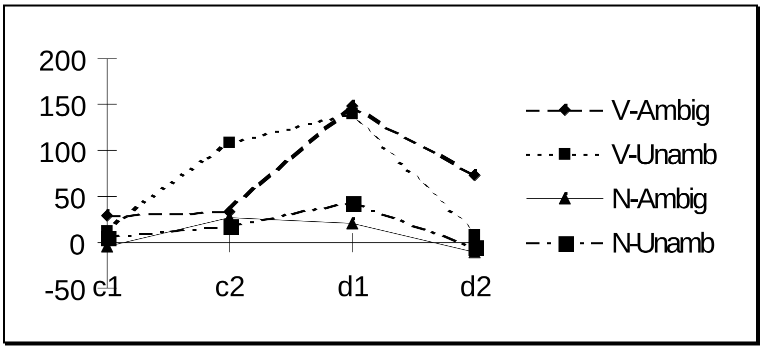

Figure 4: Experiment 2 length-adjusted reading times

A planned comparison of means for the ambiguous word (see Figure 5) reveals a significant difference in reading times for the two verb-biased conditions (F_1 = 5.24, $p < .03$; F_2 = 7.16, $p < .015$) but not for the noun-biased conditions (F_1 = 0.12, $p > .7$; F_2 = 0.10, $p > .75$). In the unambiguous verb-bias condition, subjects experience difficulty reading the lexically-ambiguous word. This is predicted by the MLCH; syntactic constraints result in a rapid reanalysis effect but do not affect the initial decision.[7]

[7] Thanks to one of the reviewers for pointing out that, as all temporarily ambiguous sentences are disambiguated towards the noun reading, it might be argued that these results arise from an experimental-internal bias. However, we believe this suggestion is implausible. If the subjects developed a strategy of preferring the noun reading when encountering ambiguous items, we would not expect to observe a significant effect at the start of the disambiguating condition in both verb bias conditions. Furthermore, this strategy would not explain the crucial observation of a reanalysis effect in the verb-bias unambiguous condition on the ambiguous word. Development of such a strategy is also unlikely due to the large number of filler items (80) compared to experimental items (24) presented to each subject.

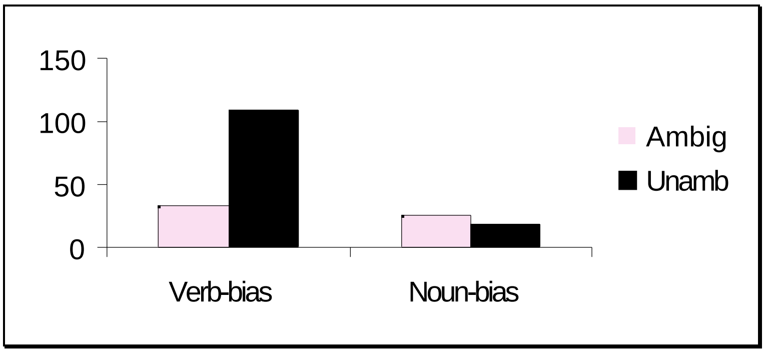

Figure 5: Experiment 2 results for the ambiguous word (c_2)

These results are predicted by and strongly support the MLCH (and the SLCH). They are not compatible with the delay strategy, which predicts a main effect of ambiguity on both the ambiguous word and the disambiguating region. These results are also incompatible with any model in which syntax determines initial lexical category decisions, including some possible interactive constraint-based models. Finally, the observed effect on the ambiguous word is not explained by a model in which reading times are sensitive to syntactic complexity (Just & Carpenter, 1980; MacDonald, 1993). Such a model might (incorrectly) predict an increased reading time on the ambiguous word in the verb-bias ambiguous condition (as a verb phrase must be constructed). However, the observed increase on the verb-bias unambiguous condition compared to the verb-bias ambiguous condition cannot arise directly from syntactic complexity.

Number agreement might have an effect even in a pre-syntactic module if contextual information affects initial decisions (as in the SLCM). This is because the lexical category sequence singular noun followed by plural verb has very low frequency. If we accept that contextual information is used, then this experiment provides evidence that it is in some ways coarse-grained. In particular, the lexical category tags used by the SLCM cannot include number.

Summary of Results

Experiment 1 strongly supported the SLCH and the results were not compatible with a model in which frequency-based bias does not affect initial lexical category decisions. Non-statistical models of lexical category disambiguation are therefore disconfirmed.

One such model is the delay model, proposed by Frazier and Rayner (1987). MacDonald's (1993) study demonstrated that Frazier and Rayner's results might have arisen from an artefact in their experiment. Experiment 1 also produced results that are incompatible with the delay model. In contrast, constraint-based models tend to be frequency-based and are therefore compatible with the results reported in experiment 1.

Experiment 2 provided clear evidence that lexical category decisions are made without regard to syntactic constraints – they are therefore pre-syntactic. This result supports the MLCH. The experiment also provided initial evidence that any contextual information used alongside lexical frequency bias (such as the category bigrams of the SLCM) in determining initial lexical category decisions must be coarse-grained.

In supporting the MLCH and SLCH, the experiments reported here also provide direct support for the more general Modular Statistical Hypothesis proposed at the beginning of this paper. In particular, the results of experiment 2 do not appear compatible with interactive models in which syntactic constraints may non-modularly resolve lexical category ambiguities.

Summary and Conclusions

We have argued that while statistical mechanisms are commonly taken to be the province of connectionist, constraint-based models of sentence processing, they are also highly consistent with a modular perspective. Rather than being unavoidable side effects of the computational machinery, we argue that statistical mechanisms will be rationally exploited by modular architectures precisely because they provide near optimal heuristic decisions in the absence of full knowledge. Indeed this is a central motivation for the use of statistical language models in computational linguistics. We have dubbed this general proposal the Modular Statistical Hypothesis (MSH).

To investigate the MSH, we proposed a specific theory of human sentence processing, in which lexical category ambiguities are resolved by a post-lexical access/pre-syntactic module. In particular we have argued for the Statistical Lexical Category Module, which adopts the standard tagger algorithm and exploits word-category unigrams and category bigrams to incrementally estimate the probability of the most likely category sequence for a given sentence. We have reviewed the operation of the SLCM, and how it accounts for relevant existing experimental findings.

We then reviewed the results of two new experiments from Corley (1998) designed to directly test both our modular and statistical claims concerning lexical category disambiguation. In both experiments, the predictions of the SLCM were confirmed, thereby supporting both our specific account of category disambiguation and the MSH more generally. The results also have implications for other theories of human sentence processing. While it is true that a constraint-based, interactive framework can be made to account for these findings, it does not *predict* them. That is, such a framework could equally have been made to account for the opposite findings, while such results would have disconfirmed our more predictive (and therefore, we argue, methodologically preferable) modular theory. Regardless, our findings do narrow the space of possible models, suggesting in particular the systematic priority of ‘bottom-up’ information (e.g. lexical frequency) over ‘top-down’ (e.g. syntactic and semantic) constraints. Again this follows directly from a modular account, and requires stipulation within a constraint-based framework (though it may follow from particular computational realisations of a constraint-based model).

The findings of experiment 2 also present a challenge for linguistic and psycholinguistic theories which deny the lexical-syntactic divide. These include syntactic theories such as Head-Driven Phrase Structure Grammar, Lexicalised Tree Adjoining Grammar, and Categorial Grammars, to the extent that they claim to be psychologically real (but see Kim and Trueswell (this volume) for a contrary view). Our findings suggest that category decisions are resolved prior to decisions concerning syntactic structures, and also suggest that the categories themselves are relatively coarse-grained, e.g. not including number features.

Finally, we suggest that there is undeniable evidence for the central role of statistical information in human sentence processing. This is a result which needs to be incorporated into the range of existing ‘symbolically-based’ models. However, such statistical mechanisms should not automatically be taken as evidence against rational and modular theories. On the contrary, statistics may be a module’s best friend.

References:

Anderson, J.R. (1991). Is human cognition adaptive? *Behavioural and Brain Sciences*, **14**, 471-517.

Brants, T. (1999). *Tagging and Parsing with Cascaded Markov Models*, Unpublished PhD dissertation, Saarland University, Germany.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics*, **21**, 543-566.

Charniak, E. (1993). *Statistical Language Learning*. MIT Press, Cambridge, MA.

Charniak, E. (1997). Statistical Techniques for Natural Language Parsing. *AI Magazine*, **18**(4), 33-44.

Charniak, E., Carroll, G., Adcock, J., Cassandra, A., Gotoh, Y., Katz, J., Littman, M. & McCann, J. (1996) Taggers for parsers, *Artificial Intelligence* **85(**1-2).

Chater, N., Crocker, M.W., & Pickering, M. (1999). The Rational Analysis of Inquiry: The Case for Parsing. In: Chater & Oaksford (eds), *Rational Analysis of Cognition*, Oxford University Press, Oxford, pages 441-468.

Corley, S. (1998). *A Statistical Model of Human Lexical Category Disambiguation*, Unpublished PhD dissertation, University of Edinburgh.

Corley, S. & Crocker, M.W. (2000). The Modular Statistical Hypothesis: Exploring Lexical Category Ambiguity. In: Crocker, Pickering & Clifton (eds*)*, *Architectures and Mechanisms for Language Processing*, pp. 135-160. CUP, England.

Crocker, M.W. (1999). Mechanisms for Sentence Processing. In: Garrod and Pickering (eds), *Language Processing*, Psychology Press.

DeRose, S.J. (1988). Grammatical Category Disambiguation by Statistical Optimization. *Computational Linguistics*, **14**, 311-39.

Ferreira, F. & Clifton Jr., C. (1986). The Independence of Syntactic Processing. *Journal of Memory and Language*, **25**, 348-368.

Fodor, Jerry A. (1983). *The modularity of mind*, MIT Press, Cambridge, MA.

Frazier, L. (1978). *On Comprehending Sentences: Syntactic Parsing Strategies*. PhD Thesis, University of Connecticut.

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

Juliano, C. & Tanenhaus, M.K. (1993). Contingent frequency effects in syntactic ambiguity resolution. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, pp. 593-598. Lawrence Erblaum Associates.

Jurafsky, D.A (1996). Probabilistic Model of Lexical and Syntactic Access and Disambiguation, *Cognitive Science*, **20**, 137-194.

Just, M. & Carpenter, P. (1980). A Theory of Reading: From Eye Fixations to Comprehansion. *Psychological Review*, **87**, 329-354.

Just, M., Carpenter, P. & Woolley, J. (1982). Paradigms and Processes in Reading Comprehension. *Journal of Experimental Psychology: General*, **111**, 228-238.

Kawamoto, A.H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity. *Journal of Memory and Language*, **32**, 474-516.

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

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

MacDonald, M.C., Pearlmutter, N.J., & Seidenberg, M.S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, **10**(4), 676-703.

Mitchell, D.C. & Brysbaert, M. (to appear). Challenges to recent theories of cross-linguistic variation in parsing: Evidence from Dutch. In D. Hillert (ed*.), Sentence Processing: A Cross-linguistic Perspective.* Academic Press.

Mitchell, D.C., Cuetos, F., Corley, M.M.B., & Brysbaert, M. (1995). Exposure-based models of human parsing: Evidence for the use of coarse-grained (non-lexical) statistical records. *Journal of Psycholinguistic Research*, **24**, 469-488.

Norris, D. (1990). Connectionism: A Case for Modularity. In D.A. Balota, G.B. Flores d'Arcais & (eds), *Comprehension Processes in Reading*. Lawrence

Erlbaum Associates, Hillsdale, New Jersey, 331-343.

Pickering, M., Traxler, M. & Crocker, M. (2000). Ambiguity Resolution in Sentence Processing: Evidence against Likelihood. *Journal of Memory and Language*, **43**(3):447-475.

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

Ratnarparkhi, A. (1996). A Maximum Entropy Model for Part-Of-Speech Tagging. In *Proceedings of the Empirical Methods in Natural Language Processing Conference,* University of Pennsylvania.

Seidenberg, M., Tanenhaus, M., Leiman, J. & Bienkowski, M. (1982). Automatic Access of the Meanings of Ambiguous Words in Context: Some Limitations on Knowledge-Based Processing. *Cognitive Psychology*, **14**, 489-537.

Seidenberg. M. (1997). Language Acquisition and Use: Learning and Applying Probabilistic Constraints. *Science*, **275**.

Spivey-Knowlton, M. & Eberhard, K. (1996). The future of modularity. In G. W. Cottrell (ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, pp. 39-40. Lawrence Erblaum Associates.

Swinney, D.A. (1979). Lexical Access during Sentence Comprehension: (Re)Construction of Context Effects*.* *Journal of Verbal Learning and Verbal Behaviour*, **18**, 645-659.

Tanenhaus, M.K. & Donnenworth-Nolan, S. (1984). *Syntactic Context and Lexical Access*. Quarterly Journal of Experimental Psychology, **36A**, 649-661.

Tanenhaus, M.K., Leiman, J.M. & Seidenberg, M.S. (1979). Evidence for Multiple Stages in the Processing of Ambiguous Words in Syntactic Contexts. *Journal of Verbal Learning and Verbal Behaviour,* **18**, 427-440.

Tanenhaus, M.K., Spivey-Knowlton, M.J., & Hanna, J.E. (in press). Modelling Discourse Context Effects: A Multiple Constraints Approach. In Crocker, Pickering & Clifton (eds.) *Architectures and Mechanisms for Language Processing*, Cambridge University Press, Cambridge, England.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., and Sedivy, J.C. (1995). Integration of Visual and Linguistic Information in Spoken Language Comprehension, *Science*, **268**, 1632-1634.

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

Weischedel, R. Meteer, M., Schwarz, R., Ramshaw, L. & Palmucci, J. (1993). Coping with ambiguity and unknown words through probabilistic models. Computational Linguistics, **19**, 359-382.