Evidence for a Tagging Model of Human Lexical Category Disambiguation.

Steffan Corley and **Matthew W. Crocker**.
Centre for Cognitive Science,
University of Edinburgh,
Edinburgh EH8 9LW, UK.
Steffan.Corley@ed.ac.uk

Abstract.

We investigate the explanatory power of very simple statistical mechanisms within a modular model of the Human Sentence Processing Mechanism. In particular, we borrow the idea of a 'part-of-speech tagger' from the field of Natural Language Processing, and use this to explain a number of existing experimental results in the area of lexical category disambiguation. Not only can each be explained without the need to posit extra mechanisms or constraints, but the exercise also suggests a novel account for some established data.

Introduction.

Much recent research into human sentence processing has concentrated on the use of experience-based statistical knowledge in making initial decisions (Mitchell & Cuetos, 1991; MacDonald, Pearlmutter & Seidenberg, 1994; Tanenhaus & Trueswell, 1994; Corley, Mitchell, Brysbaert, Cuetos & Corley, 1995). We formalise this tendency by introducing the "Statistical Hypothesis":

> *Statistical mechanisms play a central role in the Human Sentence Processor.*

It is worth establishing this as a very broad hypothesis, which avoids making a number of claims that are the subject of debate in the current literature while encompassing a range of models that do. In particular, it does not claim that *all*, or indeed any, initial decisions are made on the basis of statistics, nor that statistics play a role at any particular level of processing. Issues of granularity are also unspecified by the hypothesis.

A number of statistical models have already been proposed and therefore fall within the Statistical Hypothesis. These include Mitchell and Cuetos' (1991) "Tuning", and constraint-based models from MacDonald *et al.* (1994) and Trueswell and Tanenhaus (1994). These models share a common assumption – that all that statistics offer us is an improved heuristic for making decisions in the face of ambiguity and (in the case of the constraint-based models) for discarding parallel analysis. That is, statistics *supplement* a viable, non-statistical architecture.

We argue that statistical mechanisms are most suitable for simple low-level processes that do not form part of traditional models of the human sentence processing mechanism (*HSPM*). We also suggest that a statistical model should differ from traditional approaches in architecture, as well as in decision procedures. Evidence for these views comes from the A.I. literature, where statistical mechanisms have been used in traditional tasks such as parsing (Magerman & Marcus, 1991), but have been most successful in more constrained, low-level tasks such as lexical category disambiguation and noun phrase boundary detection (Church, 1988).

In this paper, we propose a distinct statistical process performing lexical category disambiguation within a modular *HSPM*. We briefly touch on the mathematics of such a model, and then go on to test the predictions of our model against some established experimental data. The results not only demonstrate the power of such a simple statistical technique, but also cast new light on the experimental data. We conclude with a few wider considerations and lessons learnt.

Lexical Category Disambiguation.

If we are to design a lexical category disambiguation module, the first question must be what statistics should it use?

It seems likely that the *HSPM* could gather statistics relating individual words to their lexical category (e.g. how often "post" appears as a noun or verb). Beyond that, experimental evidence supports the use of limited contextual information (Juliano & Tanenhaus, 1993). The simplest, most coarse-grained, contextual statistics are lexical category co-occurrence statistics (e.g. how often a noun follows a preposition). Given no compelling evidence for finer-grained information, we limit ourselves to these two statistics.

It happens that a simple process using exactly these statistics has been well explored in the A.I. literature. It is called a part-of-speech "tagger". Its job is to determine a preferred set of part-of-speech "tags" for a given set of words. Equation 1 is used to assign a probability to each possible tag set (or "tag path")[1].

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

This equation can be applied incrementally. That is, after 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. However, this

[1] w_i is the word at position i in the sentence, t_i is a possible tag for that word.

decision may be altered by the tagger when later words are encountered (see section 4 for further discussion).

Figure 1 depicts tagging the two words "some men". Supposing we already know the probability of "some" occurring as a determiner, noun or adjective. We can then work out the probability of each tag path in which "men" is a noun by multiplying the relevant probability for "some" by the word-tag ($P(w_i \mid t_i)$, where w_i is "men") and bigram ($P(t_i \mid t_{i-1})$) probabilities. Similar calculations can be performed for tag paths in which the tag for "men" has some other value – for instance adjective or verb. The most likely tag for "men" is the one that occurs in the most probable tag path.

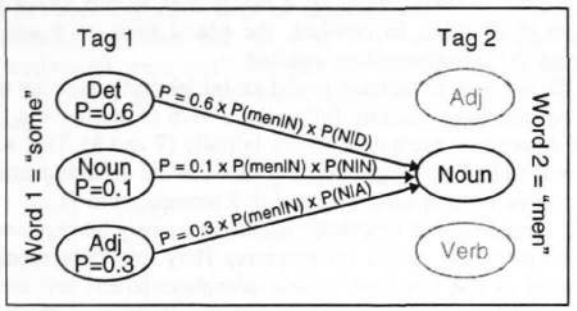

Figure 1: Tagging the words "some men".[2]

As stated, this algorithm is expensive in a real world situation, as it involves remembering all possible tag paths through an arbitrarily long input sentence. However, a large number of clear losers can rapidly be discarded. With this simplification, the algorithm is linear (Viterbi, 1967).

Taggers, in general, are extremely accurate (often 95% – see Charniak, 1993). However, they have distinctive breakdown and repair patterns, which we will argue are similar to those shown by humans.

We propose that a lexical category disambiguation module, functionally equivalent to a tagger, occurs as a distinct process with human lexical access – prior to a modular[3] syntactic component. Its purpose is to make 'quick and dirty' decisions based on limited statistical information. These decisions may then be altered at "higher" levels of processing. In sections 3 and 4 we present existing experimental evidence which supports this claim.

Tagging and Initial Decisions.

Noun-Verb Ambiguities.

Following Frazier and Rayner (1987), MacDonald (1993) investigated processing of sentences where a word is ambiguous between noun and verb readings, following another noun.

1. The union told reporters that the warehouse fires many workers each spring...
2. The union told reporters that the *corporation fires* many workers each spring...

In 1, the two words form a plausible noun compound ("warehouse fires"). As all her disambiguations favour a verb reading for the ambiguous word[4], MacDonald calls this an "unsupportive bias". In contrast, the potential noun compound in 2 ("corporation fires") is implausible, and so there is a supportive bias.

The experiment also included two unambiguous conditions in which the noun compound was ruled out on syntactic grounds. 3 and 4 are sample materials for the unsupportive and supportive bias versions of this condition.

3. The union told reporters that the *warehouses fire* many workers each spring...
4. The union told reporters that the *corporations fire* many workers each spring...

MacDonald found that bias did appear to influence the initial decision of the *HSPM*. There was a significant increase in reading time for the disambiguating region in unsupportive bias conditions (compared to the analogous unambiguous condition), but almost no difference after a supportive bias. That is, the evidence suggests that 1 is the only case in which the *HSPM* makes an initial decision in favour of the noun compound reading.

MacDonald goes on to correlate "supportive bias" with some fine-grained statistical measures, including word-word co-occurrence frequencies and the head-modifier preference of the first noun ("corporation" or "warehouse" above). The tagger model does not include such fine-grained statistics, and it is therefore clear that our predictions will be substantially different. It so happens that the frequency with which a noun follows another noun is very close to that with which a verb follows a noun in all corpora we have examined. That is, there is unlikely to be any strong contextual bias. The behaviour of the tagger will therefore depend largely on the category bias of the individual ambiguous words used.

Figure 2 represents the noun-verb bias of each of the ambiguous words in MacDonald's experiment[5]. The data was obtained from a corpus count and equation 2 was used to calculate each word's "bias" from the count.

$$bias = \log\left(\frac{noun\ count}{verb\ count}\right) \qquad (2)$$

[2] The numbers in this figure are invented for the sake of exposition and are not intended to represent real probabilities.

[3] By 'modular', we mean that processes and knowledge are somehow distinct, but we leave for later research the issue of the nature and degree of their communication.

[4] We refer to MacDonald's second experiment. The first is largely concerned with refuting Frazier and Rayner's (1987) experimental materials, and is therefore of little relevance here.

[5] The mean bias is 3.69 and the standard deviation is 1.97. This data was obtained from the British National Corpus (BNC), which contains over 100 million words of British English. The data only includes the plural ("-s") form of the word (the alternative spelling "programmes" was included in the count for "programs"). However, even if we include both base and plural forms, the results are similar (mean 2.66, standard deviation 1.79). Searching smaller corpora of American English (SUSANNE and part of the TreeBank Corpus) also gives very similar results.

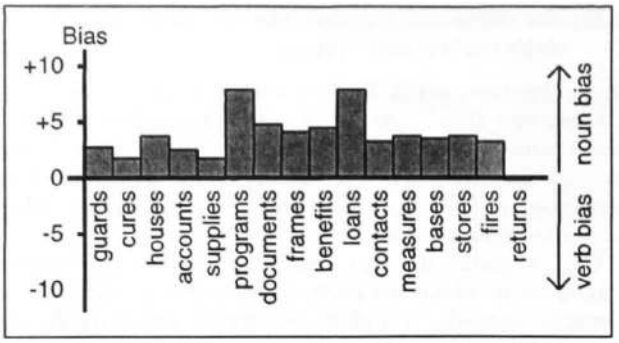

Figure 2: Bias of ambiguous words in MacDonald's (1993) experiment.

It should be clear that the vast majority of MacDonald's experimental materials were strongly biased towards a nominal reading. The initial decision of the tagger depends on two probabilities – $P(t_i \mid t_{i-1})$, the contextual bias (roughly equal for the noun and verb readings), and $P(w_i \mid t_i)$, the word bias, represented in figure 2. The tagger model therefore predicts an initial decision in favour of the noun reading for all of MacDonald's experimental items (with the possible exception of those based on the word "returns"). This decision will be rapidly revised following syntactic analysis in 3 and 4, and may be revised following pragmatic analysis in 2. We would expect these revisions to cause processing delays as the word is read.

This partially agrees with MacDonald's reported findings. We predict a similar pattern of results in the disambiguating region. However, we also predict processing delays on the ambiguous word. Fortunately, MacDonald reported the reading times for the ambiguous word (shown in figure 3).

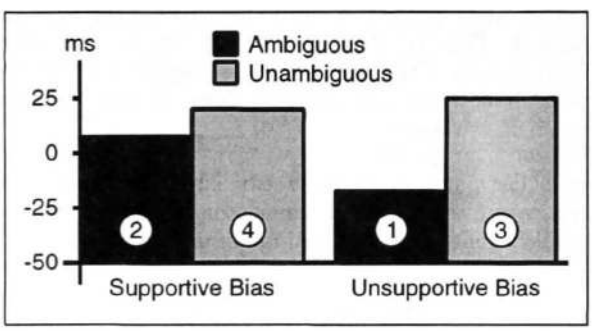

Figure 3: Length-Adjusted Reading Times for the Ambiguous Word from MacDonald (1993).

In conditions 2, 3 and 4, there is a significant processing delay on the ambiguous word compared to condition 1. MacDonald attributes this to the overhead of building the more complex verb phrase structure and calls it a "reverse ambiguity effect" (MacDonald, 1994). However, this processing delay is directly predicted by our model, without introducing complexity measures and using a simpler statistical model.

"That" Ambiguity.

Juliano and Tanenhaus (1993, experiment 1) investigated the initial decisions of the *HSPM* when faced with the ambiguous word "that" in two contexts – sentence initially and following a verb. They forced disambiguation by manipulating the number of the following noun.

5. The lawyer insisted *that experienced diplomat* would be very helpful.
6. The lawyer insisted *that experienced diplomats* would be very helpful.
7. *That experienced diplomat* would be very helpful to the lawyer.
8. *That experienced diplomats* would be very helpful made the lawyer confident.

In 5 and 7, "that" must be a determiner as the following noun is singular. In contrast, the plural noun in 6 and 8 forces the complementiser reading.

Juliano and Tanenhaus found an initial preference for the complementiser reading following a verb (5 and 6), but for the determiner reading sentence initially (7 and 8). This was demonstrated by greater reading times in the disambiguating region in 5 (compared to 6) and in 8 (compared to 7).

It would appear that these results can easily be explained in terms of the tagger architecture. They rely on a regular pattern in the language – that complementisers are more frequent following verbs than sentence initially – which is captured by the lexical category co-occurrence statistics employed by the tagger. Table 1 lists the relevant statistics.

	Prob. of Comp.	**Prob. of Det.**
Sentence Initial	0.0003	0.0652
Following Verb	0.0234	0.0296

Table 1: Estimated probabilities of complementiser and determiner in two contexts (from BNC).

In both cases the preference is in favour of the determiner reading. However, the tagger also makes use of word-tag statistics, and these are biased the other way ($P(that \mid comp)$ = 1.0, $P(that \mid det)$ = 0.171). This bias is strong enough to overcome the comparatively weak contextual bias following a verb, but not the far stronger sentence initial bias. So the predictions for the tagger model match Juliano and Tanenhaus's data – an initial decision in favour of a determiner at the beginning of a sentence, but a complementiser reading is preferred immediately following a verb.

The Tagger's Role in Reanalysis.

The results reported so far demonstrate that the initial decisions made by a tagging model match some established experimental results. However, in the tagging literature there are also good and efficient mechanisms for reassigning tags downstream; that is, reanalysis within the tagger. This section explores whether the tagger's limited reanalysis capabilities may be sufficient to explain some experimental data.

How Tagger Reanalysis Works.

We have already discussed how the tagger assigns a probability to a tag path. Returning to figure 1, suppose that $P(N \mid A)$ – the probability of a noun following an adjective – was more than twice $P(N \mid D)$ – the probability of a noun following a determiner. The tag path in which "men" is a noun and "some" is an adjective would then have a higher contingent probability than that in which "some" is a determiner. So the tagger would have altered its previous decision about the most likely tag for "some".

Such reanalysis must involve a change in the previous tag (in a bigram model). However, it is possible (though extremely unlikely) that the previous two or more tags will be revised. Figure 4 depicts tagging the two contrasting sentences "without her he was lost" and "without her contributions were lost". We plot the probabilities assigned by the tagger to the two most likely tag paths after each word.

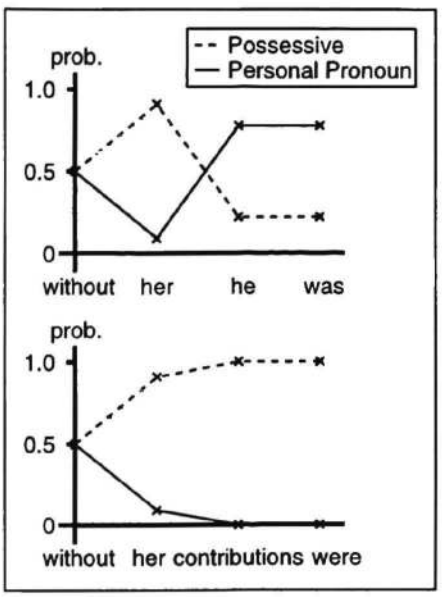

Figure 4: Tagging Two Contrasting Sentences.[6]

The tagger's initial decision when it encounters the word "her" is to favour the possessive reading. However, "he" is unambiguously a personal pronoun and the sequence possessive followed by personal pronoun is extremely unlikely. The tagger's analysis rapidly changes.

In contrast, reanalysis does not occur in the "contributions" case (possessive remains the preferred reading), so we predict a garden path effect on disambiguation. According to Pritchett (1992), this sentence produces a conscious garden path. However, we know of no published experimental evidence to confirm this prediction.

Post-Ambiguity Constraints.

MacDonald (1994) investigated a number of contextual manipulations which can make main verb/reduced relative ambiguities easier to parse. Among these were "post-ambiguity constraints".

9. The sleek greyhound *raced* at the track won four trophies.
10. The sleek greyhound *admired* at the track won four trophies.
11. The sleek greyhound *shown* at the track won four trophies.
12. The sleek greyhound *admired* all day long won four trophies.

MacDonald discovered that sentences such as 9 result in greater reading time for the disambiguating region ("won four trophies") than either 10 or 11 (the unambiguous control). However, the ambiguous region of 10 ("admired at the track") is slower to read than the same region in 9.

MacDonald argues that the difference occurs as "admired" is strongly biased towards a transitive reading. When a transitive verb is not immediately followed by a noun phrase, a strong constraint is violated and an alternative analysis may be sought. In this case, the reduced relative reading becomes the preferred analysis, as the intransitive reading is unlikely. This "post-ambiguity constraint" does not aid in processing 9 as "raced" is more frequently intransitive, and this reading is consistent with a following prepositional phrase.

The constraint MacDonald proposes here is one of lexical category co-occurrence. She also demonstrates that a "poor constraint" – where the constituent following the verb is initially ambiguous between noun phrase and other reading (as in 12) – is less helpful to the reader. These observations appear to match the reanalysis behaviour of a tagger.

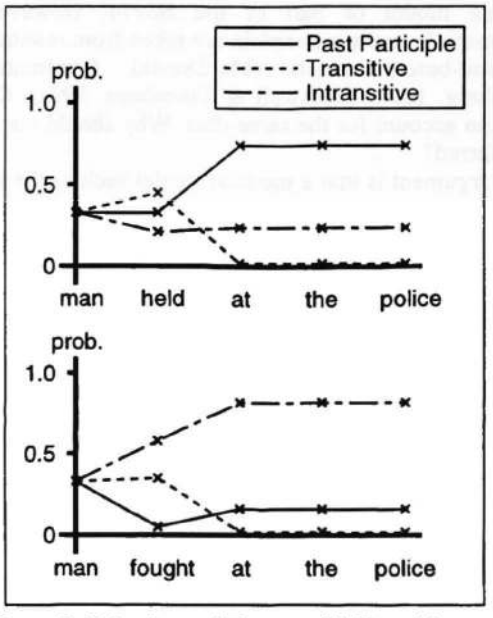

Figure 5: Behaviour of Tagger with Transitive and Intransitive Biased Verbs.

In order to simulate this behaviour, we must train a tagger to

[6] These probabilities, and those in figure 5, have been scaled to add up to 1.

assign transitivity information as part of the lexical category. We achieved this by automatically marking all verbs in the SUSANNE corpus for transitivity. Unfortunately, this marking can only be done for this particular corpus, which is rather small. While we obtained reliable tag co-occurrence statistics, we did not have sufficient lexical statistics to tag the same sentences as MacDonald used. We therefore tagged the two sentences "the man held at the police station fainted" and "the man fought at the police station fainted" – which happened, in our corpus, to be biased towards transitive and intransitive respectively. The results are shown in figure 5.

In the "held" case, the tagger initially prefers the transitive reading but rejects this in favour of the reduced relative reading on encountering the following preposition. In this case, we predict increased reading time in the ambiguous region. In contrast, the intransitive reading is preferred and there is no reanalysis in the case of "fought", so we predict increased reading time in the disambiguating region. This agrees with MacDonald's (1994) results, and so the tagger offers a simpler explanation of her "post-ambiguity constraint".

Conclusions.

The primary conclusion of this work is that a tagger model can account for some psychological data. The inclusion of a tagger within a modular model of the *HSPM* provides, at very low cost, a significant aid in ambiguity resolution. This initial study suggests such a model may be psychologically plausible.

Why not a Constraint-Based Model?

We have argued that our "tagger" account may be a plausible model of part of the *HSPM*. However, the experimental results we explain are taken from research into constraint-based models (MacDonald, Pearlmutter & Seidenberg, 1994; Trueswell & Tanenhaus, 1994). Clearly, these can account for the same data. Why should our model be preferred?

Our argument is that a modular model such as the one we are proposing is "simpler". Advocates of constraint-based approaches have argued that their models are structurally simpler (MacDonald, Pearlmutter & Seidenberg, 1994). However, within a probabilistic framework, *structural* simplicity does not seem to be the correct metric.

- In a constraint-based model, a large number of parameters may effect an initial decision during sentence processing. In our model, initial decisions are mitigated by two simple statistical counts, yet we can still account for the same data.
- The range of information types that effect initial decisions in constraint-based models mean that a huge amount of statistical information must be gathered during language learning. Our model is far more "compact".
- Due to the sparsity of some statistical data, it can be difficult to reliably estimate the parameters required by constraint-based models. The interaction of these parameters also tends to be underspecified, making it difficult to produce concrete predictions. The simpler statistics used by our model mean that it is predictive.
- Constraint-based models allow statistical information to cross levels of representation – for instance, the previous phoneme may be used as a predictor for the next word. We argue that the best predictors tend to occur on the same level of representation. This is built in to our model – again, reducing the number of parameters. Such behaviour may, at best, be emergent from constraint-based models, while it is predicted by ours.

In summary, our model uses simpler statistics, and therefore less parameters, than constraint-based models, and makes no appeal to additional mechanisms, yet still predicts the same data.

Further Conclusions and Lessons Learnt.

We have argued that our model is simpler than a constraint-based model, but can account for the same data. However, the implications are far wider than this.

We are not just using statistics to supplement the decision making process of an existing model. Instead, the use of statistics has informed the architecture of the model. The inclusion of low-level statistical mechanisms could significantly reduce the workload of the structure building component of the *HSPM*. Within a statistical approach, the syntactic module need not be unitary.

We have also learnt two lessons from undertaking this work. The first is that it can be very difficult to intuit the behaviour of a particular statistical system. In order to further the argument, we must build explicit mathematical models as well as argue general principles. The second lesson is that the complexity of behaviour possible with very simple, coarse-grained statistical models can be surprising.

Acknowledgements.

We are thankful to Chris Mellish and the three anonymous reviewers for helpful and supportive comments on earlier drafts of this paper.

The authors also gratefully acknowledge the support of the ESRC (grant R00429334081 to the first author, research fellowship H52427000394 to the second).

References.

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

Church, K.W. (1988). A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the Second Conference on Applied Natural Language Processing* (pp. 136–143). Austin, Texas. ACL.

Corley, M., Mitchell, D.C., Brysbaert, M., Cuetos, F. and Corley, S. (1995). Exploring the Rôle of Statistics in Human Natural Language Processing. In *Proceedings of the 4th International Conference on the Cognitive Science of Natural Language Processing*. Dublin, Ireland.

Frazier, L. and 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. and 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.

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(2), 157–201.

MacDonald, M.C., Pearlmutter, N.J. and Seidenberg, M.S. (1994). Lexical Nature of Syntactic Ambiguity Resolution. *Psychological Review*, 101(4), 676–703.

Magerman, D.M. and Marcus, M.P. (1991). Pearl: A Probabilistic Chart Parser. In *Proceedings: Second International Workshop on Parsing Technologies* (pp. 193–199).

Mitchell, D.C. and Cuetos, F. (1991). The Origins of Parsing Strategies. In C. Smith (Ed.), *Current Issues in Natural Language Processing*. University of Austin, Texas.

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

Trueswell, J.C. and Tanenhaus, M.K. (1994). Toward a Lexicalist Framework for Constraint-Based Syntactic Ambiguity Resolution. In C. Clifton, Jr., L. Frazier and K. Rayner (Eds.), *Perspectives on Sentence Processing* (pp. 155–179). Lawrence Erblaum Associates.

Viterbi, A.J. (1967). Error Bounds for Convolution Codes and an Asymptotically Optimal Decoding Algorithm. *IEEE Transactions on Information Theory*, 13, 260–269.