

The Influence of Discourse on Syntax

A Psycholinguistic Model of Sentence Processing

Amit Dubey

Abstract

Probabilistic models of sentence comprehension are increasingly relevant to questions concerning human language processing. However, such models are often limited to syntactic factors. This paper introduces a novel sentence processing model that consists of a parser augmented with a probabilistic logic-based model of coreference resolution, which allows us to simulate how context interacts with syntax in a reading task. Our simulations show that a Weakly Interactive cognitive architecture can explain data which had been provided as evidence for the Strongly Interactive hypothesis.

1 Introduction

Probabilistic grammars have been found to be useful for investigating the architecture of the human sentence processing mechanism (Jurafsky, 1996; Crocker and Brants, 2000; Hale, 2003; Boston et al., 2008; Levy, 2008; Demberg and Keller, 2009). For example, probabilistic models shed light on so-called locality effects: contrast the non-probabilistic hypothesis that dependants which are far away from their head *always* cause processing difficulty for readers due to the cost of storing the intervening material in memory (Gibson, 1998), compared to the probabilistic prediction that there are cases when faraway dependants facilitate processing, because readers have more time to predict the head (Levy, 2008). Using a computational model to address fundamental questions about sentence comprehension motivates the work in this paper.

So far, probabilistic models of sentence processing have been largely limited to syntactic factors. This is unfortunate because many outstanding questions in psycholinguistics concern interactions between different levels of processing. This paper addresses this gap by building a computational model which simulates the influence of discourse on syntax.

Going beyond the confines of syntax alone is a sufficiently important problem that it has attracted

attention from other authors. In the literature on probabilistic modeling, though, the bulk of this work is focused on lexical semantics (e.g. Padó et al., 2006; Narayanan and Jurafsky, 1998) or only considers syntactic decisions in the preceding text (e.g. Dubey et al., 2009; Levy and Jaeger, 2007). This is the first model we know of which introduces a broad-coverage sentence processing model which takes the effect of coreference and discourse into account.

A major question concerning discourse-syntax interactions involves the strength of communication between discourse and syntactic information. The **Weakly Interactive** (Altmann and Steedman, 1988) hypothesis states that a discourse context can reactively prune syntactic choices that have been proposed by the parser, whereas the **Strongly Interactive** hypothesis posits that context can proactively suggest choices to the syntactic processor.

Support for Weak Interaction comes from experiments in which there are temporary ambiguities, or *garden paths*, which cause processing difficulty. The general finding is that supportive contexts can reduce the effect of the garden path. However, Grodner et al. (2005) found that supportive contexts even facilitate the processing of *unambiguous* sentences. As there are no incorrect analyses to prune in unambiguous structures, the authors claimed their results were not consistent with the Weakly Interactive hypothesis, and suggested that their results were best explained by a Strongly Interactive processor.

The model we present here implements the Weakly Interactive hypothesis, but we will show that it can nonetheless successfully simulate the results of Grodner et al. (2005). There are three main parts of the model: a syntactic processor, a coreference resolution system, and a simple pragmatics processor which computes certain limited forms of discourse coherence. Following Hale (2001) and Levy (2008), among others, the syntactic processor uses an incremental probabilistic Earley parser to compute a metric which correlates with increased reading difficulty. The coreference resolution system is implemented

in a *probabilistic logic* known as Markov Logic (Richardson and Domingos, 2006). Finally, the pragmatics processing system contains a small set of probabilistic constraints which convey some intuitive facts about discourse processing. The three components form a pipeline, where each part is probabilistically dependent on the previous one. This allows us to combine all three into a single probability for each reading of an input sentence.

The rest of the paper is structured as follows. In Section 2, we discuss the details two experiments showing support of the Weakly and Strongly Interactive hypotheses: we discuss Grodner et al.'s result on unambiguous syntactic structures and we present a new experiment on involving a garden path which was designed to be similar to the Grodner et al. experiment. Section 3 introduces technical details of model, and Section 4 shows the predictions of the model on the experiments discussed in Section 2. Finally, we discuss the theoretical consequences of these predictions in Section 5.

2 Cognitive Experiments

2.1 Discourse and Ambiguity Resolution

There is a fairly large literature on garden path experiments involving context (Crain and Steedman, 1985; Mitchell et al., 1992, *ibid*). The experiments by Altmann and Steedman (1988) involved PP attachment ambiguity. Other authors (e.g. Spivey and Tanenhaus, 1998) have used reduced relative clause attachment ambiguity. In order to be more consistent with the design of the experiment in Section 2.2, however, we performed our own reading-time experiment which partially replicated previous results.¹

The experimental items all had a target sentence containing a relative clause, and one of two possible context sentences, one of which supports the relative clause reading and the other which does not.

The context sentence was one of:

- (1) a. There were two postmen, one of whom was injured and carried by paramedics, and another who was unhurt.
- b. Although there was a medical emergency at the post office earlier today, regular mail delivery was unaffected.

¹This experiment was previously reported by Dubey et al. (2010).

The target sentences, which were drawn from the experiment of McRae et al. (1998), were either the reduced or unreduced sentences similar to:

- (2) The postman *who was* carried by the paramedics was having trouble breathing.

The reduced version of the sentence is produced by removing the words *who was*. We measured reading times in the underlined region, which is the first point at which there is evidence for the relative clause interpretation. The key evidence is given by the word 'by', but the previous word is included as readers often do not fixate on short function words, but rather process them while overtly fixating on the previous word (Rayner, 1998).

The relative clauses in the target sentence act as *restrictive* relative clauses, selecting one referent from a larger set. The target sentences are therefore more coherent in a context where a restricted set and a contrast set are easily available, than one in which these sets are absent. This makes the context in Example (1-a) supportive of a reduced relative reading, and the context in Example (1-b) unsupportive of a reduced relative clause. Other experiments, for instance Spivey and Tanenhaus (1998), used an unsupportive context where only one postman was mentioned. Our experiments used a neutral context, where no postmen are mentioned, to be more similar to the Grodner et al. experiment, as described below.

Overall, there were 28 items, and 28 participants read these sentences using an EyeLink II eyetracker. Each participant read items one at a time, with fillers between subsequent items so as to obfuscate the nature of the experiment.

Results An ANOVA revealed that all conditions with a supportive context were read faster than one with a neutral context (i.e. a main effect of context), and all conditions with unambiguous syntax were read faster than those with a garden path (i.e. a main effect of ambiguity). Finally, there was a statistically significant interaction between syntax and discourse whereby context decreases reading times much more when a garden path is present compared to an unambiguous structure. In other words, a supportive context helped reduce the effect of a garden path. This is the prediction made by both the Weakly Interactive and Strongly Interactive hypothesis. The pattern of results are shown in Figure 2a in Section 4, where they are directly compared to the model results.

2.2 Discourse and Unambiguous Syntax

As mentioned in the Introduction, Grodner et al. (2005) proposed an experiment with a supportive or unsupportive discourse followed by an unambiguous target sentence. In their experiment, the target sentence was one of the following:

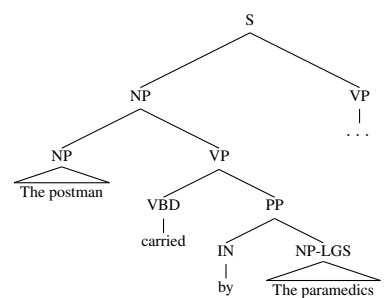
- (3) a. The director that the critics praised at a banquet announced that he was retiring to make room for young talent in the industry.
 b. The director, who the critics praised at a banquet, announced that he was retiring to make room for young talent in the industry.

They also manipulated the context, which was either supportive of the target, or a null context. The two supportive contexts are:

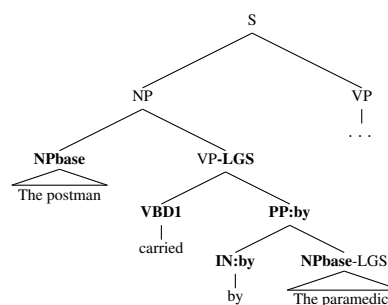
- (4) a. A group of film critics praised a director at a banquet and another director at a film premiere.
 b. A group of film critics praised a director and a producer for lifetime achievement.

The target sentence in (3-a) is a restrictive relative clause, as in the garden path experiments. However, the sentence in (3-b) is a non-restrictive relative clause, which does not assume the presence of a contrast set. Therefore, the context (4-a) is only used with the restrictive relative clause, and the context (4-b), where only one director is mentioned, is used as the context for the non-restrictive relative clause. In the conditions with a null context, the target sentence was not preceded by any contextual sentence.

Results Grodner et al. measured residual reading times, i.e. reading times compared to a baseline in the embedded subject NP ('the critics'). They found that the supportive contexts decreased reading time, and that this effect was stronger for restrictive relatives compared to non-restricted relatives. As there was no garden path, and hence no incorrect structure for the discourse processor to prune, the authors conclude that this must be evidence for the Strongly Interactive hypothesis. Unlike the garden path experiment above, these results do not appear to be consistent with a Weakly Interactive model. We plot their results in Figure 3a in Section 4, where they are



(a) Standard WSJ Tree



(b) Minimally Modified Tree

Figure 1: A schematic representation of the smallest set of grammar transformations which we found were required to accurately parse the experimental items.

directly compared to the model results. Because these results are computed as regressions against a baseline, a reading time of 0ms indicates average difficulty, with negative numbers showing some facilitation has occurred, and positive number indicating reading difficulty.

3 Model

The model comprises three parts: a parser, a coreference resolution system, and a pragmatics subsystem. Let us look at each individually.

3.1 Parser

The parser is an incremental unlexicalized probabilistic Earley parser, which is capable of computing prefix probabilities. A PCFG parser outputs the generative probability $P_{parser}(w, t)$, where w is the text and t is a parse tree. A probabilistic Earley parser can retrieve all possible derivations at word i (Stolcke, 1995), allowing us to compute the probability $P(w_i \dots w_0) = \sum_t P_{parser}(w_i \dots w_0, t)$.

Using the prefix probability, we can compute the word-by-word Surprisal (Hale, 2001), by taking the log ratio of the previous word's prefix probability against this word's prefix probability:

$$\log \left(\frac{P(w_{i-1} \dots w_0)}{P(w_i \dots w_0)} \right) \quad (1)$$

Higher Surprisal scores are interpreted as

being correlated with more reading difficulty, and likewise lower scores with greater reading ease. For most of the remainder of the paper we will simply refer to the prefix probability at word i as $P(w)$. While the prefix probability as presented here is suitable for syntax-based computations, a main technical contribution of our model, detailed in Sections 3.2 and 3.3 below, is that we include non-syntactic probabilities in the computation of Surprisal.

As per Hale’s original suggestion, our parser can compute Surprisal using an exhaustive search, which entails summing over each licensed derivation. This can be done efficiently using the packed representation of an Earley chart. However, as the coreference processor takes trees as input, we must therefore unpack parses before resolving referential ambiguity. Given the ambiguity of our grammar, this is not tractable. Therefore, we only consider an n -best list when computing Surprisal. As other authors have found that a relatively small set of analyses can give meaningful predictions (Brants and Crocker, 2000; Boston et al., 2008), we set $n = 10$.

The parser is trained on the Wall Street Journal (WSJ) section of the Penn treebank. Unfortunately, the standard WSJ grammar is not able to give correct incremental parses to our experimental items. We found we could resolve this problem by using four simple transformations, which are shown in Figure 1: (i) adding valency information to verb POS tags (e.g. VBD1 represents a transitive verb); (ii) we lexicalize ‘by’ prepositions; (iii) VPs containing a logical subject (i.e. the agent), get the -LGS label; (iv) non-recursive NPs are renamed NPbase (the coreference system treats each NPbase as a markable).

3.2 Discourse Processor

The primary function of the discourse processing module is to perform coreference resolution for each mention in an incrementally processed text. Because each mention in a coreference chains is transitive, we cannot use a simple classifier, as they cannot enforce global transitivity constraints. Therefore, this system is implemented in Markov Logic (Richardson and Domingos, 2006), a probabilistic logic, which does allow us to include such constraints.

Markov Logic attempts to combine logic with probabilities by using a Markov random field where logical formulas are features. The

Expression	Meaning
$Coref(x, y)$	x is coreferent with y .
$First(x)$	x is a first mention.
$Order(x, y)$	x occurs before y .
$SameHead(x, y)$	Do x and y share the same syntactic head?
$ExactMatch(x, y)$	x and y are same string.
$SameNumber(x, y)$	x and y match in number.
$SameGender(x, y)$	x and y match in gender.
$SamePerson(x, y)$	x and y match in person.
$Distance(x, y, d)$	The distance between x and y , in sentences.
$Pronoun(x)$	x is a pronoun.
$EntityType(x, e)$	x has entity type e (person, organization, etc.)

Table 1: Predicates used in the Markov Logic Network

Markov Logic Network (MLN) we used for our system uses similar predicates as the MLN-based coreference resolution system of Huang et al. (2009).² Our MLN uses the predicates listed in Table 1. Two of these predicates, $Coref$ and $First$, are the output of the MLN – they provide a labelling of coreference mentions into entity classes. Note that, unlike Huang et al., we assume an ordering on x and y if $Coref(x, y)$ is true: y must occur earlier in the document than x . The remaining predicates in Table 1 are a subset of features used by other coreference resolution systems (cf. Soon et al., 2001). The predicates we use involve matching strings (checking if two mentions share a head word or if they are exactly the same string), matching agreement features (if the gender, number or person of pairs of NPs are the same; especially important for pronouns), the distance between mentions, and if mentions have the same entity type (i.e. do they refer to a person, organization, etc.) As our main focus is not to produce a state-of-the-art coreference system, we do not include predicates which are irrelevant for our simulations even if they have been shown to be effective for coreference resolution. For example, we do not have predicates if two mentions are in an apposition relationship, or if two mentions are synonyms for each other.

Table 2 lists the actual logical formulae which are used as features in the MLN. It should be

²As we are not interested in unsupervised inference, the system of Poon and Domingos (2008) was unsuitable for our needs.

Description	Rule
Transitivity	$Coref(x, z) \wedge Coref(y, z) \wedge Order(x, y) \Rightarrow Coref(x, y)$
	$Coref(x, y) \wedge Coref(y, z) \Rightarrow Coref(x, z)$
	$Coref(x, y) \wedge Coref(x, z) \wedge Order(y, z) \Rightarrow Coref(y, z)$
First Mentions	$Coref(x, y) \Rightarrow \neg First(x)$
	$First(x) \Rightarrow \neg Coref(x, y)$
String Match	$ExactMatch(x, y) \Rightarrow Coref(x, y)$
	$SameHead(x, y) \Rightarrow Coref(x, y)$
Pronoun	$Pronoun(x) \wedge Pronoun(y) \wedge SameGender(x, y) \Rightarrow Coref(x, y)$
	$Pronoun(x) \wedge Pronoun(y) \wedge SameNumber(x, y) \Rightarrow Coref(x, y)$
	$Pronoun(x) \wedge Pronoun(y) \wedge SamePerson(x, y) \Rightarrow Coref(x, y)$
Other	$EntityType(x, e) \wedge EntityType(y, e) \Rightarrow Coref(x, y)$
	$Distance(x, y, +d) \Rightarrow Coref(x, y)$

Table 2: Rules used in the Markov Logic Network

noted that, because we are assuming an order on the arguments of $Coref(x, y)$, we need three formulae to capture transivity relationships. To test that the coreference resolution system was producing meaningful results, we evaluated our system on the test section of the ACE-2 dataset. Using b^3 scoring (Bagga and Baldwin, 1998), which computes the overlap of a proposed set with the gold set, the system achieves an F -score of 65.4%. While our results are not state-of-the-art, they are reasonable considering the brevity of our feature list.

The discourse model is run iteratively at each word. This allows us to find a globally best assignment at each word, which can be reanalyzed at a later point in time. It assumes there is a mention for each base NP outputted by the parser, and for all ordered pairs of mentions x, y , it outputs all the ‘observed’ predicates (i.e. everything but $First$ and $Coref$), and feeds them to the Markov Logic system. At each step, we compute both the maximum a posteriori (MAP) assignment of coreference relationships as well as the probability that each individual coreference assignment is true. Taken together, they allow us to calculate, for a coreference assignment c , $P_{coref}(c|w, t)$ where w is the text input (of the entire document until this point), and t is the parse of each tree in the document up to and including the current incremental parse. As we have previously calculated $P_{parser}(w, t)$, it is then possible to compute the joint probability $P(c, w, t)$ at each word, and therefore the prefix probability $P(w)$ due to syntax and coreference. Overall, we have:

$$\begin{aligned}
 P(w) &= \sum_c \sum_t P(c, w, t) \\
 &= \sum_c \sum_t P_{coref}(c|w, t) P_{parser}(w, t)
 \end{aligned}$$

Note that we only consider *one* possible assignment of NPs to coreference entities per parse, as we only retrieve the probabilities of the MAP solution.

3.3 Pragmatics Processor

The effect of context in the experiments described in Section 2 cannot be fully explained using a coreference resolution system alone. In the case of restrictive relative clauses, the referential ‘mismatch’ in the unsupported conditions is caused by an expectation elicited by a restrictive relative clause which is inconsistent with the previous discourse when there is no salient restricted subset of a larger set. When the larger set is not found in the discourse, the relative clause becomes incoherent given the context, causing reading difficulty. Modeling this coherence constraint is essentially a pragmatics problem, and is under the purview of the pragmatics processor in our system. The pragmatics processor is quite specialised and, although the information it encapsulates is quite intuitive, it nonetheless relies on hand-coded expert knowledge.

The pragmatics processor takes as input an incremental pragmatics configuration p and computes the probability $P_{prag}(p|w, t, c)$. The pragmatics configuration we consider is quite simple. It is a 3-tuple where one element is true if the current noun phrase being processed is a discourse new definite noun phrase, the second

element is true if the current NP is a discourse new indefinite noun phrase, and the final element is true if we encounter an unsupported restrictive relative clause. We simply conjecture that there is little processing cost (and hence a high probability) if the entire vector is false; there is a small processing cost for discourse new indefinites, a slightly larger processing cost for discourse new definites and a large processing cost for an incoherent reduced relative clause.

The first two elements of the 3-tuple depend on the identity of the determiner as recovered by the parser, and on whether the coreference system adduces the predicate *First* for the current NP. As the coreference system wasn't designed to find anaphoric contrast sets, these sets were found using a simple post-processing check. This post-processing approach worked well for our experimental items, but finding such sets is, in general, quite a difficult problem (Modjeska et al., 2003).

The distribution $P_{prag}(p|w, t, c)$ applies a processing penalty for an unsupported restrictive relative clause whenever a restrictive relative clause is in the n best list. Because Surprisal computes a ratio of probabilities, this in effect means we only pay this penalty when an unsupported restrictive relative clause first appears in the n best list (otherwise the effect is cancelled out). The penalty for discourse new entities is applied on the first word (ignoring punctuation) following the end of the NP. This spillover processing effect is simply a matter of modeling convenience: without it, we would have to compute Surprisal probabilities over regions rather than individual words. Thus, the overall prefix probability can be computed as: $P(w) = \sum_{p,c,t} P_{prag}(p|w, t, c)P_{coref}(c|w, t)P_{parser}(w, t)$, which is then substituted in Equation (1) to get a Surprisal prediction for the current word.

4 Evaluation

4.1 Method

When modeling the garden path experiment we presented in Section 2.1, we compute Surprisal values on the word 'by', which is the earliest point at which there is evidence for a relative clause interpretation. For the Grodner et al. experiment, we compute Surprisal values on the relativiser 'who' or 'that'. Again, this is the earliest point at which there is evidence for a relative clause, and depending upon the presence or absence of a pre-

ceding comma, it will be known to be restrictive or nonrestrictive clause. In addition to the overall Surprisal values, we also compute syntactic Surprisal scores, to test if there is any benefit from the discourse and pragmatics subsystems. As we are outputting n best lists for each parse, it is also straightforward to compute other measures which predict reading difficulty, including pruning (Jurafsky, 1996), whereby processing difficulty is predicted when a parse is removed from the n best list, and attention shift (Crocker and Brants, 2000), which predicts parsing difficulty at words where the most highly ranked parse flips from one interpretation to another.

For the garden path experiment, the simulation was run on each of the 28 experimental items in each of the 4 conditions, resulting in a total of 112 runs. For the Grodner et al. experiment, the simulation was run on each of the 20 items in each of the 4 conditions, resulting in a total of 80 runs. For each run, the model was reset, purging all discourse information gained while reading earlier items. As the system is not stochastic, two runs using the exact same items in the same condition will produce the same result. Therefore, we made no attempt to model by-subject variability, but we did perform by-item ANOVAs on the system output.

4.2 Results

Garden Path Experiment The simulated results of our experiment are shown in Figure 2. Comparing the full simulated results in Figure 2b to the experimental results in Figure 2a, we find that the simulation, like the actual experiment, finds both main effects and an interaction: there is a main effect of context whereby a supportive context facilitates reading, a main effect of syntax whereby the garden path slows down reading, and an interaction in that the effect of context is strongest in the garden path condition. All these effects were highly significant at $p < 0.01$. The pattern of results between the full simulation and the experiment differed in two ways. First, the simulated results suggested a much larger reading difficulty due to ambiguity than the experimental results. Also, in the unambiguous case, the model predicted a null cost of an unresponsive context on the word 'by', because the model bears the cost of an unresponsive context earlier in the sentence, and assumes no spillover to the word 'by'. Finally, we note that the syntax-only simulation, shown in Figure 2c, only produced a main effect of ambigu-

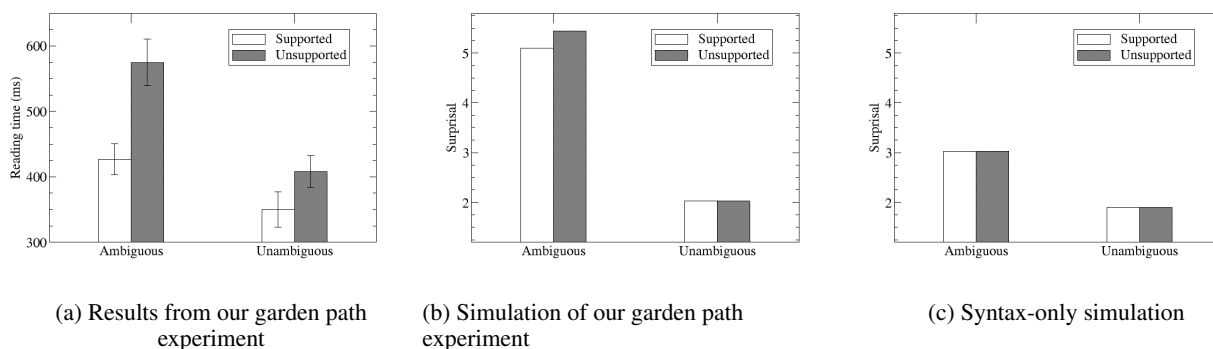


Figure 2: The simulated results predict the same interaction as the garden path experiment, but show a stronger main effect of ambiguity, and no influence of discourse in the unambiguous condition on the word ‘by’.

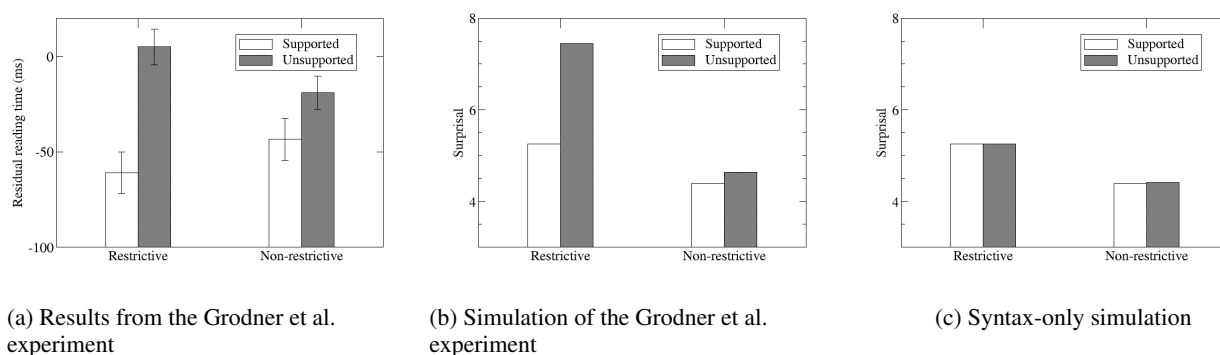


Figure 3: The simulated results predict the outcome of the Grodner et al. experiment.

ity, and was not able to model the effect of context.

Grodner et al. Experiment The simulated results of the Grodner et al. experiment are shown in Figure 3. In this experiment, the pattern of simulated results in Figure 3b showed a much closer resemblance to the experimental results in Figure 3a than the garden path experiment. There is a main effect of context, which is much stronger in the restrictive relative case compared to non-restrictive relatives. As with the garden path experiment, the ANOVA reported that all effects were significant at the $p < 0.01$ level. Again, as we can see from Figure 3c, there was no effect of context in the syntax-only simulation. The numerical trend did show a slight facilitation in the unrestricted supported condition, with a Surprisal of 4.39 compared to 4.41 in the supported case, but this difference was not significant.

4.3 Discussion

We have shown that our incremental sentence processor augmented with discourse processing can

successfully simulate syntax-discourse interaction effects which have been shown in the literature. The difference between a Weakly Interactive and Strongly Interactive model can be thought of computationally in terms of a pipeline architecture versus joint inference. In a weaker sense, even a pipeline architecture where the discourse can influence syntactic probabilities could be claimed to be a Strongly Interactive model. However, as our model uses a pipeline where syntactic probabilities are independent of the discourse, we claim that our model is Weakly Interactive. Unlike Altmann and Steedman, who posited that the discourse processor actually removes parsing hypotheses, we were able to simulate this pruning behaviour by simply re-weighting parses in our coreference and pragmatics modules.

The fact that a Weakly Interactive system can simulate the result of an experiment proposed in support of the Strongly Interactive hypothesis is initially counter-intuitive. However, this naturally falls out from our decision to use a probabilistic

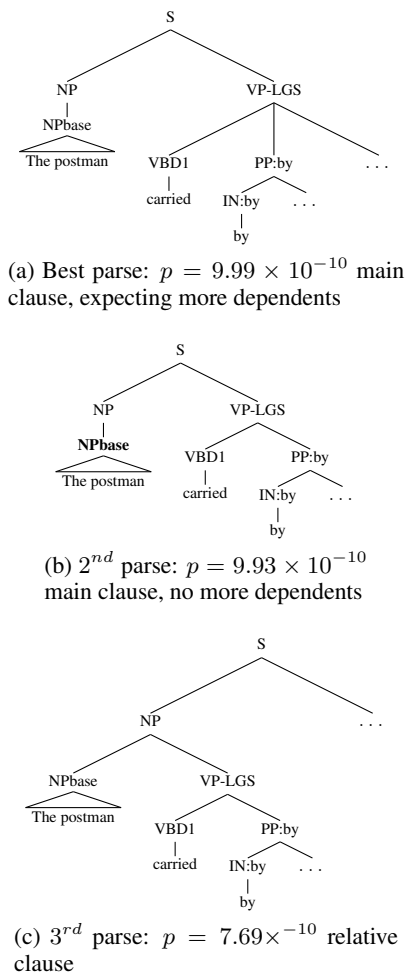


Figure 4: The top three parses on the word ‘by’ in the our first experimental item.

model: a lower probability, even in an unambiguous structure, is associated with increased reading difficulty. As an aside, we note that when using realistic computational grammars, even the structures used in the Grodner et al. experiment are not unambiguous. In the restrictive relative clause condition, even though there was not any competition between a relative and main clause reading, our n best list was at all times filled with analyses. For example, on the word ‘who’ in the restricted relative clause condition, the parser is already predicting both the subject-relative (‘the postman who was bit by the dog’) and object-relative (‘the postman who the dog bit’) readings.

Overall, these results are supportive of the growing importance of probabilistic reasoning as a model of human cognitive behaviour. Therefore, especially with respect to sentence processing, it is necessary to have a proper understanding of how probabilities are linked to real-world behaviours. We note that Surprisal does indeed

show processing difficulty on the word ‘by’ in the garden path experiment. However, Figure 4 (which shows the top three parses on the word ‘by’) indicates that not only are there still main clause interpretations present, but in fact, the top two parses are main clause interpretations. This is also true if we limit ourselves to syntactic probabilities (which are the probabilities listed in Figure 4). This suggests that neither Jurafsky (1996)’s notion of pruning as processing difficulty nor Crocker and Brants (2000) notion of attention shifts would correctly predict higher reading times on a region containing the word ‘by’. In fact, the main clause interpretation remains the highest-ranked interpretation until it is finally pruned at an auxiliary of the main verb of the sentence (‘The postman carried by the paramedics was *having*’).

This result is curious as our experimental items closely match some of those simulated by Crocker and Brants (2000). We conjecture that the difference between our attention shift prediction and theirs is due to differences in the grammar. It is possible that using a more highly tuned grammar would result in attention shift making the correct prediction, but this possibly shows one benefit of using Surprisal as a linking hypothesis. Because Surprisal sums over several derivations, it is not as reliant upon the grammar as the attention shift or pruning linking hypotheses.

5 Conclusions

The main result of this paper is that it is possible to produce a Surprisal-based sentence processing model which can simulate the influence of discourse on syntax in both garden path and unambiguous sentences. Computationally, the inclusion of Markov Logic allowed the discourse module to compute well-formed coreference chains, and opens two avenues of future research. First, it ought to be possible to make the probabilistic logic more naturally incremental, rather than re-running from scratch at each word. Second, we would like to make greater use of the logical elements by applying it to problems where inference is necessary, such as resolving bridging anaphora (Haviland and Clark, 1974).

Our primary cognitive finding that our model, which assumes the Weakly Interactive hypothesis (whereby discourse is influenced by syntax in a reactive manner), is nonetheless able to simulate the experimental results of Grodner et al. (2005), which were claimed by the authors to be in

support of the Strongly Interactive hypothesis. This suggests that the evidence is in favour of the Strongly Interactive hypothesis may be weaker than thought.

Finally, we found that the attention shift (Crocker and Brants, 2000) and pruning (Jurafsky, 1996) linking theories are unable to correctly simulate the results of the garden path experiment. Although our main results above underscore the usefulness of probabilistic modeling, this observation emphasizes the importance of finding a tenable link between probabilities and behaviours.

Acknowledgements

We would like to thank Frank Keller, Patrick Sturt, Alex Lascarides, Mark Steedman, Mirella Lapata and the anonymous reviewers for their insightful comments. We would also like to thank ESRC for their financial supporting on grant RES-062-23-1450.

References

- Gerry Altmann and Mark Steedman. Interaction with context during human sentence processing. *Cognition*, 30:191–238, 1988.
- Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference (LREC 98)*, 1998.
- Marisa Ferrara Boston, John T. Hale, Reinhold Kliegl, and Shravan Vasishth. Surprising parser actions and reading difficulty. In *Proceedings of ACL-08:HLT, Short Papers*, pages 5–8, 2008.
- Thorsten Brants and Matthew Crocker. Probabilistic parsing and psychological plausibility. In *Proceedings of 18th International Conference on Computational Linguistics (COLING-2000)*, pages 111–117, 2000.
- Stephen Crain and Mark Steedman. On not being led down the garden path: the use of context by the psychological syntax processor. In D. Dowty, L. Karttunen, and A. Zwicky, editors, *Natural language parsing: Psychological, computational, and theoretical perspectives*. Cambridge University Press, 1985.
- Matthew Crocker and Thorsten Brants. Wide coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, 29(6): 647–669, 2000.
- Vera Demberg and Frank Keller. A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of the 29th meeting of the Cognitive Science Society (CogSci-09)*, 2009.
- Amit Dubey, Frank Keller, and Patrick Sturt. A probabilistic corpus-based model of parallelism. *Cognition*, 109(2):193–210, 2009.
- Amit Dubey, Patrick Sturt, and Frank Keller. The effect of discourse inferences on syntactic ambiguity resolution. In *Proceedings of the 23rd Annual CUNY Conference on Human Sentence Processing (CUNY 2010)*, page 151, 2010.
- Ted Gibson. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68:1–76, 1998.
- Daniel J. Grodner, Edward A. F. Gibson, and Duane Watson. The influence of contextual contrast on syntactic processing: Evidence for strong-interaction in sentence comprehension. *Cognition*, 95(3):275–296, 2005.
- John T. Hale. A probabilistic earley parser as a psycholinguistic model. In *In Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001.
- John T. Hale. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2):101–123, 2003.
- Susan E. Haviland and Herbert H. Clark. What's new? acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, 13:512–521, 1974.
- Shujian Huang, Yabing Zhang, Junsheng Zhou, and Jiajun Chen. Coreference resolution using markov logic. In *Proceedings of the 2009 Conference on Intelligent Text Processing and Computational Linguistics (CICLing 09)*, 2009.
- D. Jurafsky. A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20:137–194, 1996.
- Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, March 2008.
- Roger Levy and T. Florian Jaeger. Speakers optimize information density through syntactic reduction. In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*, 2007.

- Ken McRae, Michael J. Spivey-Knowlton, and Michael K. Tanenhaus. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38:283–312, 1998.
- Don C. Mitchell, Martin M. B. Corley, and Alan Garnham. Effects of context in human sentence parsing: Evidence against a discourse-based proposal mechanism. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18(1):69–88, 1992.
- Natalia N. Modjeska, Katja Markert, and Malvina Nissim. Using the web in machine learning for other-anaphora resolution. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*, pages 176–183, Sapporo, Japan, 2003.
- Shrini Narayanan and Daniel Jurafsky. Bayesian models of human sentence processing. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society (CogSci 98)*, 1998.
- Ulrike Padó, Matthew Crocker, and Frank Keller. Modelling semantic role plausability in human sentence processing. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci 2006)*, pages 657–662, 2006.
- Hoifung Poon and Pedro Domingos. Joint unsupervised coreference resolution with markov logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-08)*, 2008.
- Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998.
- Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.
- W. M. Soon, H. T. Ng, and D. C. Y. Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.
- M. J. Spivey and M. K. Tanenhaus. Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24(6):1521–1543, 1998.
- Andreas Stolcke. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2):165–201, 1995.