Probabilistic Modeling of Discourse-aware Sentence Processing

Amit Dubey
amit@dubey.ca
Google Inc.

Frank Keller
keller@inf.ed.ac.uk
University of Edinburgh

Patrick Sturt
patrick.sturt@ed.ac.uk
University of Edinburgh

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 restriction is unrealistic in light of experimental results suggesting interactions between syntax and other forms of linguistic information in human sentence processing. To address this limitation, this paper introduces two sentence processing models which augment a syntactic component with information about discourse co-reference. The novel combination of probabilistic syntactic components with co-reference classifiers permits them to more closely mimic human behavior than existing models. The first model uses a deep model of linguistics, based in part on probabilistic logic, allowing it to make qualitative predictions on experimental data; the second model uses shallow processing to make quantitative predictions on a broad-coverage reading-time corpus.

1. Introduction

One of the key issues in contemporary cognitive science is the question of how humans are able to integrate such a wide variety of information in real time, during language processing. By studying how this integration occurs, we can examine the extent to which the various sources of information interact with each other, thus gaining insights into the overall mental architecture of the cognitive system.

A paradigm example of the real-time integration of diverse sources of information is the use of contextual information during syntactic processing, which has provided a valuable testing ground for experimental psycholinguistics over the last three decades (see, for example, Crain & Steedman, 1985; Grodner et al., 2005; D. C. Mitchell et al., 1992; Spivey & Tanenhaus, 1998; Tanenhaus et al., 1995). However, despite this wealth of experimental data, many central questions about the relevant cognitive mechanisms and underlying mental architecture remain unresolved. We believe that one reason for this that an essential piece of the puzzle is missing, namely, there are no large scale computational models of syntax-discourse integration. This is unfortunate, as computational models are ideal tools for helping us to clarify theoretical issues, and for generating empirical predictions.

This paper addresses this gap by describing two computational models, building on results of Dubey (2010) and Dubey et al. (2011), which simulate the influence of discourse on syntax, using probabilistic information. Probabilistic grammars have been found to be useful for investigating the architecture of the

The research reported here was supported by ESRC grant RES-062-23-1450.

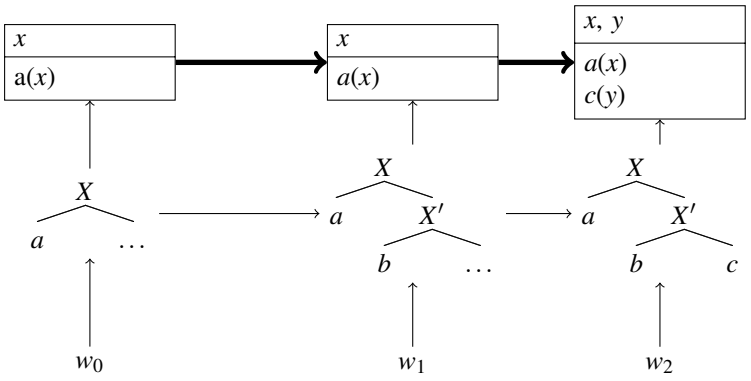


Figure 1. A schematic view of the Paired Model, showing words on the bottom row, syntactic trees on the middle row, and discourse structures on the top row. Arrows show the information used to algorithmically build a new structure. In the Paired Model, discourse structures are created from previous discourse structures (the arrows between structures on the top row) and the incremental syntactic structure (the arrows from trees to discourse structures). *In the Pipeline Model, discourse structures are derived from the syntax alone (i.e. there are no arrows on the top row).*

human sentence processing mechanism (Jurafsky, 1996; Crocker & Brants, 2000; Hale, 2003; Boston et al., 2011; Levy, 2008; Demberg & Keller, 2009), because they allow a natural characterization of graded phenomena, while also being suitable for implementation as large scale models capable of processing realistic text.

So far, however, probabilistic models of sentence processing have been largely limited to syntactic factors. The present work fits within a growing literature which attempts to address this shortcoming. Padó et al. (2006) and Narayanan & Jurafsky (1998) augmented a syntactic parser with the ability to use lexical semantic information. This included information such as a verb's subcategorization preferences, and how to assign theta roles to a verb's arguments. Some authors have attempted to look at the preceding discourse to improve a model's fit to data, but use a limited view of the discourse. Dubey et al. (2009) observed evidence for syntactic priming in corpora, and Levy & Jaeger (2007) predicted the deletion of the relativizer that based upon not any linguistic entities per se, but based upon a probabilistic measure of the information density of the preceding text. J. Mitchell et al. (2010) combined an incremental parser with a vector-space model of semantics. By using a vector-space model, J. Mitchell et al.'s (2010) notion of semantics is restricted to lexical meaning approximated by word co-occurrences. In particular, this approach is not able to account for discourse-level effects which stem from nominal co-reference.

In contrast to earlier work, the models that we propose in this paper integrate contextual information, in the form of discourse entities, into a probabilistic syntactic processing architecture. To our knowledge, they represent the first attempt to account for co-reference or discourse information in a wide-coverage probabilistic model. Both of our models are based upon an intuitive idea: just as probabilistic context-free grammar provides a distribution over syntactic structures, namely trees, our models provide distributions over structures representing discourse co-reference. Practically, this means that both models maintain a list of noun phrases which have been mentioned in the document so far, and also keep track of which noun phrases refer to the same entity.

The two models that we introduce here update the discourse representation incrementally on a sentence-by-sentence basis, but within each sentence, the first model re-generates the entire discourse representation at each word based solely on the syntax, whereas the second one bases it upon the representation of the previous word. A succinct way of describing the difference between the two models is that the first

model can be thought of as a *Pipeline Model* whereby syntax determines the discourse structure. In contrast, the second model has a closer pairing of discourse and syntax. In this model, the discourse representation for a new word is based upon both the syntax and the discourse representation at the previous word. We call the second model the *Paired Model*. The paired model uses a greedy strategy to resolve discourse coreference. This difference is illustrated in Figure 1. In the figure, the particular structures used are irrelevant, but the key point is illustrated by the arrows, which represent information flow.[1] A third type of model, a *Joint Model*, which we do not explore, would extend a Paired Model with a more global view of discourse structure.

The two models are complementary in that they address different aspects of computational sentence processing. The Pipeline Model is a linguistically deep model which is useful for making qualitative predictions concerning the strength of communication between discourse and syntactic information during sentence processing. The Paired Model, being the simpler of the two, is able to address a major shortcoming of the Pipeline Model, namely, its coverage, allowing it to make quantitative predictions.

The main question addressed by the Pipeline Model involves two hypotheses that have been put forward to explain discourse-syntax interactions. The **Weakly Interactive** hypothesis (Altmann & Steedman, 1988) 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 initial choices to the syntactic processor. There is experimental evidence in favor of both hypotheses. Evidence compatible with Weak Interaction comes from experiments in which temporary syntactic ambiguities are manipulated to induce *garden path* effects, where an initially preferred syntactic analysis is disconfirmed by later input in the sentence, causing processing difficulty. The general finding is that supportive contexts can reduce the difficulty of the garden path, relative to appropriate baseline conditions. 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 utility of the Pipeline Model in investigating these hypotheses arises from the fact that the Weakly Interactive hypothesis entails a pipelined structure, and yet the Pipeline Model can nonetheless successfully simulate the results of Grodner et al. (2005).

Our goal in introducing the Pipeline Model was to simulate experimentally observed contrasts. However, we found that the model was too complex to accurately produce broad-coverage quantitative predictions which could be evaluated on a large scale corpus of naturalistic reading beheavior. Therefore, we produced a much simpler model, the Paired Model, which can make such predictions. While the inspiration behind the Pipeline Model was to make theoretical claims about human sentence processing, the goal of the Paired Model was much simpler: it aims to use methods of computational modeling to produce a usable model, but we do not use the model to directly draw any theoretical conclusions.

We begin by introducing the Pipeline and Paired Models in Sections 3 and 4, then move to reviewing the experiments in support of Weak and Strong Interaction in Section 2. In Section 5, these experiments are then used to evaluate the models. Finally, Section 6 discusses an evaluation of the Paired model on broad-coverage parsing behavior. The paper closes with a general discussion in Section 7.

[1] For illustrative purposes, the discourse structures are represented similarly to the boxes used in discourse representation theory (Kamp & Reyle, 1993).

2. Cognitive Experiments

2.1. Discourse and Ambiguity Resolution

There is a large literature on garden path experiments involving context (see, for example, Crain & Steedman, 1985; D. C. Mitchell et al., 1992). These experiments typically involve measuring the effect of different types of context on the processing difficulty induced by a locally ambiguous garden path sentence. The experiments by Altmann & Steedman (1988) involved Prepositional Phrase attachment ambiguity. Other authors (e.g. Spivey & Tanenhaus, 1998) have used reduced relative clause attachment ambiguities. In order to replicate previous results, and to provide data for model evaluation, we performed our own reading-time experiment, which used reduced relative clause ambiguities.[2]

The experimental items all had a target sentence containing a relative clause, and one of two possible context sentences, of which one (see (1-a)) supports the relative clause reading and the other (see (1-b)) does not:

(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.

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 (locally ambiguous) 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 often process them while overtly fixating on the previous word (Rayner, 1998).

The relative clause in the target sentence acts as a *restrictive* relative clause, 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 & 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.

Results Using ANOVA to analyze by-subject total fixation time on the critical *verb+by* region revealed that all conditions with a supportive context were read faster (373 ms) than those with a neutral context (477 ms), a significant main effect of context ($n = 28$, $F = 34.737$, $p < .001$). All conditions with unambiguous syntax were read faster (348 ms) than those with a garden path (501 ms), a significant main effect of ambiguity ($F = 36.173$, $p < .001$). Finally, there was a statistically significant interaction between syntax and discourse, whereby the processing cost for ambiguity was reduced when the supportive context was present, relative to when it was not ($F = 4.904$, $p = .035$). The main effects and interaction were also significant for the by-item analysis ($n = 28$, $F = 32.971$ for context, $F = 51.189$ for ambiguity and $F = 7.783$ for the interaction). The pattern of results, shown in Figure 5(a), could be argued to be compatible with

[2]This experiment was previously reported by Dubey et al. (2010).

either the Weakly Interactive or the Strongly Interactive hypothesis. We will discuss the results in more detail when we evaluate the Pipeline Model below.

2.2. Discourse and Unambiguous Syntax

As mentioned in the Introduction, Grodner et al. (2005) reported 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 reading times for 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 with non-restricted relatives. As there was no garden path, and hence no contextually inappropriate structure for the discourse processor to prune, the authors concluded that this must be evidence for the Strongly Interactive hypothesis. Unlike the garden path experiment above, these results, at first glance, do not appear to be consistent with a Weakly Interactive model. We plot their results in Figure 6(a), and compare them to the predictions of the Pipeline and Paired Models below. Before discussing the modeling results, we give a detailed description of the models.[3]

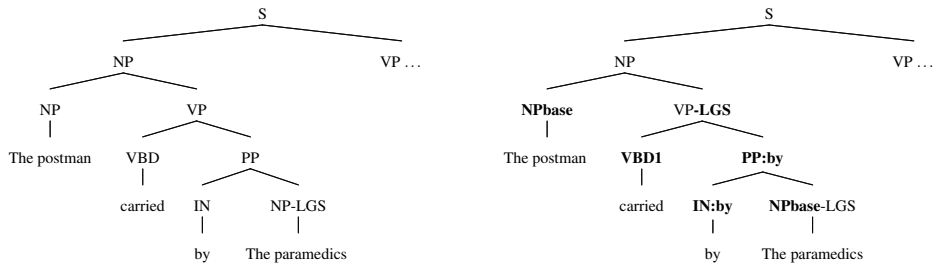
3. Pipeline Model

The Pipeline Model comprises three parts: a parser, a co-reference resolution system, and a pragmatics subsystem. Let us look at each individually.

3.1. Syntactic Component

The parser operates incrementally, building fully connected structures on a word-by-word basis, computing probabilities as it goes. Many authors have suggested various forms of probabilistic parsing algorithms, but most are not capable of producing fully connected structures theorized to be relevant for modeling human sentence processing (Sturt & Lombardo, 2005). A notable exception is the work of Stolcke (1995),

[3]Note that the results in Figure 6(a) are plotted as residual reading times, in which the effect of region length is statistically removed via a linear regression. This means that a negative values are possible, indicating that the region was read faster that what would be predicted by its length. However, for the purposes of theoretical interpretation, it is the *difference* between conditions that is important, rather than the absolute values.



(a) Standard WSJ Tree (b) Minimally Modified Tree

Figure 2. A schematic representation of the smallest set of grammar transformations which we found were required to accurately parse the experimental items.

who modified the Earley top-down parsing algorithm to compute probabilities. Starting with the start symbol *S*, Stolcke's parser incrementally predicts structures, using the input text to validate its hypotheses. Stolcke's approach allows, at each word i, to compute the probability of observing a string up until that word. This probability is known as a *prefix probability*. If the probability of the entire string is $P_{parser}(w,t)$, where w is the text and t is a parse tree, then 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 conditional probability of this word's prefix probability given the previous word's prefix probability:

$$\log P(w_{i-1} \dots w_0) - \log P(w_i \dots w_0) \tag{1}$$

Surprisal scores are interpreted as correlating with reading difficulty, such that the higher the score, the greater difficulty is predicted. For most of the remainder of the paper we will simply refer to the prefix probability at word i as $P(w)$. When using context-free grammars, there is an efficient dynamic programming algorithm to compute Surprisal scores based on an exhaustive search. Once discourse co-reference is added, however, exhaustive search is unfeasible, and we therefore use n-best search. As other authors have found that a relatively small set of analyses can give meaningful predictions (Brants & Crocker, 2000; Boston et al., 2011), we set $n = 10$ for Surprisal computation (the parser does, though, keep more analyses in its chart).

The parser is trained on the Wall Street Journal (WSJ) section of the Penn treebank. Unfortunately, the standard WSJ grammar is not able to assign correct incremental parses to all of our experimental items. We found we could resolve this problem by using four simple transformations: which are shown in Figure 2: (i) adding valency information to verb POS tags (e.g. VBD1 represents a transitive verb); (ii) we lexicalize prepositions and prepositional phrases (e.g. PP:by instead of PP) so that a context-free grammar can find by-phrases; (iii) VPs containing a logical subject (i.e. the agent), are renamed to VP-LGS; (iv) non-recursive NPs are renamed NPbase. Note that the co-reference system treats each NPbase as a markable.

3.2. Co-Reference Component

When processing a document, humans are usually able to effortlessly decide what a word refers to in the world, and, in doing so, decide which words refer to the same thing. The task of simulating this computationally is known as *co-reference resolution*.

Consider the following example sentence from the Dundee Corpus:

(5) Ten years later *the magazine*$_A$ was maintaining *its image*$_B$ with *raunchy coverlines*$_C$. But *it*$_D$ also

Expression	Meaning
$Coref(x,y)$	x is co-referent with y.
$First(x)$	x is a first mention.
$Order(x,y)$	x occurs before y.
$SameHead(x,y)$	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

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: Network templates used in the Markov Logic Network

had *Jill Tweedie*$_E$ confirming the *magazine*$_F$'s *commitment*$_G$ both to *good writing*$_H$ and *the goal of financial independence*$_I$ for *its readers*$_J$.

Here, the subscripts are indices for mentions. The problem of co-reference resolution can be either cast as linking these mentions to conceptual entities, or to simply deciding which of these mentions co-refer with previous ones, without having an explicit extentional model of entities. For the Pipeline Model, we will consider the latter of these two alternatives, and will not explicitly model entities.

The existence of a co-referential relationship between two mentions, say, A and F, is modeled using the predicate $Coref(A,F)$. If the mention A is discourse new, we posit that the predicate $First(A)$ is true. Each mention in a co-reference chain is transitively linked to all others: if $Coref(A,F)$ is true and $Coref(A,D)$ is true, then $Coref(D,F)$ is probably true. As transitivity cannot be enforced by a simple classifier, we use Markov Logic (Richardson & Domingos, 2006), a probabilistic logic, which does allow us to include such soft constraints.

Markov Logic attempts to combine logic with probabilities by using a Markov random field whereby a logical database D containing grounded logical formulae F_i is interpreted probabilistically according to the formula:

$$P(D) = \frac{1}{Z} exp^{\Sigma_i w_i n_i(D)}$$

Where $n_i(D)$ is the number of times that F_i is true in D, and w_i is a weight set in training. Compared to strict first-order logic, Markov logic relaxes the requirement that a predicate be true. Whereas, given some grounded facts, each F_i in D can be either true or false, Markov logic allows gradients of truth, from full certainty to full impossibility.

The Markov Logic Network (MLN) we used for our system uses similar predicates as the MLN-based co-reference resolution system of Huang et al. (2009).[4] Our MLN uses the predicates listed in Table 1. These serve as the F_i that we wish to compute probabilities over.

Two predicates, *Coref* and *First*, are the output of the MLN – they provide a labelling of co-reference 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 co-reference 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 mentions 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.) These features capture some of the basic intuitions about constraints on co-reference resolution, and are enough to correctly predict co-reference in our experimental data.

As our main focus is not to produce a state-of-the-art co-reference system, we do not include predicates which are irrelevant for our simulations, even if they have been shown to be effective for co-reference resolution. For example, the experimental data contains no appositions, so we leave out predicates which fire when two mentions are in an apposition relationship.

Table 2 lists the actual logical formulae which are used as features in the MLN. It should be noted that, because we are assuming an order on the arguments of $Coref(x,y)$, we need three formulae to capture transitivity relationships. To test that the co-reference resolution system was producing meaningful results, we trained our system on the training section of a corpus annotated with co-reference, which is known as the ACE-2 corpus. This corpus explicitly lists mentions and entities, and states which mentions belong to which entities. Using b^3 scoring (Bagga & 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 the parser outputs, 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 co-reference relationships as well as the probability that each individual co-reference assignment is true. Taken together, they allow us to calculate, for a co-reference 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. We sum over all trees, but note that we only consider *one* possible assignment of NPs to co-reference entities per parse, as we only retrieve the probabilities of the MAP solution. Overall, we have:

[4] Poon & Domingos (2008) also developed a Markov logic co-reference resolver, but it was geared toward unsupervised inference, and therefore was not suitable for our needs.

$$P(w) = \sum_c \sum_t P(c,w,t) \approx \sum_t P_{coref}(c|w,t) P_{parser}(w,t)$$

3.3. Pragmatics Component

The effect of context in the experiments described in Section 2 cannot be fully explained using a co-reference 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 specialized 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 three tuple where the first 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 conjecture that there is little processing cost (and hence a high probability) if each element in the 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 relative frequency of discourse old versus new can be readily estimated from a corpus. In the ACE-2 corpus, we found these to occur with probability .58 (discourse old) and .41 (discourse new). Incoherent structures, in the form of unsupported restrictive relative clauses, are rare and hard to find in corpora. Thus, we estimate this value from expert knowledge (Pearl, 1989), and define a probability of .8 for coherent structures and .2 for incoherent ones.

The first two elements of the 3-tuple are provided by the parser and co-reference system, respectively. The third is found by a simple post-processing check to find anaphoric contrast sets. In general, though, finding contrast sets is quite a difficult problem (Modjeska et al., 2003).

The overall prefix probability can then be computed as:

$$P(w) = \sum_{p,c,t} P_{prag}(p|w,t,c) P_{coref}(c|w,t) P_{parser}(w,t)$$

This quantity is then substituted in Equation (1) to get a Surprisal prediction for the current word.

4. Paired Model

The Pipeline Model is suitable for modeling small data sets such as the sets of materials used in psycholinguistic experiments. However, the model is not designed to make predictions on a broad-coverage corpus such as the Dundee Corpus. A particular problem is that the Pipeline Model is geared towards modeling particular relative clause constructions which must be regarded as outliers in corpora, as noted by Roland et al. (2012).

The Paired Model, in order to capture broad-coverage behavior, is based upon a much simpler architecture than the linguistically deep Pipeline Model. The Paired Model utilizes an NP chunker based on a probabilistic finite state machine known as a hidden Markov model (HMM). An HMM, an example of which is shown in Figure 3, has a hidden state associated with each word. The probability of a word depends upon the state, and the probability of each state depends upon the previous state. This probabilistic dependency is shown via the arrows in Figure 3. In our case, the state associated with a word is a finite-state approximation of a syntactic constituent which spans the word.

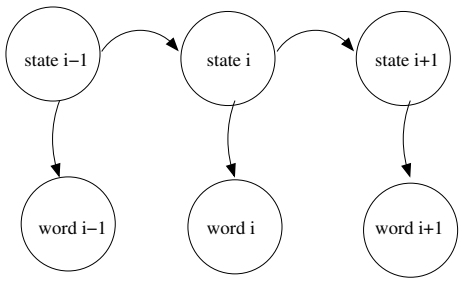


Figure 3. A schematic of a hidden Markov model.

Using a simple model such as an HMM facilitates the integration of a co-reference component, and simplifies the computation of forward probabilities, which are required to compute Surprisal scores. The key insight in this model is that human sentence processing is facilitated when an NP refers to a previously-mentioned discourse entity, relative to when it introduces a new discourse entity. This facilitation depends upon keeping track of a list of previously-mentioned entities, which requires (at the least) shallow syntactic information, yet the facilitation itself is modeled in terms of Surprisal, primarily as a lexical phenomenon.

4.1. Syntactic Component

A key feature of the co-reference component of our model (described below) is that syntactic analysis and co-reference resolution happen simultaneously. As noted above, we use a finite state approximation of the syntax, which is found using an HMM. The particular approximation we use is to label each word as either belonging to an NP or not. In this manner, we can re-create the NP 'chunks' of the utterance. Finding the NP chunks is sufficient to extract NP discourse mentions and, as we show below, Surprisal values computed using HMM chunks provide a useful fit on the Dundee eye-movement data.

NP chunks are found by augmenting part-of-speech tags with the labels B-NP, I-NP, and O, where B-NP denotes the start of a base-level NP chunk, I-NP denotes a word inside a base-level NP chunk, and O denotes a word inside any other category.

It is not possible to find relative clauses using an NP chunker. In the case of our experimental items, though, we can approximate this using two additional piece of information: we note during training when *by* introduces a logical subject, and attempt to guess this when interpreting input. In addition, we use the presence of *who* to guess where a relative clause begins.

The model uses unsmoothed bi-gram transition probabilities, along with a maximum entropy distribution to guess unknown word features. The resulting distribution has the form $P(tag|word)$ and is therefore unsuitable for computing Surprisal values. However, using Bayes' theorem we can compute $P(word|tag) = P(tag|word)P(word)/P(tag)$, which is what we need for Surprisal. The primary information from this probability comes from $P(tag|word)$, however, reasonable estimates of $P(tag)$ and $P(word)$ are required to ensure the probability distribution is proper. $P(tag)$ may be estimated on a parsed treebank, and $P(word)$ is estimated for words observed during training. However, word probabilities are hard to estimate for unseen words. Given that our training data contains approximately 10^6 words, we assume that this probability must be bounded above by 10^{-6}. As an approximation, we use this upper bound as the probability of $P(word)$ when the word is unseen.

Training The chunker is trained on sections 2–22 of the Wall Street Journal section of the Penn Treebank. As we discuss under *Name Recognizer* below, a modification was used for names. Evaluating the perfor-

mance of the chunker alone, we get an F-score of 85.5%. To compare this to previous results, we note that CoNLL 2000 included chunking as a shared task (Tjong Kim Sang & Buchholz, 2000). Our chunker is not comparable to the systems in the shared task for several reasons: we use more training data and we tag simultaneously (the CoNLL systems used gold standard tags). The best performing chunker from CoNLL 2000 achieved an F-score of 93.5%, and the worst performing system an F-score of 85.8%.

4.2. Co-Reference Component

In a standard HMM, the emission probabilities are computed as $P(w_i|s_i)$ where w_i is the i^{th} word and s_i is the i^{th} state. In our model, we replace this with a choice between two alternatives:

$$P(w_i|s_i) = \begin{cases} \lambda P_{\text{seen before}}(w_i|s_i) \\ (1-\lambda) P_{\text{discourse new}}(w_i|s_i) \end{cases} \quad (2)$$

One alternative is the case that the word is from an NP which was seen before, and the other that the current word comes from an NP which is discourse new. While the probabilities could be summed to one, each represents a different hypothesis about the discourse structures, and these hypotheses are kept separate. The 'discourse new' probability distribution is the standard HMM emission distribution. The 'seen before' distribution is more complicated. In general, the 'seen before' distribution could handle all forms of co-reference. In this model, we restrict it to two: word-for-word co-reference and mentions of people (including both proper names and pronouns). The word-for-word co-reference model is in part based upon caching language models (e.g. Kuhn & De Mori, 1990). A language model attempts to assign probabilities to words in a document, and a caching language model keeps track of previously seen words, and assigns them a higher probability than unseen words. The Paired Model differs from caching language models in two ways. First, the contents of the cache are not individual words but rather a collection of NPs mentioned so far in the document. Second, cache entries are annotated with gender and number attributes. Third, pronouns are allowed to refer to cache entries based upon agreeing with the gender and number attributes.

If an NP co-refers to a previous mention, there is no ambiguity about which mentions it is co-referent with: pronouns are co-referent with the closest NP which matches gender and number, and full NPs are co-refer-rent with NPs that have a prefix word-for-word match. A Joint Model, which we do not explore here, would consider a broader collection of NPs as co-reference candidates, possibly using a similar set of features as the Pipeline Model.

At the end of each sentence, the NPs of the Viterbi parse are added to the list of discourse entities[5] after having their leading articles stripped. After being augmented with probabilistically guessed gender and number information, this list is the Paired Model's representation of the discourse. A weakness of the algorithm is that mentions are only added on a sentence-by-sentence basis (disallowing within-sentence references). Although the algorithm is intended to find whole-string matches, in practice, it will count any NP whose prefix matches as being co-referent.

Because of the simplicity of the syntactic component, we ignore syntactic constraints on pronoun resolution such as the Binding Principles, and simply match a pronoun to the nearest compatible antecedent when using the 'seen before' distribution. While the notion of a nearest match is used to build a co-reference chain, it is not used in the probabilistic model. The only notion of salience in the model is an overall trend to reduce the saliency of all NPs as more NPs are introduced into the discourse.

A consequence of Equation 2 is that co-reference resolution is handled at the same time as HMM decoding. Whenever the 'seen before' distribution is applied, an NP is co-referent with one occurring earlier, whenever the 'discourse new' distribution is used, it does not have an antecedent. As these are the only two options, the decoder therefore also selects a chain of co-referent entities.

[5]Technically, they are stored in a trie rather than a list.

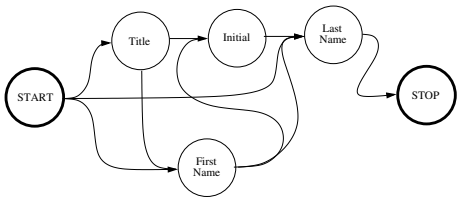


Figure 4. The automaton used to modify the training data. Not shown are the junk states, which sit between nodes on the above graph and which can expand to arbitrary words.

Name Recognizer Any time an NP is added to the salience list, its syntactic gender and number features are guessed as well. In most cases, we tag the gender as Unknown, and use the last noun to guess the number (the WSJ annotation format encodes the number as part of the tag). An exception is the case of proper names. A list of titles and first names associated with a particular gender were extracted from the Yago knowledge base (Suchanek et al., 2007). In addition, we also extracted middle initial and last names, although without gender. This gives us a distribution of words given gender and title; words given gender and firstname; and words given lastname. Thus, whenever these states occur in the HMM, the alternative distribution from Yago is used as opposed to the treebank distribution. As gender is part of the HMM state for firstnames and titles, this allows us to guess the gender during HMM decoding. This leaves open the problem of deciding when to use the new states of title (plus gender), firstname (plus gender) and lastname. This was accomplished by modifying the training data using the regular automaton depicted in Figure 4. In short, words were generated using the regular automaton as well as using the standard proper noun distribution. If the name had higher probability, the name states replaced the proper noun tag in the training data. HMM Training then proceeded as normal.

Training The co-reference component only has one free parameter, λ, which is estimated from the ACE-2 corpus. The estimate is computed by counting how often a repeated NP actually is discourse new. In the current implementation of the model, λ is constant throughout the test runs. However, we have also attempted simulations where λ varied with the nature of the NP (full NP or pronoun), but we did not notice a substantial difference in results. However, it is also possible to use λ as a function of the previous discourse, allowing for more complicated classification probabilities.

4.3. Pragmatics Component

While the focus of the Paired Model is to simulate broad-coverage behavior, it is also possible to use the model to simulate the results of the same experiments the Pipeline Model is used for. This is done by recognizing when a relative clause may cause incoherence. This rarely happens in corpora, but using a similar method as the Pipeline Model, it is possible to include this in the model. As the co-reference system in the Paired Model already has a discount for discourse new mentions, we only need to recognize incoherent relative clauses. This is done using the same mechanism as the Paired Model, but it requires a method for detecting relative clauses. This cannot be done using an HMM in the general case, but for the experimental items of interest, we can detect the relevant cases by using the information about by-phrases which introduce logical subjects, and cases when *who* acts as a relativizer, which are returned by the chunker with high accuracy on the experimental items. The pragmatics component is turned off for the broad-coverage experiments.

5. Evaluation: Experimental Modeling

In this section, we describe the evaluation of the Pipeline and Paired Models on the experimental data from Section 2. For both models, we compute word-by-word Surprisal scores against the garden path and the Grodner et al. (2005) experiments.

5.1. Method

The syntactic components of the Pipeline and Paired models were trained on the Wall Street Journal section of the Penn Treebank. In addition to the transformations mentioned in Section 3, trees were stripped of co-indexation information, empty nodes, and all grammatical functions other than -LGS. In the case of the Pipeline Model, the co-reference component was trained on the training section of ACE-2.

When simulating the experimental sentences, context sentences were treated differently than target sentences in two ways. First, we did not compute word-by-word Surprisal scores for the context sentence, to avoid unnecessary bookkeeping. Second, to further simplify the model, the simulator was (correctly) told in advance that NPs in the context sentence were discourse new.

When modeling the garden path experiment, 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 at the token after the initial NP. This is the earliest point at which there is evidence for a relative clause modifying the initial NP, and presence or absence of the comma signals the presence of a restrictive or nonrestrictive clause. Theoretically, it would also be possible to measure Surprisal over the entire relative clause, a much larger region. While this may be reasonable to do for a full syntactic parser, the results of the Paired Model, which uses an HMM, would likely not be interpretable, as it does not have enough syntactic information to correctly find the boundaries of a relative clause.

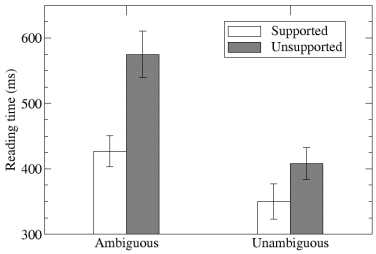
In addition to the overall Surprisal values, we also compute syntactic Surprisal scores, to test if there is any benefit from the discourse and pragmatics subsystems. As we are outputting n best lists for each parse, it is also straightforward to compute other measures which predict reading difficulty,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

For our 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 per model. 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 per model. For each run, the model was reset, purging all discourse information gained while reading earlier items. As neither model is 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.

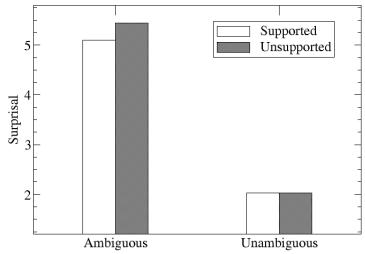
5.2. Results

Garden Path Experiment The results of the simulation of our experiment are shown in Figure 5. Comparing the full Pipeline Model simulation results in Figure 5(b) and the full Paired Model simulation results in Figure 5(c) to the experimental results in Figure 5(a), we find that the models, like the actual experiment, find 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 (ambiguity) slows down reading, and an interaction, such that the effect of context is strongest in the ambiguous condition. All these effects were significant at $p < 0.01$ for both models.

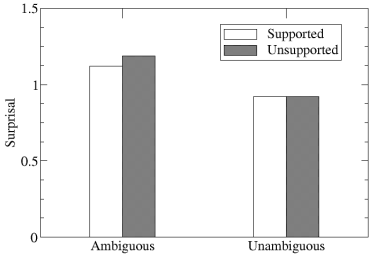
The pattern of results between the full simulation and the experiment differed in two ways. First the results of the simulation suggested a much larger reading difficulty due to ambiguity than the experimental



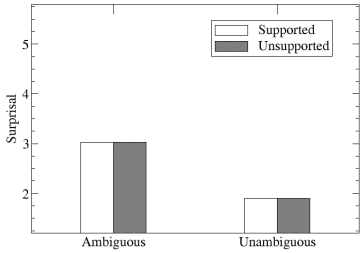
(a) Results from our garden path experiment



(b) Pipeline Model: Simulation of our garden path experiment



(c) Paired Model: Simulation of our garden path experiment



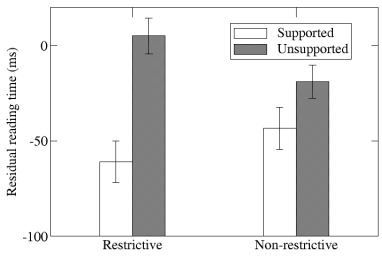
(d) Pipeline Model: Syntax-only simulation

Figure 5. The results of the simulation 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*.

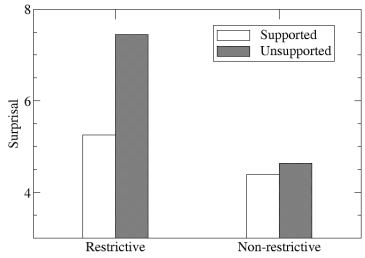
results. The magnitude of the difference was larger for the Pipeline Model than the Paired Model, due to its simpler representation of syntax. Also, in the unambiguous case, the models predicted a null cost of an unsupportive context on the word *by*, because both models bear the cost of an unsupportive context earlier in the sentence, and assume no spillover to the word *by*.

Figure 5(d) shows the result of the syntax-only simulation for the Pipeline Model. The syntax-only results for the Paired Model are similar. We observe that the syntax-only simulations only produced a main effect of ambiguity, and was not able to model the effect of context.

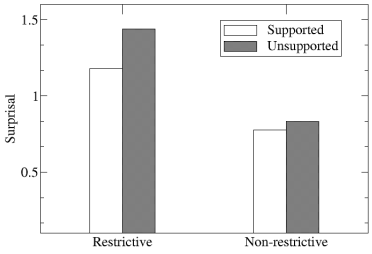
Grodner et al. Experiment The simulation results of the Grodner et al. experiment are shown in Figure 6. In this experiment, the pattern of Pipeline Model simulation results in Figure 6(b) showed a closer resemblance to the experimental results in Figure 6(a) than the garden path experiment. A similar pattern holds for the Paired Model in Figure 6(c). In both cases, 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 6(d), there was no effect of context in the Pipeline 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.



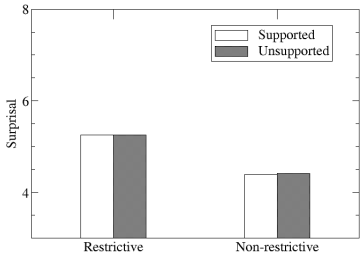
(a) Results from the Grodner et al. experiment



(b) Pipeline Model: Simulation of the Grodner et al. experiment



(c) Paired Model: Simulation of the Grodner et al. experiment



(d) Pipeline Model: Syntax-only simulation

Figure 6. The results of the simulation predict the outcome of the Grodner et al. experiment.

5.3. Discussion

We have shown that, for both the Pipeline and Paired Models, our incremental sentence processor augmented with discourse processing can successfully simulate syntax-discourse interaction effects which have been shown in the literature. The success of the Pipeline Model is of particular interest, with regards to the debate between those who posit a Weakly Interactive model of human sentence processing, and those who argue for a and Strongly Interactive model. The difference between strong and weak interaction can be thought of computationally in terms of a pipeline architecture versus joint inference. Note that the distinction we make between a Weakly and Strongly Interactive model follows closely the claims laid out by Altmann and Steedman (1988): the parser in a Strongly Interactive model is guided by discourse information, the parser in a Weakly Interactive model is not. In our case, the syntactic parser of the Pipeline Model does not make use of discourse information, so it fits into Altmann and Steedman's definition of a Weak Interaction. Unlike Altmann and Steedman, who posited that the discourse processor actually removes parsing hypotheses, we were able to simulate this pruning behavior by simply re-weighting parses in our co-reference 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 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.

6. Evaluation: Broad-coverage Modeling

In the previous section, we found that both the Pipeline and Paired Models are able to simulate the experimental results containing discourse/syntax interactions. In this section, we move on to modeling broad-coverage behavior. From this point on, we deal exclusively with the Paired Model, as the Pipeline Model is not designed for broad coverage processing.

6.1. Data and Evaluation

Our evaluation experiments were conducted upon the Dundee Corpus (Kennedy et al., 2003), which contains the eye-movement record of 10 participants each reading 2,368 sentences of newspaper text. This data set has previously been used to validate probabilistic sentence processing models by several authors, including Demberg & Keller (2008) and Frank (2009).

Eye tracking data is noisy for a number of reasons, including the fact that experimental participants can look at any word which is currently displayed. While English is normally read in a left-to-right manner, readers often skip words or make regressions (i.e., look at a word to the left of the one they are currently fixating). Deviations from a strict left-to-right progression of fixations motivate the need for several different measures of eye movement. The Paired Model predicts the Total Time that participants spent looking at a region, which includes any re-fixations after the reader looks away (we found a similar pattern of results in other measures). We compute several Surprisal scores: one for the syntax-only HMM, which does not have access to co-reference information (henceforth referred to as 'HMM') and another for the full Paired Model, which combines the syntax-only HMM with the co-reference model (henceforth 'HMM+Ref').

To determine if our Dundee Corpus simulations provide a reasonable model of human sentence processing, we perform a regression analysis with the Dundee Corpus reading time measure as the dependent variable and the Surprisal scores as the independent variable. To account for noise in the corpus, we also use a number of additional explanatory variables which are known to strongly influence reading times. These include the logarithm of the frequency of a word (measured in occurrences per million) and the length of a word in letters. Two additional explanatory variables were available in the Dundee corpus, which we also included in the regression model. These were the position of a word on a line, and which line in a document a word appeared in. As participants could only view one line at a time (i.e., one line per screen), these covariates are known as line position and screen position, respectively.

Our choice of covariates were informed by our choice of reading time measure, which in turn was informed by the particular class of behaviors we were investigating. Our experimental results showed that effects of discourse were often observed in reading-time measures which include refixations. Other authors, for example Demberg & Keller (2008), utilize reading time measures which do not include refixations. To explain the variance due to refixations, they need to add additional covariates such as launch distance (a saccade launched from the same word will result in a shorter fixation, as some lexical processing will have already occurred), or and fixation landing position (landing near the edge of the word may trigger a re-fixation near the middle of the word). As these behaviors are not relevant when all fixations are counted, they were not included in the model.

All the covariates, including the Surprisal estimates, were centered before including them in the regression model. Because the HMM and HMM+Ref Surprisal values are highly collinear, the HMM+Ref Surprisal values were added as residuals of the HMM Surprisal values.

From	To	Δ AIC	Δ BIC	Δ logLik	χ^2	Significance
Baseline	HMM	-80	-69	41	82.11	$p < .001$
Baseline	HMM+Ref	-99	-89	51	101.54	$p < .001$
HMM	HMM+Ref	-19	-8	11	21.42	$p < .001$

Table 3: Model comparison for the Paired Model

In a traditional regression analysis, one must assume that either the sample of participants or the particular choice of items contribute variance to the experiment, and this means either that the responses of each participant are analyzed, averaging over items (i.e. treating participants as a random factor), or otherwise, that responses made in response to each item are analyzed, averaging over participants (i.e. treating items as a random factor). However, in the present analysis we utilize a mixed effects model, which allows both items and participants to be treated as random factors simultaneously.[6]

There are a number of criteria which can be used to test the efficacy of one regression model over another. These include the Aikake Information Criterion (AIC), the Bayesian Information Criterion (BIC), which trade off model fit against the number of model parameters (lower scores are better). It is also common to compare the log-likelihood of the models (higher log-likelihood is better), in which case a χ^2 can be used to evaluate if a model offers a significantly better fit, given the number of parameters is uses. We test three models: (i) a baseline, with only low-level factors as independent variables; (ii) the HMM model, with the baseline factors plus Surprisal computed by the syntax-only HMM; and (iii) the HMM+Ref model which includes the raw Surprisal values of the syntax-only HMM and the Surprisal of the HMM+Ref models as computed as a residual of the HMM Surprisal score. We compare the HMM and HMM+Ref to the baseline, and the HMM+Ref model against the HMM model.

Some of the data needed to be trimmed. If, due to data sparsity, the Surprisal of a word goes to infinity for one of the models, we entirely remove that word from the analysis. This occurred seven times with the HMM+Ref model, but did not occur at all with the HMM model. Some of the eye-movement data was trimmed, as well. Fixations on the first and last words of a line were excluded, as were tracklosses. However, we did not trim any items due to abnormally short or abnormally long fixation durations.

6.2. Results

The result of the model comparison on Total Time reading data is summarized in Table 3. We found that both the HMM and HMM+Ref provide a significantly better fit with the reading time data than the Baseline model; all three criteria agree: AIC and BIC are lower than for the baseline, and log-likelihood is higher. Moreover, the HMM+Ref Paired Model provides a significantly better fit than the HMM model, which demonstrates the benefit of co-reference information for modeling reading times. Again, all three measures provide the same result. Table 4 corroborates this result. It lists the mixed-model coefficients for the HMM+Ref model and shows that all factors are significant predictors, including both HMM Surprisal and residualized HMM+Ref Surprisal.

[6]We assume that each participant and item bias the reading time of the experiment individually. Such an analysis is known as having random intercepts of participant and item. It is also possible to assume a more involved analysis, using *random slope* parameters, where the participants and items bias the slope of the predictor. The mixed effects model did not converge when using random intercept and slopes on both participant and item. If random slopes on items were left out, the HMM regression model did converge, but the HMM+Ref model did not. As the HMM+Ref is the model of interest random slopes were left out entirely to allow a like-with-like comparison between the HMM and HMM+Ref regression models.

Coefficient	Estimate	Std Error	t-value
(Intercept)	991.43	23.79	41.66
log(Word Frequency)	-55.30	1.48	-37.29
Word Length	128.62	1.46	87.63
Screen Position	-1.77	0.13	-13.40
Line Position	10.15	0.73	13.75
HMM	12.12	1.33	9.07
HMM+Ref	19.27	4.16	4.63

Table 4: Coefficients of the HMM+Ref Paired Model on Total Reading Times. Note that $t > 2$ indicates that the factor in question is a significant predictor.

7. Discussion

The main result of this paper is that it is possible to produce Surprisal-based sentence processing models which can simulate the influence of discourse on syntactic processing. These models are the result of a novel combination of combining syntactic analyzers with probabilistic models of co-reference and discourse coherence.

As far as we are aware, the work presented here represents the first attempt to create a wide-coverage model of human syntactic processing that incorporates discourse information. In fact, only a very few previous computational models have tackled the question of how discourse information is used in human syntactic parsing, and they have been limited in scope in various ways. For example, Niv (1993) proposed a model based on Combinatory Categorial Grammar (Steedman, 1989), which was able to use discourse information as a filter on syntactic analyses proposed by a parser, thus embodying *weak interaction*. However, although Niv's model does make predictions about which reading of a syntactic ambiguity would be selected given information about the context, it does not incorporate probabilistic information, and it is not obvious how it might account for graded processing costs in unambiguous sentences, such as seen in the data reported Grodner et al. (2005), in contrast to the Pipeline model described above. Other work has attempted to make quantitative predictions of reading times of syntactically ambiguous sentences, using multiple constraints, including discourse information (e.g. Spivey & Tanenhaus, 1998). However, this work was been limited to the consideration of particular sentence types, and would not generate reading-time predictions for unrestricted text, in contrast to the Paired model described above. Thus, we believe that the models described in this paper make a substantial contribution over and above previous work in this area.

In particular, the Pipeline Model was able to simulate both garden path and unambiguous sentences, and the use of Markov Logic allowed the module to compute well-formed co reference chains. Our primary cognitive finding is that the Pipeline 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 favor of the Strongly Interactive hypothesis may be weaker than thought, and that conceptually simpler models may be adequate to capture the data in this domain. More generally, a major contribution of this work is to emphasize the utility of interpreting experimental data in the light of the predictions of explicit computational models. While probabilistic models have been successfully used at lexical and syntactic levels, we demonstrate that they can be fruitfully employed for the investigation of cross-domain phenomena. Indeed, in such a complex domain, computational models are invaluable in clarifying the consequences of abstract theoretical models, as we have shown here.

In light of the Paired Model's success in modeling eye tracking data from the Dundee Corpus, it is important to bear in mind the strengths and weaknesses of each model. Indeed, an interesting direction lies in creating a synthesis of the two models. On one hand, extending the Paired Model's co-reference classifier with a globally correct one is a difficult enterprise. Because the Paired Model closely ties syntactic analysis with co-reference resolution, finding a globally correct co-reference solution requires keeping track of syntactic states in all sentences in the discourse. As a document is processed incrementally, it may be necessary to backtrack to previous sentences in order to explore the global search space.

On the other hand, much of the predictive power of the Pipeline Model is the pragmatic coherence classifier, which would be quite simple to apply to the Paired Model. In addition, with a suitable aggressive search strategy (which does not explore the entire search space), it ought to be possible to extend the Paired Model to full parsing. Given the growing interest in probabilistic models of sentence processing, we believe this is a promising area of future research, and may allow contributions not only to psycholinguistic modeling, but also to engineering systems that deal with parsing and reference resolution.

References

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, *30*, 191–238.

Bagga, A., & Baldwin, B. (1998). Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference (LREC 98).*

Boston, M. F., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, *26*, 301–349.

Brants, T., & Crocker, M. (2000). Probabilistic parsing and psychological plausibility. In *Proceedings of 18th international conference on computational linguistics (COLING-2000)* (pp. 111–117).

Crain, S., & Steedman, M. (1985). On not being led down the garden path: the use of context by the psychological syntax processor. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives.* Cambridge University Press.

Crocker, M., & Brants, T. (2000). Wide coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, *29*(6), 647–669.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*, 192–210.

Demberg, V., & Keller, F. (2009). 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).*

Dubey, A. (2010). The influence of discourse on syntax: A psycholinguistic model of sentence processing. In *Proceedings of the 48th annual meeting of the association for computational linguistics (ACL 2010).* Uppsala, Sweden.

Dubey, A., Keller, F., & Sturt, P. (2009). A probabilistic corpus-based model of parallelism. *Cognition*, *109*(2), 193–210.

Dubey, A., Keller, F., & Sturt, P. (2011). A model of discourse predictions in human sentence processing. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP 2011).* Edinburgh.

Dubey, A., Sturt, P., & Keller, F. (2010). The effect of discourse inferences on syntactic ambiguity resolution. In *Proceedings of the 23rd annual cuny conference on human sentence processing (CUNY 2010)* (p. 151). New York City.

Frank, S. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *31st annual conference of the cognitive science society (cogsci 2009).* Amsterdam, The Netherlands.

Grodner, D. J., Gibson, E. A. F., & Watson, D. (2005). The influence of contextual constrast on syntactic processing: Evidence for strong-interaction in sentence comprehension. *Cognition*, *95*(3), 275–296.

Hale, J. T. (2001). A probabilistic earley parser as a psycholinguistic model. In *In proceedings of the second meeting of the north american chapter of the asssociation for computational linguistics.*

Hale, J. T. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, *32*(2), 101–123.

Huang, S., Zhang, Y., Zhou, J., & Chen, J. (2009). Coreference resolution using markov logic. In *Proceedings of the 2009 conference on intelligent text processing and computational linguistics (CICLing 09).*

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

Kamp, H., & Reyle, U. (1993). *From discourse to logic: Introduction to model theoretic semantics of natural language, formal logic and discourse representation theory*. Spinger.

Kennedy, A., Hill, R., & Pynte, J. (2003). The dundee corpus. In *Proceedings of the 12th european conference on eye movement.*

Kuhn, R., & De Mori, R. (1990). A cache-based natural language model for speech recognition. In (Vol. 12).

Levy, R. (2008, March). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Proceedings of the twentieth annual conference on neural information processing systems.*

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

Mitchell, D. C., Corley, M. M. B., & Garnham, A. (1992). Effects of context in human sentence parsing: Evidence against a discourse-baed proposal mechanism. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*(1), 69–88.

Mitchell, J., Lapata, M., Demberg, V., & Keller, F. (2010). Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th annual meeting of the association for computational linguistics.* Uppsala, Sweden.

Modjeska, N. N., Markert, K., & Nissim, M. (2003). 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)* (pp. 176–183). Sapporo, Japan.

Narayanan, S., & Jurafsky, D. (1998). Bayesian models of human sentence processing. In *Proceedings of the 20th annual conference of the cognitive science society (CogSci 98).*

Niv, M. (1993). *A computational model of syntactic processing: Ambiguity resolution from interpretation*. Unpublished doctoral dissertation, University of Pennsylvania.

Padó, U., Crocker, M., & Keller, F. (2006). Modelling semantic role plausability in human sentence processing. In *Proceedings of the 28th annual conference of the cognitive science society (CogSci 2006)* (pp. 657–662).

Pearl, J. (1989). *Probabilistic reasoning in intelligent systems - networks of plausible inference*. Morgan Kaufmann.

Poon, H., & Domingos, P. (2008). Joint unsupervised coreference resolution with markov logic. In *Proceedings of the 2008 conference on empirical methods in natural language processing (EMNLP 2008).*

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Richardson, M., & Domingos, P. (2006). Markov logic networks. *Machine Learning*, *62*(1-2), 107–136.

Roland, D., Mauner, G., OMeara, C., & Yun, H. (2012). Discourse expectations and relative clause processing. *Journal of Memory and Language*, *66*(3), 479 - 508.

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, *27*(4), 521–544.

Spivey, M. J., & Tanenhaus, M. K. (1998). 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.

Steedman, M. J. (1989). Grammar, interpretation and processing from the lexicon. In W. M. Wilson (Ed.), *Lexical representation and process* (pp. 463–504). MIT Press.

Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, *21*(2), 165–201.

Sturt, P., & Lombardo, V. (2005). Processing coordinate structures: Incrementality and connectedness. *Cognitive Science*, *29*, 291–305.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: A Core of Semantic Knowledge. In *16th international world wide web conference (www 2007).* New York, NY, USA: ACM Press.

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

Tjong Kim Sang, E. F., & Buchholz, S. (2000). Introduction to the conll-2000 shared task: Chunking. In *Proceedings of CoNLL-2000 and LLL-2000* (pp. 127–132). Lisbon, Portugal.