Running head: A PROBABILISTIC MODEL OF PARALLELISM

# A Probabilistic Corpus-based Model of Syntactic Parallelism

Amit Dubey and Frank Keller

School of Informatics, University of Edinburgh

10 Crichton Street, Edinburgh EH8 9AB, UK

phone: +44-131-650-4407, fax: +44-131-650-6626

email: {amit.dubey, frank.keller}@ed.ac.uk

Patrick Sturt

School of Philosophy, Psychology and Language Sciences, University of Edinburgh

7 George Square, Edinburgh EH8 9JZ, UK

phone: +44-131-651-1712, fax: +44-131-650-3461

email: patrick.sturt@ed.ac.uk

**Abstract**

Work in experimental psycholinguistics has shown that the processing of coordinate structures is facilitated when the two conjuncts share the same syntactic structure (Frazier, Munn, & Clifton, 2000). In the present paper, we argue that this parallelism effect is a specific case of the more general phenomenon of syntactic priming—the tendency to repeat recently used syntactic structures. We show that there is a significant tendency for structural repetition in corpora, and that this tendency is not limited to syntactic environments involving coordination, though it is greater in these environments. We present two different implementations of a syntactic priming mechanism in a probabilistic parsing model and test their predictions against experimental data on NP parallelism in English. Based on these results, we argue that a general purpose priming mechanism is preferred over a special mechanism limited to coordination. Finally, we show how notions of activation and decay from ACT-R can be incorporated in the model, enabling it to account for a set of experimental data on sentential parallelism in German.

**Keywords:** syntactic priming, syntactic parallelism, cognitive architectures, probabilistic grammars, corpora.

# A Probabilistic Corpus-based Model of Syntactic Parallelism

## 1. Introduction

Over the last two decades, the psycholinguistic literature has provided a wealth of experimental evidence for *syntactic priming*, an effect in which processing facilitation is observed when syntactic structures are re-used. Most work on syntactic priming has been concerned with sentence production (e.g., Bock, 1986; Pickering & Branigan, 1998). Such studies typically show that people prefer to produce sentences using syntactic structures that have recently been processed.

There has been less experimental work investigating the effect of syntactic priming on language comprehension. Work on comprehension priming has shown that, under certain conditions, the processing of a target sentence is faster when that target sentence includes a structure repeated from the prime. Branigan, Pickering, Liversedge, Stewart, and Urbach (1995) showed this effect in whole-sentence reading times for garden path sentences, while Ledoux, Traxler, and Swaab (2007) have more recently shown similar effects using Event Related Potentials. Moreover, work using a picture matching paradigm (Branigan, Pickering, & McLean, 2005) has shown evidence for the priming for prepositional phrase attachments.

A phenomenon closely related to syntactic priming in comprehension is the so-called *parallelism effect* demonstrated by Frazier et al. (2000): speakers processes coordinated structures more quickly when the second conjunct repeats the syntactic structure of the first conjunct. The parallelism preference in NP coordination can be illustrated using Frazier et al.'s (2000) Experiment 3, which recorded subjects' eye-movements while they read sentences like (1):

(1) a. Terry wrote a long novel and a short poem during her sabbatical.

b. Terry wrote a novel and a short poem during her sabbatical.

Total reading times for the underlined region were faster in (1-a), where *short poem* is coordinated with a syntactically parallel noun phrase (*a long novel*), compared to (1-b), where it is coordinated with a syntactically non-parallel phrase.

In this paper, we will contrast two alternative accounts of the parallelism effect. According to one account, the parallelism effect is simply an instance of a pervasive syntactic priming mechanism in human parsing. This priming account predicts that parallelism effects should be obtainable even in the absence of syntactic environments involving coordination.

According to the alternative account, the parallelism effect is due to a specialized *copy mechanism*, which is applied in coordination and related environments. One such mechanism is *copy*-α, proposed by Frazier and Clifton (2001). Unlike priming, this mechanism is highly specialized and only applies in certain syntactic contexts involving coordination. When the second conjunct is encountered, instead of building new structure, the language processor simply copies the structure of the first conjunct to provide a template into which the input of the second conjunct is mapped. Frazier and Clifton (2001) originally intended *copy*-α to apply in a highly restricted range of contexts, particularly in cases where the scope of the left conjunct is unambiguously marked. An example are gapping structures, where parallelism effects are well documented (Carlson, 2002). However, it is clear that a mechanism like *copy*-α could potentially provide an account for parallelism phenomena if allowed to apply to coordination more generally. This is because in a parallel coordination environment, the linguistic input of the second conjunct matches the copied structure, while in a non-parallel case it does not, yielding faster reading times for parallel structures. If the copying account is correct, then we would expect parallelism effects to be restricted to coordinate structures and would not apply in other contexts.

There is some experimental evidence that bears on the issues outlined above. In addition to testing items involving coordination like (1) above, Frazier et al. (2000) also report an experiment in which they manipulated the syntactic parallelism of two noun phrases separated by a verb (as in *a (strange) man noticed a tall woman*). Unlike the coordination case in (1), no parallelism

advantage was observed in this experiment. Taken together, the two experiments appear to favor an account in which parallelism effects are indeed restricted to coordination, as would be predicted by a model based on a copy mechanism. However, the results should be interpreted with caution, because the coordination experiment, which showed the parallelism effect (Frazier et al.'s (2000) Experiment 3) used a very sensitive eye-tracking technique, while the non-coordination experiment, which showed no such effect (Frazier et al.'s (2000) Experiment 4) used the less sensitive technique of self-paced reading.

Apel, Knöferle, and Crocker (2007) described two experiments in German, which had a similar design to the Frazier et al. (2000) experiments summarized above. Like Frazier et al. (2000), they found evidence for parallelism when the two relevant noun phrases were coordinated (their Experiment 1), but not when they were subject and object of the same verb (their Experiment 2). However, although both of Apel et al.'s (2007) experiments used eye-tracking, their conclusion relies on a cross-experiment comparison. Moreover, the non-coordinating contexts considered by both Frazier et al. (2000) and Apel et al. (2007) used sentences in which the relevant noun phrases differed in grammatical function (e.g., *a (strange) man noticed a tall woman*), while two coordinated phrases share the same grammatical function by definition. This may have affected the size of the parallelism effect.

The aim of the present paper is to compare the two accounts outlined above using a series of corpus studies and computational simulations. The basis for our modeling studies is a probabilistic parser similar to those proposed by Jurafsky (1996) and Crocker and Brants (2000). We integrate both the priming account and the copying account of parallelism into this parser, and then evaluate the predictions of the resulting models against reading time patterns such as those obtained by Frazier et al. (2000). Apart from accounting for the parallelism effect, our model simulates two important aspects of human parsing: (i) it is broad-coverage (rather than only covering specific experimental items) and (ii) it processes sentences incrementally.

This paper is structured as follows. In Section 2, we provide evidence for parallelism effects

in corpus data. We first explain how we measure parallelism, and then present two corpus studies that demonstrate the existence of a parallelism effect in coordinate structures and in non-coordinate structures, both within and between sentences. These corpus results are a crucial prerequisite for our modeling effort, as the probabilistic parsing model that we present is trained on corpus data. Such a model is only able to exhibit a parallelism preference if such a preference exists in its training data, i.e., the syntactically annotated corpora we explore in Section 2.

In Section 3, we present probabilistic models that are designed to account for the parallelism effect. We first present a formalization of the priming and copying accounts of parallelism and integrate them into an incremental probabilistic parser. We then evaluate this parser against reading time patterns in Frazier et al.'s (2000) parallelism experiments. Based on a consideration of the role of distance in priming, we then develop more cognitively plausible parallelism model in Section 4 inspired by Anderson et al.'s (2004) ACT-R framework. This model is again evaluated against the experimental items of Frazier et al. (2000). To show the generality of the model across languages and syntactic structures, we also test this new model against the experimental items of Knöferle and Crocker (2006) which cover parallelism in sentential coordination in German. We conclude with a general discussion in Section 5.

## 2. Corpus Studies

### *2.1. Adaptation*

Psycholinguistic studies have shown that priming affects both production (Bock, 1986) and comprehension (Branigan et al., 2005). The importance of comprehension priming at the lexical level has also been noted by the speech recognition community (Kuhn & Mori, 1990), who use so-called caching language models to improve the performance of speech comprehension software. The concept of caching language models is quite simple: a cache of recently seen words is maintained, and the probability of words in the cache is higher than those outside the cache.

While the performance of caching language models is judged by their success in improving

speech recognition accuracy, it is also possible to use an abstract measure to diagnose their efficacy more closely. Church (2000) introduces such a diagnostic for lexical priming: adaptation probabilities. Adaptation probabilities provide a method to separate the general problem of priming from a particular implementation (i.e., caching models). They measure the amount of priming that occurs for a given construction, and therefore provide an upper limit for the performance of models such as caching models.

Adaptation is based upon three concepts. First is the *prior*, which serves as a baseline. The prior measures the probability of a word appearing, ignoring the presence or absence of a prime. Second is the *positive adaptation*, which is the probability of a word appearing given that it has been primed. Third is the *negative adaptation*, the probability of a word appearing given it has not been primed.

In Church's case, the prior and adaptation probabilities are defined as follows. If a corpus is divided into individual documents, then each document is then split in half. We refer to the first half as the prime (half) and to the second half as the target (half).[1] If $\pi$ is a random variable denoting the appearance of a word in the prime half, and $\tau$ is a random variable denoting the appearance of a word $w$ in the target half, then we define the prior probability $P_{Prior}(w)$ as:

$$\begin{aligned} P_{Prior}(w) &= P(\tau = w) \\ &= P(\tau = w|\pi = w)P(\pi = w) + P(\tau = w|\pi \neq w)P(\pi \neq w) \end{aligned} \tag{1}$$

Intuitively, $P_{Prior}(w)$ is the probability that $w$ occurs in the target, independently of whether it has occurred in the prime. As indicated in equation (1), this can be computed by summing the relevant conditional probabilities: the probability that $w$ occurs in the target given that it has occurred in the prime, and the probability that $w$ occurs in the target given that it has not occurred in the prime. According to the rule of total probability, each conditional probability has to be multiplied by the independent probability of the conditioning variable ($P(\pi = w)$ and $P(\pi \neq w)$, respectively).

The positive adaptation probability $P_+(w)$ and the negative adaptation $P_-(w)$ can then be

defined as follows:

$$P_+(w) = P(\tau = w | \pi = w) \tag{2}$$

$$P_-(w) = P(\tau = w | \pi \neq w) \tag{3}$$

In other words, $P_+(w)$ is the conditional probability that the word *w* occurs in the target, given that it also occurred in the prime. Conversely, $P_-(w)$ is the probability that *w* occurs in the target, given that it did not occur in the prime.

In the case of lexical priming, Church observes that $P_+ \gg P_{Prior} > P_-$. In fact, even in cases when $P_{Prior}$ is quite small, $P_+$ may be higher than 0.8. Intuitively, a positive adaptation which is higher than the prior entails that a word is likely to reappear in the target given that it has already appeared in the prime. In order to obtain corpus evidence for priming, we need to demonstrate that the adaptation probabilities for syntactic constructions behave similarly to those for lexical items, showing positive adaptation $P_+$ greater than the prior. As $P_-$ must become smaller than $P_{Prior}$ whenever $P_+$ is larger than $P_{Prior}$, we only report the positive adaptation $P_+$ and the prior $P_{Prior}$.

### *2.2. Estimation*

There are several methods available for estimating the prior and adaptation probabilities from corpora. The most straightforward approach is to compute the maximum likelihood empirical distribution, which can be achieved by simply counting the number of times a word *w* occurs (or fails to occur) in the prime and target positions in the corpus. However, this approach has two shortcomings which preclude us from using it here.

First, the existing literature on priming in corpora (e.g., Gries, 2005; Szmrecsanyi, 2005; Reitter, Moore, & Keller, 2006; Jaeger, 2006a, 2006b; ?, ?) reports that a variety of factors can influence priming, including the distance between prime and target, the type and genre of the corpus, and whether prime and target are uttered by the same speaker. Previous work has used multiple regression methods to study priming in corpora; this approach is particularly useful when

several factors are confounded, as regression makes it possible to quantify the relative contribution of each factor.

A second argument against a simple maximum likelihood approach emerges when we want to carry out statistical significance tests based on the word counts obtained from the corpus. Such tests often require that the occurrence of a word $w_i$ is independent of the occurrence of another word $w_j$. However, this independence assumption is trivially false in our case: if $w_i$ occurs in a certain context, then we know that $w_j$ does not occur in that context. This implies that $P(\tau = w|\pi = w_i)$ and $P(\tau = w|\pi = w_j)$ are not statistically independent (if $w_i$ and $w_j$ are in the same context), therefore we are not able to apply independent statistical tests to these two probabilities (or the underlying corpus counts).

Both these shortcomings can be overcome by using multinomial logistic regression to estimate prior and adaptation probabilities. In multinomial logistic regression, a *response variable* is defined as a function of a set of *predictor variables*. The response variable is multinomial, i.e., it is drawn from a set of discrete categories (in contrast to binary logistic regression, where the response variable can take on only two different values). The predictor variables can either be categorical or continuous. In the case of measuring Church-like lexical adaptation, each possible word corresponds to a category, and the response variable describes the occurrence of word $w_j$ in the target position, while the predictor variable describes the occurrence of word $w_i$ in the prime position. In the priming case, $w_i = w_j$, in the non-priming case, $w_i \neq w_j$.

A multinomial logistic regression model includes a parameter vector $\beta$ for each predictor variable. In the case of lexical adaptation, this means that there is a parameter $\beta_i$ for each word $w_i$ which may occur in the priming position. One of the values of the predictor variable (the value $w_0$) serves as the reference category, which is assumed to have a parameter estimate of 0. After parameter fitting, we have an estimated value $\hat{\beta}$ for the parameter vector, and the positive

adaptation probability can then be computed according to the following formula:

$$\hat{P}_{+}(\tau = w_i) = \frac{\exp(X\hat{\beta}_i)}{1 + \sum_j \exp(X\hat{\beta}_j)} \tag{4}$$

Here, $X$ is a vector in which the $i$-th element takes on the value 1 if $\pi = w_i$, and the value 0 otherwise. Note in particular that we can straightforwardly include predictor variables other than the prime word in our regression model; technically, this corresponds to adding extra parameters to $\beta$ and their corresponding explanatory variables to $X$; mathematically, this amounts to computing the conditional probability $P(\tau = w_i | \pi = w_i, \sigma)$, where $\sigma$ is an additional predictor variable.

The negative adaptation is estimated in the same way as the positive adaption. The prior is estimated as follows, based on the definition in equation (1):

$$\hat{P}_{Prior}(w_i) = \hat{P}_{+}(\tau = w_i)\hat{P}(\pi = w_i) + \hat{P}_{-}(\tau = w_i)\hat{P}(\pi \neq w_i) \tag{5}$$

Where $\hat{P}(\pi = w_i)$ and $\hat{P}(\pi \neq w_i)$ can be estimated directly from the corpus using maximum likelihood estimation, and the positive and negative adaptation terms are estimated using the regression coefficients.

Multinomial logistic regression makes it possible to compute prior and adaptation probabilities while solving both of the problems with maximum likelihood estimation noted above. First, we can freely include additional predictor variables if we want to determine the influence of possible confounding variables on priming. Second, statistical significance tests can now be performed for each predictor variable in the regression, without requiring independence between the categories that a variable can take.

There is one potentially confounding variable in Studies 1 and 2 that we need to pay particular attention to. This is the general tendency of speakers to order syntactic phrases such that short phrases precede long phrases (Hawkins, 1994). This short-before-long preference, sometimes known as Behagel's law, is also attested in corpus data on coordination (Levy, 2002). The short-before-long preference can potentially amplify the parallelism effect. For example, consider a case in which the first conjunct consists of a relatively long phrase. Here, Behagel's law would

predict that the second conjunct should also be long (in fact, it should be even longer than the first). Assuming that long constituents tend to be generated by a certain specific set of rules, this would mean that the two conjuncts would have an above-chance tendency to be parallel in structure, and this tendency could be attributable to Behagel's law alone.

Studies 1 and 2 have the aim of validating our approach, and laying the ground for our modeling work in Studies 3–5. We will use Church adaptation probabilities estimated using multinomial regression to investigate parallelism effects within coordination (Study 1) and outside coordination (Study 2), both within sentences and between sentences. It is important to show that adaptation effects exist in corpus data before we can build a model that learns the adaptation of syntactic rules from corpus frequencies. Our studies build on previous corpus-based work which demonstrated parallelism in coordination (Levy, 2002; Cooper & Hale, 2005), as well as between-sentence priming effects. Gries (2005), Szmrecsanyi (2005), Jaeger (2006a, 2006b), and ? (?) investigated priming in corpora for cases of structural choice (e.g., between a dative object and a PP object or between active and passive constructions). These results have been generalized by Reitter et al. (2006), who showed that arbitrary rules in a corpus can be subject to priming.

### *2.3. Study 1: Parallelism in Coordination*

In this section, we test the hypothesis that coordinated noun phrases in a corpus are more likely to be structurally parallel than we would expect by chance. We show how Church's adaptation probabilities, as defined in the previous section, can be used to measure syntactic parallelism in coordinate structures. We restrict our study to the constructions used in Frazier et al.'s (2000) experiments, all of which involve two coordinated NPs. This ensures that a direct comparison between our corpus results and the experimental findings is possible.

#### *2.3.1. Method*

This study was carried out on the English Penn Treebank (Release 2, annotated in Treebank II style; Marcus et al., 1994), a collection of documents which have been annotated with

parse trees by automatically parsing and then manually correcting the parses. This treebank comprises multiple parts drawn from distinct text types. To ensure that our results are not limited to a particular genre, we used two parts of the treebank: the Wall Street Journal (WSJ) corpus of newspaper text and the Brown corpus of written text balanced across genres. In both cases, we used the entire corpus for our experiments.

In the Penn Treebank, NP coordination is annotated using the rule $\mathsf{NP} \rightarrow \mathsf{NP}_1\ \mathsf{CC}\ \mathsf{NP}_2$ where CC represents a coordinator such as *and*. The application of the adaptation metric introduced in Section 2.1 to such a rule is straightforward: we pick $\mathsf{NP}_1$ as the prime $\pi$ and $\mathsf{NP}_2$ as the target $\tau$. We restrict our investigation to the following syntactic rules:

**SBAR** An NP with a relative clause, i.e., $\mathsf{NP} \rightarrow \mathsf{NP\ SBAR}$.

**PP** An NP with a PP modifier, i.e., $\mathsf{NP} \rightarrow \mathsf{NP\ PP}$.

**N** An NP with a single noun, i.e., $\mathsf{NP} \rightarrow \mathsf{N}$.

**Det N** An NP with a determiner and a noun, i.e., $\mathsf{NP} \rightarrow \mathsf{Det\ N}$.

**Det Adj N** An NP with a determiner, an adjective, and a noun, i.e., $\mathsf{NP} \rightarrow \mathsf{Det\ Adj\ N}$.

Our study focuses on $\mathsf{NP} \rightarrow \mathsf{Det\ Adj\ N}$ and $\mathsf{NP} \rightarrow \mathsf{Det\ N}$ as these are the rules used in the items of Frazier et al. (2000); $\mathsf{NP} \rightarrow \mathsf{N}$ is a more frequent variant; $\mathsf{NP} \rightarrow \mathsf{NP\ PP}$ and $\mathsf{NP} \rightarrow \mathsf{NP\ SBAR}$ were added as they are the two most common NP rules that include non-terminals on the right-hand side.

To count the relative number of occurrences of each prime and target pair, we iterate through each parsed sentences in the corpus. Each time the expansion $\mathsf{NP} \rightarrow \mathsf{NP}_1\ \mathsf{CC}\ \mathsf{NP}_2$ occurs in a tree, we check if one (or both) of the NP daughters of this expansion matches one of the five rules listed above. Each of these five rules constitutes a category of the response and predictor variables in the multinomial logistic regression (see Section 2.2); we also use an additional category ‘other’ that comprises all others rules. This category serves as the reference category for the regression.

As noted in Section 2.1, a possible confound for a corpus study of parallelism is Behagel's law, which states that there is a preference to order short phrases before long ones. We deal with this confound by including an additional predictor in the multinomial regression: the indicator variable $\sigma$ takes the value $\sigma = 1$ if the first conjunct is shorter than the right conjunct, and $\sigma = 0$ if this is not the case. The regression vector $X$ is then augmented to include this indicator variable, and an additional parameter is likewise added to the parameter vector $\beta$ to accommodate the short-before-long predictor variable. In addition to the main effect of short-before-long, we also include a predictor that represents the interaction between adaption and short-before-long.

The addition of the interaction term makes the computation of the prior and adaptation probabilities slightly more complicated. The prior can now be computed according to the following formula:

$$\begin{aligned} P_{Prior}(w) &= P(\tau = w|\pi = w, \sigma = 0)P(\pi = w, \sigma = 0) + \\ &\quad P(\tau = w|\pi \neq w, \sigma = 0)P(\pi \neq w, \sigma = 0) + \\ &\quad P(\tau = w|\pi = w, \sigma = 1)P(\pi = w, \sigma = 1) + \\ &\quad P(\tau = w|\pi \neq w, \sigma = 1)P(\pi \neq w, \sigma = 1) \end{aligned} \tag{6}$$

The positive adaptation probability must likewise include both cases or short-before-long and long-before-short:

$$\begin{aligned} \hat{P}_{+}(\tau = w_i) &= P(\tau = w|\pi = w, \sigma = 0)P(\pi = w, \sigma = 0) + \\ &\quad P(\tau = w|\pi = w, \sigma = 1)P(\pi = w, \sigma = 1) \end{aligned} \tag{7}$$

In addition, we compute two additional probabilities, the first of which is a prior for cases where the short phrase precedes the long one (which we will refer to as the short-before-long prior). This is the probability of seeing a rule in the second conjunct (regardless if it occurred in the first conjunct or not) in cases where short rules precede long rules:

$$\begin{aligned} P_{Short}(w) &= P(\tau = w|\pi = w, \sigma = 1)P(\pi = w, \sigma = 1) + \end{aligned} \tag{8}$$

$$P(\tau = w|\pi \neq w, \sigma = 1)P(\pi \neq w, \sigma = 1)$$

The second probability is $P(\tau = w|\pi = w, \sigma = 1)$, the probability of priming given that the prime is shorter than the target (i.e., $\sigma = 1$). This probability, which in the regression directly corresponds to the coefficient of the interaction between the predictors adaptation and short-before-long, will be referred to as the short-before-long adaptation probability.

Based on the parameter estimates of the multinomial logistic regression, we compute a total of 14 probability values for each corpus. We report the prior probability and positive adaptation probability for all rules; as Behagel's Law only confounds the adaptation of complex rules, we report the short-before-long prior and the short-before-long adaptation probabilities only for these rules (noun phrases of type PP and SBAR).

Our main hypothesis is that the adaptation probability is higher than the prior probability in a given corpus. A stronger hypothesis is that the short-before-long adaptation probability as defined above is also higher than the short-before-long prior probability, i.e., that the parallelism preference holds even in cases that match the preferred phrase order in coordinate structure (i.e., the shorter conjunct precedes the longer one). To test these hypotheses, we perform $\chi^2$ tests which compare the log-likelihood difference between a model that includes the relevant predictor and a model without that predictor. We report whether the following predictors are significant: the main effect of adaptation (corresponding to $P(\tau = w|\pi = w)$), the main effect of short-before-long, and the interaction of adaptation and short-before-long (corresponding to $P(\tau = w|\pi = w, \sigma = 1)$).

*2.3.2. Results*

The probabilities computed using multinomial regression are shown in Figure 1 for the Brown corpus and Figure 2 for the Wall Street Journal corpus. The results of the significance tests are given in Table 1 for Brown and in Table 2 for WSJ. Each figure shows the prior probability (Prior) and the adaptation probability (Adapt) for all five constructions: single common noun (N), determiner and noun (Det N), determiner, adjective, and noun (Det Adj N), NP with PP modifier

(PP), NP with relative clause (SBAR). In addition, the figures show the short-before-long prior probability (Short Prior) and short-before-long adaptation probability (Short Adapt) for PP and SBAR categories.

For both corpora, we observe a strong adaptation effect: the adaptation probability is consistently higher than the prior probability across all five rules. According to the log-likelihood $\chi^2$ tests (see Tables 1 and 2), there was a highly significant effect of adaptation in both corpora.

Turning to the short-before-long adaptation probabilities, we note that these also are consistently higher than the short-before-long prior probabilities in both corpora for the rules PP and SBAR. Short-before-long probabilities cannot be computed for the rules N, Det N, and Det Adj N, as they are of fixed length (one, two, and three words respectively), due to the fact that they only contain pre-terminals on the right-hand side. The log-likelihood $\chi^2$ tests also show a significant interaction of adaptation and short-before-long, consistent with the observation that there is more adaptation in the short-before-long case (see Figures 1 and Figures 2). There is also a main effect of short-before-long, which confirms that there is a short-before-long preference, but this is not the focus of the present study.

*2.3.3. Discussion*

The main conclusion we draw is that the parallelism effect in corpora mirrors the one found experimentally by Frazier et al. (2000), if we assume higher probabilities are correlated with easier human processing. This conclusion is important, as the experiments of Frazier et al. (2000) only provided evidence for parallelism in *comprehension* data. Corpus data, however, are *production* data, which means that our findings provide an important generalization of Frazier et al.'s experimental findings. Furthermore, we were able to show that the parallelism effect in corpora persists even if the preference for short phrases to precede long ones is taking into account.

### *2.4. Study 2: Parallelism in Non-coordinated Structures*

The results in the previous section showed that the parallelism effect found experimentally by Frazier et al. (2000) is also present in corpus data. As discussed in Section 1, there are two possible explanation for the effect: one in terms of a construction-specific copying mechanism, and one in terms of a generalized syntactic priming effect. In the first case, we predict that the parallelism effect is restricted to coordinate structures, while in the second case, we expect that parallelism (a) is independent of coordination, i.e., also applies to non-coordinate structures, and (b) occurs in the wider discourse, i.e., not only within sentences but also between sentences. The purpose of the present corpus study is to test these two predictions.

#### *2.4.1. Method*

The method used was the same as in Study 1 (see Section 2.3.1), with the exception that the prime and the target are no longer restricted to being the first and second conjunct in a coordinate structure. We investigated two levels of granularity: *within-sentence parallelism* occurs when the prime NP and the target NP appear within the same sentence, but stand in an arbitrary structural relationship. Coordinate NPs were excluded from this analysis, so as to make sure that any parallelism effects are not confounded by coordination parallelism as established in Study 1. For *between-sentence parallelism*, the prime NP occurs in the sentence immediately preceding the sentence containing the target NP.

The data for both types of parallelism analysis was gathered as follows. Any NP in the corpus was considered a target NP if it was preceded by another NP in the relevant context (for within-sentence parallelism, the context was the same sentence, for between-sentence parallelism, it was the preceding sentence). If there was more than one possible prime NP (because the target was preceded by more than one NP in the context), then the closest NP was used as the prime. In the case of ties, the dominating NP was chosen to accurately account for PP and SBAR priming (i.e. in in the case of a noun with a relative clause, the entire NP will be chosen over the object of

the relative clause). The NP pairs gathered this way could either constitute positive examples (both NPs instantiate the same rule) or negative example (they do not instantiate the same rule). As in Study 1, the NPs were grouped into six categories: the five NP rules listed in Section 2.3.1, and 'other' for all other NP rules. A multinomial logistic regression model was then fit on this data set.

### *2.4.2. Results*

The results for the within-sentence analysis for non-coordinate structures are graphed in Figures 3 and 4 for the Brown and WSJ corpus, respectively. The results of the statistical tests are given in Tables 3 and 4. We find that there is a parallelism effect in both corpora across NP types. The adaptation probabilities are higher than the prior probabilities, except for two cases: SBAR rules in Brown, and PP rules in WSJ. However, even in these cases, the short-before-long adaptation is higher than the short-before-long prior, which indicates that there is a parallelism effect for structures in which the short phrase precedes the long one. Furthermore, the log-likelihood $\chi^2$ tests show significant effects of adaptation for both corpora, as well as significant main effects of short-before-long, and a significant interaction of adaptation and short-before-long.

A parallelism effect was also found in the between-sentence analysis, as shown by Figures 5 and 6, with the corresponding statistical tests summarized in Tables 5 and 6. In both corpora and for all structures, we found that the adaptation probability was higher than the prior probability. In addition, the short-before-long adaptation was higher than the short-before-long prior, confirming the finding that the parallelism effect persists when the short-before-long preference for phrases is taken into account. (Recall that only the SBAR and PP noun phrases can differ in length, so short-before-long probabilities are only available for these rule types.)

### *2.4.3. Discussion*

This experiment demonstrated that the parallelism effect is not restricted to coordinate structures. Rather, we found that it holds across the board: for NPs that occur in the same sentence (and are not part of a coordinate structure) as well as for NPs that occur in adjacent sentences. Just

as for coordination, we found that this effect persists if we only consider pairs of NPs that respect the short-before-long preference. However, this study also indicated that the parallelism effect is weaker in within-sentence and between-sentence configurations compared to in coordination: The differences between the prior probabilities and the adaptation probabilities are markedly smaller than those uncovered for parallelism in coordinate structure. (Note that Figures 1 and 2 range from 0 to 1 on the x-axis, while Figures 3–6 range from 0 to 0.25.)

The fact that parallelism is a pervasive phenomenon, rather than being limited to coordinate structures, is compatible with the claim that it is an instance of a general syntactic priming mechanism, which has been an established feature of accounts of the human sentence production system for a while (e.g., Bock, 1986). This runs counter to claims by Frazier et al. (2000), who argue that parallelism only occurs in coordinate structures. (It is important to bear in mind, however, that Frazier et al. only make explicit claims about comprehension, not about production.)

The question of the relationship between comprehension and production data is an interesting one. One way of looking at a comprehension-based priming mechanism may be in terms of a more general sensitivity of comprehenders towards distributional information. According to such a hypothesis, processing should easier if the current input is more predictable given previous experience (Reali & Christiansen, 2007). The corpus studies described here and in previous work have shown that similar structures do tend to appear near to each other more often than would be expected by chance. If comprehenders are sensitive to this fact, then this could be the basis for the priming effect. This is an attractive hypothesis, as it requires no additional mechanism other than prediction, and provides a very general explanation that is potentially able to unify parallelism and priming effects with experience-based sentence processing in general, as advocated, for instance, by constraint-based lexicalist models (MacDonald, Pearlmutter, & Seidenberg, 1994) or by the tuning hypothesis (Mitchell, Cuetos, Corley, & Brysbaert, 1996). The probabilistic model we will propose in the remainder of this paper is one possible instantiation of such an experience-based approach.

## 3. Modeling Studies

### *3.1. Priming Models*

In Section 2, we provided corpus evidence for syntactic parallelism at varying levels of granularity. Focusing on NP rules, we found the parallelism effect in coordinate structures, but also in non-coordinate structures, and between adjacent sentences. These corpus results form an important basis for the modeling studies to be presented in the rest of this paper. Our modeling approach uses a probabilistic parser, which obtains probabilities from corpus data. Therefore, we first needed to ascertain that the corpus data includes evidence for parallelism. If there was no parallelism in the training data of our model, it would be unlikely that the model would be able to account for the parallelism effect.

Having verified that parallelism is present in the corpus data, in this section we will propose a set of models designed to capture the priming hypothesis and the copy hypothesis of parallelism, respectively. To keep the models as simple as possible, each formulation is based on an unlexicalized probabilistic context-free grammar (PCFG), which also serves as a baseline for evaluating more sophisticated models. In this section, we describe the baseline, the copy model, and the priming model in turn. We will also discuss the design of the probabilistic parser used to evaluate the models.

#### *3.1.1. Baseline Model*

PCFGs serve as a suitable baseline for our modeling efforts as they have a number of compelling and well-understood properties. For instance, PCFGs make a probabilistic independence assumption which closely corresponds to the context-free assumption: each rule used in a parse is conditionally independent of other rules, given its parent. This independence assumption makes it relatively simple to estimate the probability of context-free rules. The

probability of a rule $N \rightarrow \zeta$ is estimated as:

$$\hat{P}(\zeta|N) = \frac{c(N \rightarrow \zeta)}{c(N)} \tag{9}$$

Where the function $c(\cdot)$ counts the number of times a rule or the left-hand side of a rule occurs in a training corpus.

### *3.1.2. Copy Model*

The first model we introduce is a probabilistic variant of Frazier and Clifton's (2001) copying mechanism: it models parallelism only in coordination. This is achieved by assuming that the default operation upon observing a coordinator (assumed to be anything marked up with a CC tag in the corpus, e.g., *and*) is to copy the full subtree of the preceding coordinate sister. The copying operation is depicted in Figure 7: upon reading the *and*, the parser attempts to copy the subtree for *a novel* to the second NP *a book*.

Obviously, copying has an impact on how the parser works (see Section 3.1.5 for details). However, in a probabilistic setting, our primary interest is to model the copy operation by altering probabilities compared to a standard PCFG. Intuitively, the characteristic we desire is a higher probability when the conjuncts are identical. More formally, if $P_{PCFG}(t)$ is the probability of a tree $t$ according to a standard PCFG and $P_{Copy}(t)$ is the probability of copying the tree $t$, then if two the subtrees $t_1$ and $t_2$ are parallel, we can say that $P_{Copy}(t_2) > P_{PCFG}(t_2)$. For simplicity's sake, we assume that $P_{Copy}(t)$ is a parameter which can be estimated from the training corpus.

A naive probability assignment would decide between copying with probability $P_{Copy}(t_2)$ or analyzing the subtree rule-by-rule with the probability $(1 - P_{Copy}(t_2)) \cdot P_{PCFG}(t_2)$. However, the PCFG distribution assigns some probability to all trees, including a tree which is equivalent to $t_2$ 'by chance'. The probability that $t_1$ and $t_2$ are equal 'by chance' is $P_{PCFG}(t_1)$. We must therefore properly account for the probability of these chance derivations. This is done by formalizing the notion that identical subtrees could be due to either a copying operation or by chance, giving the

following probability for identical trees:

$$P_{Copy}(t_2) + P_{PCFG}(t_1) \tag{10}$$

Similarly, the probability of a non-identical tree is:

$$\frac{1 - P_{PCFG}(t_1) - p_{Copy}}{1 - P_{PCFG}(t_1)} \cdot P_{PCFG}(t_2) \tag{11}$$

This accounts for both a copy mismatch and a PCFG derivation mismatch, and assures the probabilities still sum to one. These definitions for the probabilities of parallel and non-parallel coordinate sisters therefore form the basis of the Copy model.

*Estimation.* We saw in Section 2.3 that the parallelism effect can be observed in corpus data. We make use of this fact to estimate the non-PCFG parameter of the Copy model, $\hat{P}_{Copy}$ (the PCFG parameters are estimated in the same way as for a standard PCFG, as explained above). While we cannot observe copying directly because of the 'chance' derivations, we can use Equations (10) and (11) above to derive a likelihood equation, which can then be maximized using a numerical method. A common approach to numerical optimization is the gradient ascent algorithm (Press, Teukolsky, Vetterling, & Flannery, 1988), which requires a gradient of the likelihood (i.e., the partial derivative of the likelihood with respect to $P_{Copy}$). If we let $c_{ident}$ be the number of trees in the corpus which are identical, and if $j$ counts through the non-identical trees (and $t_{j_1}$ is the first conjunct of the $j$-th non-identical tree in the corpus), then the gradient of the log-likelihood equation (with respect to $P_{Copy}$) is:

$$\bigtriangledown = \frac{c_{ident}}{P_{Copy}} - \sum_j \frac{1 - P_{PCFG}(t_{j_1})}{1 - P_{Copy} - P_{PCFG}(t_{j_1})} \tag{12}$$

This equation is then fed to the gradient ascent algorithm, producing an estimate $\hat{P}_{Copy}$ which maximizes the likelihood of the training corpus. This approach ensures that the copy parameter $\hat{P}_{Copy}$ is set to the optimal value, i.e., the value that results in the best fit with the training data, and thus maximizes our chance of correctly accounting for the parallelism effect.

### *3.1.3. Between Model*

While the Copy model limits itself to parallelism in coordination, the next two models simulate structural priming in general. Both are similar in design, and are based on a simple insight: we can condition a PCFG rule expansion on whether the rule occurred in some previous context. If *Prime* is a binary-valued random variable denoting if a rule occurred in the context, then we can define an adaptation probability for PCFG rules as:

$$\hat{P}(\zeta|N, Prime) = \frac{c(N \rightarrow \zeta, Prime)}{c(N, Prime)} \tag{13}$$

This is an instantiation of Church's (2000) adaptation probability, used in a similar fashion as in our corpus studies in Section 2. Our aim here is not to show that certain factors are significant predictors; rather, we want to estimate the parameters of a model that simulates the parallelism effect by incrementally predicting sentence probabilities. Therefore, unlike in the corpus studies, we do not need to carry out hypothesis testing and we can simply use the empirical distribution to estimate our parameters, rather than relying on multinomial logistic regression.

For our first model, the context is the previous sentence. Thus, the model can be said to capture the degree to which rule use is primed between sentences. We henceforth refer to this as the Between model. Each rule acts once as a target (i.e., the event of interest) and once as a prime.

### *3.1.4. Within Model*

Just as the Between model conditions on rules from the previous sentence, the Within sentence model conditions on rules from earlier in the current sentence. Each rule acts once as a target, and possibly several times as a prime (for each subsequent rule in the sentence). A rule is considered 'used' once the parser passes the word on the leftmost corner of the rule. Because the Within model is finer grained than the Between model, it should be able to capture the parallelism effect in coordination. In other words, this model could explain parallelism in coordination as an instance of a more general priming effect.

### *3.1.5. Parser*

Reading time experiments, including the parallelism studies of Frazier et al. (2000), measure the time taken to read sentences on a word-by-word basis. Slower reading times are assumed to indicate processing difficulty, and faster reading times (as is the case with parallel structures) are assumed to indicate processing ease, and the location of the effect (which word or words it occurs on) can be used to draw conclusions about the nature of the difficulty.

As our main aim is to build a psycholinguistic model of structural repetition, the most important feature of the parsing model is to build structures incrementally, i.e., on a word-by-word basis. In order to achieve incrementality, we need a parser which has the prefix property, i.e., it is able to assign probabilities to arbitrary left-most substrings of the input string.

We use an Earley-style probabilistic parser, which has these properties and outputs the most probable parses (Stolcke, 1995). Furthermore, we make a number of modifications to the grammar to speed up parsing time. The treebank trees contain annotations for grammatical functions (i.e., subject, object, different types of modifier) and co-indexed empty nodes denoting long-distance dependencies, both of which we removed.

The Earley algorithm requires a modification to support the Copy model. We implemented a copying mechanism that activates any time the parser comes across a CC tag in the input string, indicating a coordinate structure, as shown in Figure 7. Before copying, though, the parser looks ahead to check if the part-of-speech tags after the CC are equivalent to those inside first conjunct. In the example in Figure 7, the copy operation succeeds because the tags of the NPs ‘a book’ and ‘a novel’ are both Det N.

Mathematically, the copying operation is guaranteed to return the most probable parse because an incremental parser is guaranteed to know the most likely parse of the first conjunct by the time it reaches the coordinator.

The Baseline and Within models also require a change to the parser. In particular, they require a cache or history of recently used rules. This raises a dilemma: whenever a parsing error

occurs, the accuracy of the contextual history is compromised. However, the experimental items used were simple enough that no parsing errors occurred. Thus, it was always possible to fill the cache using rules from the best incremental parse so far.[2]

### *3.2. Study 3: Modeling Parallelism Experiments*

The purpose of this study is to evaluate the models described in the previous section by using them to simulate the results of a reading time experiment on syntactic parallelism. We will test the hypothesis that our models can correctly predict the pattern of results found in the experiment study. We will restrict ourselves to evaluating the qualitative pattern of results, rather than modeling the reading times directly.

Frazier et al. (2000) reported a series of experiments that examined the parallelism preference in reading. In their Experiment 3, they monitored subjects' eye-movements while they read sentences like (2):

(2) a. Terry wrote a long novel and a short poem during her sabbatical.

b. Terry wrote a novel and a short poem during her sabbatical.

They found that total reading times were faster on the phrase *a short poem* in (2-a), where the coordinated noun phrases are parallel in structure, compared with in (2-b), where they are not.

The probabilistic models presented here do not directly make predictions about total reading times as reported by Frazier et al.. Therefore, a *linking hypothesis* is required to link the predictions of the model (e.g., in the form of probabilities) to experimentally observed behavior (e.g., in the form of processing difficulty). The literature on probabilistic modeling contains a number of different linking hypotheses. For example, one possibility is to use an incremental parser with beam search (e.g., an *n*-best approach). Processing difficulty is predicted at points in the input string where the current best parse is replaced by an alternative derivation, and garden-pathing occurs when the ultimately correct parse has dropped out of the beam (Jurafsky, 1996; Crocker &

Brants, 2000). However, this approach is only suited to ambiguous structures.

An alternative is to keep track of all derivations, and predict difficulty if there is a change in the probability distributions computed by the parser. One way of conceptualizing this is Hale's (2001) notion of surprisal. Intuitively, surprisal measures the change in probability mass as structural predictions are disconfirmed when a new word is processed. If the new word disconfirms predictions with a large probability mass (high surprisal), then high processing complexity is predicted, corresponding to increased processing difficulty. If the new word only disconfirms predictions with a low probability mass (low surprisal), then we expect low processing complexity and reduced processing difficulty. Technically, the surprisal $S_k$ at input word $w_k$ corresponds to the difference between the logarithm of the prefix probabilities of word $w_{k-1}$ and $w_k$ (for a detailed derivation, see Levy, 2007):

$$S_k = \log P(w_1 \cdots w_k) - \log P(w_1 \cdots w_{k-1}) \tag{14}$$

The standard definition of surprisal given in Equation (14) is useful for investigating word-by-word reading time effects. In the present parallelism studies, however, we are interested in capturing reading time differences in regions containing several words. Therefore, we introduce a more general notion of surprisal computed over an $m$ word region spanning from $w_{k+1}$ to $w_{k+m+1}$:

$$S_{k\ldots k+m} = \log P(w_1 \cdots w_{k+m}) - \log P(w_1 \cdots w_{k-1}) \tag{15}$$

Subsequent uses of 'surprisal' will refer to this region-based surprisal quantity, and the term 'word surprisal' will be reserved for the traditional word-by-word measure. Both word surprisal and region surprisal have the useful property that they can be easily computed from the prefix probabilities returned by our parser.

In addition to surprisal, we also compute a simpler metric: we calculate the probability of the best parse of the whole sentence (Stolcke, 1995). Low probabilities are assumed to correspond to high processing difficulty, and high probabilities predict low processing difficulty. As we use

log-transformed sentence probabilities, this metric hypothesizes a log-linear relationship between model probability and processing difficulty.

*3.2.1. Method*

The item set we used for evaluation was adapted from that of Frazier et al. The original two relevant conditions of their experiment (see (2-a) and (2-b)) differ in terms of length. This results in a confound in the PCFG-based framework, because longer sentences tend to result in lower probabilities (as the parses tend to involve more rules). To control for such length differences, we adapted the materials by adding two extra conditions in which the relation between syntactic parallelism and length was reversed. This resulted in the following four conditions:

(3) a. Det Adj N and Det Adj N (parallel)

Terry wrote a long novel and a short poem during her sabbatical.

b. Det N and Det Adj N (non-parallel)

Terry wrote a novel and a short poem during her sabbatical.

c. Det Adj N and Det N (non-parallel)

Terry wrote a long novel and a poem during her sabbatical.

d. Det N and Det N (parallel)

Terry wrote a novel and a poem during her sabbatical.

In order to account for Frazier et al.'s parallelism effect a probabilistic model should predict a greater difference in probability between (3-a) and (3-b) than between (3-c) and (3-d) (i.e., for the reading times holds: (3-a)$-$(3-b) $>$ (3-c)$-$(3-d)). This effect will not be confounded with length, because the relation between length and parallelism is reversed between (3-a), (3-b) and (3-c), (3-d). In order to obtain a more extensive evaluation set for our models, we added eight items to the original Frazier et al. materials, resulting in a new set of 24 items similar to (3).[3]

The models we evaluated were the Baseline, the Within and the Copy models, trained as described in Section 3.1. We tested these three PCFG-based models on all 24 experimental sentences across four conditions. Each sentence was input as a sequence of correct part-of-speech tags, and the surprisal of the sentence as well as the probability of the best parse was computed.

Note that we do not attempt to predict reading time data directly. Rather, our model predictions are evaluated against reading times averaged over experimental conditions. This means that we predict qualitative patterns in the data, rather than obtaining a quantitative measure of model fit, such as $R^2$. Qualitative evaluation is standard practice in the psycholinguistic modeling literature. It is also important to note that reading times are contaminated by non-syntactic factors such as word length and word frequency (Rayner, 1998) that parsing models are not designed to account for.

### *3.2.2. Results and Discussion*

Table 7 shows the mean log surprisal values estimated by the models for the four experimental conditions, along with the differences between parallel and non-parallel conditions. Table 8 presents the mean log probabilities of the best parse in the same way.

The results indicate that both the Within and the Copy model predict a parallelism advantage. We used a Wilcoxon signed rank test to establish if the difference in surprisal and probability values were statistically different for the parallel and the non-parallel conditions, i.e., we compared the values for (3-a)−(3-b) with those for (3-c)−(3-d).[4] In the surprisal case, significant results were obtained for both the Within model ($N = 24$, $Z = 2.55$, $p < .01$, one-tailed) and the Copy model ($N = 24$, $Z = 3.87$, $p < .001$, one-tailed). Using the probability of the best parse, a statistically significant difference was again found for both the Within ($N = 24$, $Z = 1.67$, $p < .05$, one-tailed) and the Copy model ($N = 24$, $Z = 4.27$, $p < .001$, one-tailed).

The qualitative pattern of results is therefore the same for both models: the Within and the Copy model both predict that parallel structures are easier to process than non-parallel ones.

However, there are quantitative differences between the surprisal and the best-parse implementations of the models. In the surprisal case, the parallelism effect for the Within model is larger than the parallelism effect for the Copy model. This difference is significant ($N = 24$, $Z = 2.93$, $p < .01$, one-tailed). In the case of the best-parse implementation, we observe the opposite pattern: the Copy model predicts a significantly larger parallelism advantage than the Within model ($N = 24$, $Z = 4.27$, $p < .001$, one-tailed).

The Baseline model was not evaluated statistically, because by definition it predicts a constant value for (3-a)−(3-b) and (3-c)−(3-d) across all items (there are small differences due to floating point underflow). This is simply a consequence of the PCFG independence assumption, coupled with fact that the four conditions of each experimental item differ only in the occurrences of two NP rules.

Overall, the results show that the approach taken here can be successfully applied to model experimental data. Moreover, the effect is robust to parameter changes: we found a significant parallelism effect for both the Within and the Copy model, in both the surprisal and the best-parse implementation. It is perhaps not surprising that the Copy model shows a parallelism advantage for the Frazier et al. (2000) items, as this model was explicitly designed to prefer structurally parallel conjuncts. The more interesting result is the parallelism effect we found for the Within model, which shows that such an effect can arise from a more general probabilistic priming mechanism. We also found that the surprisal implementation of the Within model predicts a larger parallelism effect than the best-parse implementation (relative to the Copy model, see Tables 7 and 8). This indicates that using all available parses (as in the surprisal case) amplifies the effect of syntactic repetition, perhaps because it takes into account repetition effects in all possible syntactic structures that the parser considers, rather than only in the most probable structure (as in the best parse implementation).

In spite of this difference in effect size, we can conclude, however, that best-parse probabilities are a good approximation of surprisal values for the sentences under consideration,

while being much simpler to compute. We therefore focus on best-parse probabilities for the remainder of the paper.

## 4. Modeling Priming and Decay

### *4.1. A Model of Priming inspired by ACT-R*

The Within and Between priming models introduced in Sections 2 and 3 make the assumption that priming is based upon a binary decision: a target item is either primed or not. Moreover, the model 'forgets' what was primed once it finishes processing the target region, i.e., no learning occurs. As we saw in Section 3.2, these modeling assumptions were sufficient to model a set of standard experimental data. However, Section 2 showed that the parallelism effect applies to a range of different levels of granularity. Using a single binary decision makes it impossible to build a model which can simultaneously account for priming at multiple levels of granularity. In this section, we introduce the Decay model, which is able to account for priming without making an arbitrary choice about the size of the priming region.

The structure of this model is inspired by two observations. First, we found in Section 2 that priming effects were smaller as the size of the priming region increased (from coordinate structures to arbitrary structures within sentences, to arbitrary structures between sentences). Second, a number of authors (e.g., Gries, 2005; Szmrecsanyi, 2005; Reitter et al., 2006) found in corpus studies that the strength of priming decays over time (but not Jaeger, 2006b and ?, ?, who controlled for speaker differences). Intuitively, these two effects are related: by selecting a larger priming region, we effectively increase the time between the onset of the prime and onset of the target. Therefore, by accounting for decay effects, it may be possible to remove the arbitrary choice of the size of the priming region from our model. Given that we are assuming that priming is due to a general cognitive mechanism, it is a logical next step to model decay effects using a general, integrated cognitive architecture. We will therefore build on concepts from the ACT-R framework, which has been successfully used to account for a wide range of experimental data on human

cognition (Anderson, 1983; Anderson et al., 2004).

The ACT-R system has two main elements: a planner and a model of memory which places a cost on accesses to declarative memory (where declarative facts, also called 'chunks', are stored) and on procedural memory (which contains information on how to carry out planning actions). The Decay model uses the ACT-R memory system to store grammar rules but we eschew the planner, instead continuing to use the incremental Earley parser of previous experiments.

Following earlier work on parsing with ACT-R (Lewis & Vasishth, 2005), we assume that grammar rules are stored in procedural memory. Lewis and Vasishth (2005) fully commit to the ACT-R architecture, implementing their incremental parser in the ACT-R planning system and therefore storing partially constructed parses in declarative memory. Instead, we will assume that the Decay model uses the same underlying chart parser presented in Section 3.1.5 instead of a planning system. This means we make no particular claims about the memory cost of accessing partially constructed parses.[5]

A goal of the full ACT-R architecture is to model the time course of cognitive behavior. Our restricted ACT-R-inspired simulation, though, is limited to modeling the *probability* of certain memory accesses, and, via our linking hypothesis, we only make qualitative predictions about processing difficulty. This restriction is motivated by the desire to maintain the underlying architecture and approach developed for the Within and Copy models, which was successfully evaluated in the previous section.

In ACT-R, the probability of memory access depends upon recency information both for declarative chunks, and, following the work of Lovett (1998), for production rules. The probability of a production rule (or more specifically, the probability of successfully applying a production rule) depends on the number of past successes and failures of that rule. Lovett (1998) argues that the times of these successes and failures should be taken into account. For our ACT-R-inspired Decay model, a 'success' will be a successful application of a rule $N \rightarrow \zeta$, and a 'failure' will be an application of any other rule with the same left-hand side $N$. As we will see below, this choice

of success or failure makes the Decay model a simple generalization of a PCFG.

Successes are counted in a similar manner to primed rules in the Between model: after each word, we compute the most probable parse, and compute the set of rules used in this parse. Any new rules (compared to the set from the previous word) is considered to be a success at this word.[6]

ACT-R assumes that the probability of declarative chunks is based upon their activation, and that this activation directly influences the time taken to retrieve chunks from long-term memory into buffers, where they can be acted upon. There is no such direct influence of the probability of success on the time to retrieve or act on a production rule. However, there is a clear, if indirect influence on time: if the rule required for producing a parse has a high probability, the expected number of incorrect rules attempted will be low, leading to the prediction of lower processing difficulty; on the other hand, if the required rule has a low probability, the expected number of incorrect rules attempted will be high, leading to a prediction of high processing difficulty. This is no different than the intuition behind the linking hypothesis of Section 3.2, where we assumed that low probability (or high surprisal) corresponds to high processing difficult. The novelty of the Decay model is that the time sensitivity of production rules now affects the calculated probabilities.

#### *4.1.1. Parametrizing the Model*

Recall that in a probabilistic context-free grammar the probability of a grammar rule $N \rightarrow \zeta$ is estimated as follows:

$$\hat{P}(\zeta|N) = \frac{c(N \rightarrow \zeta)}{c(N)} \tag{16}$$

As before, the function $c(\cdot)$ counts the number of times the event occurs. The Lovett model postulates that the probability of a certain production rule being picked is:

$$\hat{P} = \frac{\sum_{i \in Successes} t_i^d}{\sum_{i \in Successes} t_i^d + \sum_{j \in Failures} t_j^d} \tag{17}$$

Here, $d$ is the decay constant, *Successes* is the set of successful rule applications, *Failures* is the set of unsuccessful rule applications, and $t_i$ is time at which rule application $i$ occurred. As described

below, this time parameter can estimated from a corpus. Following convention in the ACT-R literature (see Lewis & Vasishth, 2005), we set $d$ to 0.5.

As noted above, the model uses a notion of success and failure appropriate for use with a probabilistic context-free grammar: a success is an application of a rule $N \rightarrow \zeta$, and a failure is all other rule applications with the same left-hand side $N$. The standard ACT-R model defines successes and failures in terms of a higher-level task. In our case, the task is finding the best way to rewrite an $N$ to get a correct parse. Using this choice for success and failure, we can rewrite the probability of a rule as:

$$\hat{P}(\zeta|N) = \frac{\sum_{i \in Successes} t_i^d}{\sum_{i \in Successes} t_i^d + \sum_{j \in Failures} t_j^d} \tag{18}$$

For example, if the rule NP $\rightarrow$ Det Adj N is used to parse the tag sequence Det Adj N, then this rule will get a success at time $t_i$ while all other NP rules get a failure. Notice that if the parameter $d$ is set to 0, all the exponentiated time parameters $t_i^d$ are set to 1, giving the maximum likelihood estimator for a PCFG in equation (16), which is what we used as our Baseline model. In other words, our Decay model has a standard PCFG as a special case.

Standard ACT-R uses the activation of procedural rules as an intermediate step toward calculating time course information. However, the model presented here does not make any time course predictions. This choice was made due to our focus on syntactic processing behavior: obviously, time is also spent doing semantic, pragmatic and discourse inferences, which we do not attempt to model. Although this simplifies the model, it does pose a problem. One of the model parameters, $t_i$, is expressed in units of time, and cannot be observed directly in the corpus. To overcome this difficulty, we assume each word uniformly takes 500 ms to read. This is meant as an approximation of the average total reading time of a word.[7] Because the previous occurrence of a constituent can be several sentences away, we expect that local inaccuracies will average out over a sufficiently large training corpus.

### *4.2. Study 4: Modeling Parallelism Experiments using the Decay Model*

The purpose of this experiment is to evaluate the ACT-R-inspired Decay model described in the previous section. Our hypothesis is that adding decay to the model, while increasing its cognitive realism, does not impair the model's ability to predict the pattern of results in experiments of syntactic parallelism.

#### *4.2.1. Method*

As in Study 3, we estimated the model probabilities using the WSJ corpus. Similar to the Within model, parameter estimation requires traversing the rules in the same order the parser does, here to get accurate statistics for the time parameter.

#### *4.2.2. Results and Discussion*

Following the method of Study 3, we test the model on the extended set of experimental stimuli based on Frazier et al. (2000). As described in Section 3.2, we use 24 items in four conditions, and compute the probability of the best parse for the sentences in each conditions (see (3) in Section 3.2 for example sentences). The hypothesis of interest is again that the difference between (3-a) and (3-b) is greater than the difference between (3-c) and (3-d).

The results for the Decay model are shown in Table 9, with the results of the Baseline, Within, and Copy model from Study 3 as comparison. We find that the Decay model gives a significant parallelism effect using the Wilcoxon signed rank test ($N = 24$, $Z = 4.27$, $p < .001$, one-tailed). Like the Within model, the effect size is quite small, as a general mechanism is used to predict the parallelism effect, rather than a specialized one as in the Copy model.

However, the Within model and the Decay model do not make identical predictions. The differences between the two models become clear upon closer examination of the experimental items. Some of the materials have a Det N as the subject, for instance *The nurse checked a young woman and a patient before going home*. In this example, the Within model will predict a speedup

for *a patient* even though Frazier et al. (2000) would not consider it to be a parallel sentence. In such cases, the Decay model predicts some facilitation at the target NP, but the effect is weaker because of the greater distance from the target to the subject NP. This example illustrates how the Decay model benefits from the fact that it incorporates decay and therefore captures the granularity of the priming effect more accurately. This contrasts with the coarse-grained binary primed/unprimed distinction made by the Within model. We will return to this observation in the next section.

Moreover, the Decay model is more cognitively plausible than the Within model because it is grounded in research on cognitive architectures: we were able to re-use model parameters proposed in the ACT-R literature (such as the decay parameter) without resorting to stipulation or parameter tuning.

### *4.3. Study 5: The Parallelism Effect in German Sentence Coordination*

Sections 3.2 and 4.1 introduced models which were able to simulate Experiment 3 of Frazier et al. (2000). The Frazier et al. experimental items are limited to English noun phrases. This raises the question whether our models generalize to other constructions and to other languages. The purpose of the present study is to address this questions by modeling additional experimental data, viz., Knöferle and Crocker's (2006) experiment on parallelism in German sentence coordination. Our hypothesis is that both the Copy model and the Decay model will be able to account for the German data.

Knöferle and Crocker's (2006) items take advantage of German word order. Declarative sentences normally have an subject-verb-object (SVO) order, such as (4-a). A temporal modifier can appear before the object, as illustrated in (4-b). However, word order is flexible in German. The temporal modifier may be focused by bringing it to the front of the sentence, a process known as topicalization. The topicalized version of the last sentence is (4-c). We will refer to this as an VSO (verb-subject-object) or *subject-first* order. A more marked word order would be topicalized

verb-object-subject (VOS), as in (4-d). We will refer to such sentences as VOS or *object-first*.

(4) a. Der Geiger lobte den Sänger.
'The violinist complimented the singer.'

b. Der Geiger lobte vor ein paar Minuten den Sänger.
'The violinist complimented the singer several minutes ago.'

c. Vor ein paar Minuten lobte der Geiger den Sänger.
'Several minutes ago, the violinist complimented the singer.'

d. Vor ein paar Minuten lobte den Sänger der Geiger.
'Several minutes ago, it was the singer that the violinist complimented.'

In the experiment of Knöferle and Crocker, each item contains two coordinated sentences, each of which is either subject-first or object-first. The experiment uses a $2 \times 2$ design: either subject-first or object-first in the first conjunct with either subject-first or object-first in the second conjunct. This leads to two parallel and two non-parallel conditions, as shown in (5) below. (In reality, Knöferle and Crocker's (2006) items contain a spillover region which we removed, as explained below.)

(5) a. Vor ein paar Minuten lobte der Geiger den Sänger und in diesem Augenblick preist der Trommler den Dichter.
'Several minutes ago, the violinist complimented the singer and at this moment the drummer is praising the poet.'

b. Vor ein paar Minuten lobte den Sänger der Geiger und in diesem Augenblick preist der Trommler den Dichter.
'Several minutes ago, it was the singer that the violinist complimented and at this moment the drummer is praising the poet.'

c. Vor ein paar Minuten lobte der Geiger den Sänger und in diesem Augenblick preist

den Dichter der Trommler.

'Several minutes ago, the violinist complimented the singer and at this moment it is poet that the drummer is praising.'

d. Vor ein paar Minuten lobte den Sänger der Geiger und in diesem Augenblick preist den Dichter der Trommler.

'Several minutes ago, it was the singer that the violinist complimented and at this moment it is the poet the drummer is praising.'

The object-first condition is rare and generally considered marked in the psycholinguistic literature on German. We will therefore refer to conditions (5-b) and (5-d), where the object-first clause is in the second conjunct, as the marked condition. The alternative conditions, (5-a) and (5-c), are henceforth referred to as unmarked. We refer to conditions (5-a) and (5-d), which have the same order of subject and object as the parallel conditions. Likewise, (5-b) and (5-c), which have a different order, are the non-parallel conditions. Knöferle and Crocker (2006) found that, overall, the unmarked (subject-first) conjunct was faster to read, but this markedness effect was dominated by a parallelism effect. In other words, the marked parallel conditions had a lower reading time than the marked non-parallel condition. In the following we will investigate if our models are able to replicate this result.

*4.3.1. Method*

This study was largely set up in a manner similar to Studies 3 and 4 on English data. As the present study aims to analyze German data, it was necessary to train the parser on German text. Therefore, the Tiger corpus (Brants, Dipper, Hansen, Lezius, & Smith, 2002) of German newspaper text was used in lieu of the the Wall Street Journal corpus used in earlier studies. Of the models which have been presented in earlier studies, only the Copy and Decay models are used here (there is no need to test the Within model as it is subsumed by the Decay model).

Modeling the data of Knöferle and Crocker (2006) poses a challenge to the computational

models, for several reasons. First, as noted above, these data contain a spillover region. This region is problematic as it creates an attachment ambiguity: it may attach low to the second conjunct or high to the main clause. While the ambiguity apparently causes few problems for human subjects, it proves difficult for an automatic parser to analyze unambiguously. We therefore decided to remove the spillover region from the original items for the present modeling study. As we emphasized in Section 3.2, the purpose of our modeling studies is to account for qualitative patterns in the data, rather than modeling individual reading times. Therefore, we have no way of predicting spillover effects using the current approach.

An additional challenge posed to the parser is that the verbs in the experimental items are ditransitive, and the accusative and nominative objects need to be disambiguated by their articles. Reading a grammar directly from the Tiger treebank does not encode this information, but both subcategorization information (Baldewein & Keller, 2004) and case information (Dubey, 2004) can be encoded in the grammar by way of treebank transformations. These unambiguous transformations create a grammar in which verb and NP nodes are enriched with the relevant information, and are exemplified in Figure 9 and Figure 10, respectively.

A third transformation, based on German topological field theory, is also necessary due to the flat annotation style of the Tiger corpus. We explicitly add a *Vorfeld* (first position) phrase to the grammar, as shown in Figure 11. This corresponds to the part of a sentence preceding the verb. In an untopicalized declarative sentence, the subject is in the Vorfeld. In topicalized sentences, such as the Knöferle and Crocker (2006) items, the Vorfeld contains the topic. This transformation is necessary because the annotation style of the Tiger corpus would otherwise make it difficult to model any kind of word-order priming. We have hypothesized priming occurs on the level of syntactic rules, and the Tiger corpus uses a flat annotation style for sentences. So, if the topic in the prime and target conjuncts are of different grammatical categories, a naive model would predict no priming. This is averted by putting the topic in a category of its own, which is an uncontroversial assumption not only in topological field theory (used here), but also in X-bar theory, which would

posit a covert complementizer whose specifier contains the topic.

The choice of applying or not applying this third transformation corresponds to an instance of the Grain problem (Mitchell et al., 1996) with two different choices of grain size for estimating syntactic frequencies. Without the transformation, we are making the assumption that both conjuncts must be equivalent on the coarse-grained sentential level hypothesized by the Tiger annotators. By applying the transformation, we allow a finer-grained parallelism effect which is somewhat more independent of particular annotation strategies.

The Grain problem also appears in Frazier and Clifton's (2001) Copy hypothesis. What exactly is meant by 'copying'? As Frazier and Clifton do not claim otherwise, we have assumed that the entire structure is copied. It is possible that Frazier et al. did intend for their Copy hypothesis to operate at a more fine-grained level, but they did not specify how this might be done. Therefore, we do not make any modification to the Copy model for this experiment (of course, there is also no change to the Decay model, either).

Just as in the modeling studies presented on the Frazier et al. (2000) items, we measure the probability difference between pairs of parallel and non-parallel conditions. In this case, we measure (5-a)−(5-b) compared to (5-c)−(5-d). If there is no statistical difference between the quantities, we conclude there is no parallelism effect. On the other hand, we may conclude there is a parallelism effect if the former is greater than the latter.

### *4.3.2. Results and Discussion*

The results are shown in Table 10. All three models were able to parse the experimental materials unambiguously. We performed a Wilcoxon signed rank test on the difference between (5-a)−(5-b) and (5-c)−(5-d). We found no significant difference between the two conditions for the Copy model. However, the Decay model gave a statistically significant result ($N = 32$, $Z = 1.72$, $p < .05$).

This finding provides support for the hypothesis that the Decay model generalizes to other

structural configurations and to a new language. An interesting fact about Knöferle and Crocker's (2006) materials is that the marked VOS word order occurs quite infrequently in the Tiger corpus, and never occurs twice in the same sentence (coordinated or not). This fact does not affect the Copy or Decay models: the Copy operation does not inspect the rules other than to check that they are identical, and the Decay model dynamically updates rules probabilities depending on context. However, the Within model strongly depends upon observing particular rules repeatedly, and therefore it would fail to deliver any parse at all if the rules have never been observed repeating.

Moreover, the results here show that our straightforward implementation of the Copy model does not easily generalize beyond NP experiments. A key problem is that Knöferle and Crocker's parallel condition contained sentences which had marginally different structures: while the NP word orders were parallel, each conjunct contained a modifier which in many cases did not have parallel structure. The non-parallel modifiers resulted in the Copy model ignoring such conjuncts as candidates to be copied. A more precise statement of the Copy hypothesis is required to model clause-level parallelism in general. In particular, the Copy model faces a grain problem, which will need to be addressed in future work.

## 5. General Discussion

We began this paper by showing how *adaptation probabilities* can be defined as a measure of structural repetition. Using multinomial logistic regression, we demonstrated that there is a robust, pervasive effect of parallelism for a variety of noun phrase types. We found this tendency for structural repetition in two different corpora of written English. We showed that the effect occurs in a number of contexts: coordinate NPs (Study 1), non-coordinate NPs within the same sentence (Study 2), and NPs in two adjacent sentences (Study 2). We were also able to show that the parallelism effect persists in complex noun phrases (those containing a PP or an SBAR), even if the preference for short phrases to precede long ones is taken into account. Taken together, the findings of Studies 1 and 2 strongly suggest that the parallelism effect is an instance of a general

processing mechanism, such as syntactic priming (Bock, 1986), rather than specific to coordination, as suggested by (Frazier et al., 2000). Frazier et al. (2000) base their claim on the failure to find a parallelism effect between the subject and the object NP in the same sentence. This is not sufficient to argue against a priming explanation for the parallelism effect, as our results for within and between sentence priming show.

We also observed marked differences in the effect sizes of Study 1 and Study 2: we found that the parallelism effect is strongest for coordinate structures, and weaker for non-coordinate structures within the same sentence and in adjacent sentences. There are a number of possible explanations for this difference. Priming has been argued to be subject to distance-based decay (e.g., Gries, 2005; Szmrecsanyi, 2005; Reitter et al., 2006). This may be a relevant factor as prime and target are relatively close together in coordination (only separated by one word), while the mean distance between prime and target is larger for priming in non-coordinate structures within the same sentence, and even larger for priming between sentences.

Previous experimental work has found parallelism effects only in comprehension data. The present work demonstrates that parallelism effects also occur in production data, replicating the results of previous corpus studies (Levy, 2002; Cooper & Hale, 2005). This raises the interesting question of the relationship between the two data types. It has been hypothesized that the human language processing system is tuned to mirror the probability distributions in its environment, including the probabilities of syntactic structures (Mitchell et al., 1996). If this *tuning hypothesis* is correct, then the parallelism effect in comprehension data can be explained as an adaptation of the human parser to the prevalence of parallel structures in its environment (as approximated by corpus data), as found in the present set of studies.

Note that the results in this paper not only have an impact on theoretical issues regarding human sentence processing, but also on engineering problems in natural language processing, such as probabilistic parsing. To avoid sparse data problems, probabilistic parsing models make strong independence assumptions; in particular, they generally assume that sentences are independent of

each other. This is partly due to the fact it is difficult to parametrize the many possible dependencies which may occur between adjacent sentences. However, in this paper, we show that structure re-use is one possible way in which the independence assumption is broken. A simple and principled approach to handling structure re-use is to use adaptation probabilities for probabilistic grammar rules, analogous to cache probabilities used in caching language models (?, ?; Kuhn & Mori, 1990), which is what we proposed in this paper.

The use of adaptation probabilities leads directly to the second contribution of this paper, which is to show that an incremental parser can simulate syntactic parallelism effects in human parsing by incorporating a probabilistic account of rule re-use. Frazier et al. (2000) argued that the best account of the parallelism advantage was a model in which parallelism is limited to particular structural configurations such as coordination. To test this hypothesis, we explored a probabilistic variant of Frazier and Clifton's (2001) *copy*-$\alpha$ mechanism, along with two more general models based on within- and between-sentence priming. Although the copy mechanism provided a stronger parallelism effect when we used it to simulate the patterns in the human reading time data, the effect was also successfully simulated by a general within-sentence priming model. On the basis of Occam's razor, we therefore argue that it is preferable to assume a simpler and more general mechanism, and that the copy mechanism is not needed. We explored also an alternative implementation of our models which uses Hale's (2001) surprisal to predict processing difficulty. We found that the parallelism effect can be captured both by the surprisal implementation and by a more straightforward implementation that uses the probability of the best parse as a measure of processing difficulty.

All the models we proposed offer a broad-coverage account of human parsing, not just a limited model of a hand-selected set of examples. This is in line with recent developments in the literature on probabilistic models of human language processing, which has seen a shift of focus away from construction-specific models to broad-coverage models (Crocker & Brants, 2000; Hale, 2001; Padó, Keller, & Crocker, 2006; Padó, Crocker, & Keller, 2006).

The third and final contribution of the present paper is the development an ACT-R-inspired Decay model of syntactic priming. This model is based on the observation in the literature that the strength of the priming effect shows an exponential decay with the temporal distance between the prime and the target. The Decay model of priming incorporates a decay of rule probabilities inspired by ACT-R's model of procedural memory, and is able to offer a more realistic account of priming that should be able to cover a wider range of parallelism phenomena. We validated this by training the Decay model on a different language (German) and testing it on a new data set that includes sentential coordination rather than NP coordination. We also found that the Copy model in its current form is not able to account for the German parallelism data.

In the research reported in this paper, we have adopted a simple model based on an unlexicalized PCFG. In future research, we intend to explore the consequences of introducing lexicalization into the parser. This is particularly interesting from the point of view of psycholinguistic modeling, because there are well known interactions between lexical repetition and syntactic priming, which require lexicalization for a proper treatment. Another area for future work is the implementation of a more cognitively realistic version of our model that predicts reading times directly, e.g., by making us of ACT-R real-time capabilities. Such a model could then also be applied to reading time data for domains other than NP parallelism.

## References

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of mind. *Psychological Review*, *111*(4), 1036–1060.

Apel, J., Knöferle, P., & Crocker, M. W. (2007). Processing parallel structure: Evidence from eye-tracking and a computational model. In S. Vosniadou, D. Kayser, & A. Protopapas (Eds.), *Proceedings of the European Cognitive Science Conference 2007* (pp. 125–131). London: Taylor and Francis.

Baldewein, U., & Keller, F. (2004). Modeling attachment decisions with a probabilistic parser: The case of head final structures. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 73–78). Chicago: Cognitive Science Society.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*, 355–387.

Branigan, H. P., Pickering, M. J., Liversedge, S. P., Stewart, A. J., & Urbach, T. P. (1995). Syntactic priming: Investigating the mental representation of language. *Journal of Psycholinguistic Research*, *24*, 489–506.

Branigan, H. P., Pickering, M. J., & McLean, J. F. (2005). Priming prepositional-phrase attachment during comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*(3), 468–481.

Brants, S., Dipper, S., Hansen, S., Lezius, W., & Smith, G. (2002). The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories.* Sozopol: Bulgarian Academy of Sciences.

Carlson, K. (2002). The effects of parallelism and prosody on the processing of gapping structures. *Language and Speech*, *44*(1), 1–26.

Church, K. W. (2000, August). Empirical estimates of adaptation: the chance of two Noriegas is

closer to $p/2$ than $p^2$. In *Proceedings of the 17th Conference on Computational Linguistics* (pp. 180–186). Saarbrücken, Germany: Association for Computational Linguistics.

Cooper, A. A., & Hale, J. T. (2005). Promotion of disfluency in syntactic parallelism. In *Proceedings of the workshop on disfluency in spontaneous speech* (pp. 59–63). Aix-en-Provence: International Speech Communication Association.

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

Dubey, A. (2004). *Statistical parsing for German: Modeling syntactic properties and annotation differences*. Unpublished doctoral dissertation, Saarland University.

Dubey, A., Keller, F., & Sturt, P. (2006). Integrating syntactic priming into an incremental probabilistic parser, with an application to psycholinguistic modeling. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (pp. 417–424). Sydney: Association for Computational Linguistics.

Dubey, A., Sturt, P., & Keller, F. (2005). Parallelism in coordination as an instance of syntactic priming: Evidence from corpus-based modeling. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing* (pp. 827–834). Vancouver: Association for Computational Linguistics.

Frazier, L., & Clifton, C. (2001). Parsing coordinates and ellipsis: Copy α. *Syntax*, *4*(1), 1–22.

Frazier, L., Munn, A., & Clifton, C. (2000). Processing coordinate structures. *Journal of Psycholinguistic Research*, *29*(4), 343–370.

Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, *35*.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics.* Pittsburgh, PA: Association for Computational Linguistics.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Jaeger, T. F. (2006a). *Redundancy and syntactic reduction in spontaneous speech*. Unpublished doctoral dissertation, Stanford University.

Jaeger, T. F. (2006b). *Syntactic persistence in real life (spontaneous speech).* Talk at the 19th Annual CUNY Conference on Human Sentence Processing, New York.

Jaeger, T. F., & Snider, N. (2007). *Implicit learning and syntactic persistence: Surprisal and cumulativity.* Talk at the 20th Annual CUNY Conference on Human Sentence Processing, San Diego.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137–194.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, *45*, 153–168.

Knöferle, P., & Crocker, M. (2006). *Clause-level and constituent-level parallelism: eye-tracking coordinate structures in German.* Presented at the 12th Conference on Architectures and Mechanisms for Language Processing, Edinburgh. Nijmegen, Netherlands.

Kuhn, R., & Mori, R. de. (1990). A cache-based natural language model for speech recognition. *IEEE Transanctions on Pattern Analysis and Machine Intelligence*, *12*(6), 570–583.

Ledoux, K., Traxler, M. J., & Swaab, T. Y. (2007). Syntactic priming in comprehension. *Psychological Science*, *18*(2).

Levy, R. (2002). *The statistical distribution of english coordinate noun phrases: Parallelism and weight effects.* Talk at the 31st Conference on New Ways of Analyzing Variation, Stanford.

Levy, R. (2007). Expectation-based syntactic comprehension. *Cognition, to appear*.

Lewis, R. L., & Vasishth, S. (2005, May). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 1–45.

Lovett, M. C. (1998). The atomic components of thought. In J. R. Anderson & C. Lebiere (Eds.),

(pp. 255–296). Mahwah, NJ: Erlbaum.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*, 676–703.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., et al. (1994). The Penn Treebank: Annotating predicate-argument structure. In *ARPA human language technology workshop*.

Mitchell, D. C., Cuetos, F., Corley, M. M. B., & Brysbaert, M. (1996). Exposure-based models of human parsing: Evidence for the use of coarse-grained (non-lexical) statistical records. *Journal of Psycholinguistic Research, 24*(6), 469–488.

Padó, U., Crocker, M., & Keller, F. (2006). Combining syntax and thematic fit in a probabilistic model of sentence processing. In R. Sun (Ed.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 657–662). Vancouver: Cognitive Science Society.

Padó, U., Keller, F., & Crocker, M. (2006). Modeling semantic role plausibility in human sentence processing. In *Proceedings of the 11th conference of the european chapter of the association for computational linguistics* (pp. 345–352). Trento: Association for Computational Linguistics.

Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language, 39*(4), 633–651.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1988). *Numerical recipes in C*. Cambridge, UK: Cambridge University Press.

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

Reali, F., & Christiansen, M. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language, 57*, 1–23.

Reitter, D., Moore, J., & Keller, F. (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In R. Sun (Ed.), *Proceedings of the 28th Annual Conference of*

*the Cognitive Science Society* (pp. 685–690). Vancouver: Cognitive Science Society.

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

Sun, R. (Ed.). (2006). *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Vancouver: Cognitive Science Society.

Szmrecsanyi, B. (2005). Creatures of habit: A corpus-linguistic analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory*, *1*(1), 113-149.

## Author Note

The work reported here was supported by Leverhulme Trust research project grant. We would like to thank Matthew Crocker, Pia Knoeferle, Roger Levy, Martin Pickering, and two anonymous reviewers for comments and feedback on this paper. A preliminary version of Studies 1–3 was published previously in Dubey, Sturt, and Keller (2005) and Dubey, Keller, and Sturt (2006).

## Footnotes

[1]Our terminology differs from that of Church, who uses 'history' to describe the first half, and 'test' to describe the second. Our terms avoid the ambiguity of the phrase 'test set' and coincide with the common usage in the psycholinguistic literature.

[2]In unrestricted text, where parsing errors are more common, alternative strategies are required. One possibility is to use an accurate lexicalized parser. A second possibility arises if an 'oracle' can suggest the correct parse in advance: then the cache may be filled with the correct rules suggested by the oracle.

[3]To ensure the new materials and conditions did not alter the parallelism effect, we carried out a preliminary eye-tracking study based on an identical design to the modeling study (see (3)), with 36 participants. The interaction predicted by parallelism ((3-a)−(3-b) > (3-c)−(3-d)) was obtained in probability of regression from the region immediately following the second conjunct (*during her sabbatical*) and also in second-pass reading times on a region consisting of *and* followed by the second conjunct (e.g., *and a short poem*).

[4]The Wilcoxon signed rank test can be thought of as a non-parametric version to the paired *t*-test for repeated measurements on a single sample. The test was used because it does not require strong assumptions about the distribution of the log probabilities.

[5]This assumption is reasonable on our low-memory load items, but may be untenable for parsing high-memory load constructions such as object relative clauses.

[6]If a structure is ambiguous and requires reanalysis, we make the assumption that the initial incorrect analysis acts as a prime. However, because the items of the parallelism experiment are unambiguous, this assumption has no effect for our data set.

[7]This value is arbitrary, but could be made precise using eye-tracking corpora which provide estimates for word reading times in continuous text (e.g., Kennedy & Pynte, 2005). Any constant value would produce the same modeling result. In general, frequency and length effects on reading time are well documented (Rayner, 1998) and could be incorporated into the model.

| Predictor | $\chi^2$ | df | Probability |
|---|---|---|---|
| Adaptation | 4055 | 25 | $p < 0.0001$ |
| Short-before-long | 2726 | 5 | $p < 0.0001$ |
| Interaction | 201 | 25 | $p < 0.0001$ |

Table 1

*Summary of the log-likelihood $\chi^2$ statistics for the predictors in the multinomial regression for coordinate structures in the Brown corpus*

| Predictor | $\chi^2$ | df | Probability |
|---|---|---|---|
| Adaptation | 1686 | 25 | $p < 0.0001$ |
| Short-before-long | 1002 | 5 | $p < 0.0001$ |
| Interaction | 146 | 25 | $p < 0.0001$ |

Table 2

*Summary of the log-likelihood $\chi^2$ statistics for the predictors in the multinomial regression for coordinate structures in the WSJ corpus*

| Predictor | $\chi^2$ | df | Probability |
|---|---|---|---|
| Adaptation | 13782 | 25 | $p < 0.0001$ |
| Short-before-long | 239732 | 5 | $p < 0.0001$ |
| Interaction | 49226 | 25 | $p < 0.0001$ |

Table 3

*Summary of the log-likelihood $\chi^2$ statistics for the predictors in the multinomial regression for within-sentence priming in the Brown corpus*

| Predictor | $\chi^2$ | df | Probability |
|---|---|---|---|
| Adaptation | 37612 | 25 | $p < 0.0001$ |
| Short-before-long | 447652 | 5 | $p < 0.0001$ |
| Interaction | 64112 | 25 | $p < 0.0001$ |

Table 4

*Summary of the log-likelihood $\chi^2$ statistics for the predictors in the multinomial regression for within-sentence priming in the WSJ corpus*

| Predictor | $\chi^2$ | df | Probability |
|---|---|---|---|
| Adaptation | 21952 | 25 | $p < 0.0001$ |
| Short-before-long | 699252 | 5 | $p < 0.0001$ |
| Interaction | 55943 | 25 | $p < 0.0001$ |

Table 5

*Summary of the log-likelihood $\chi^2$ statistics for the predictors in the multinomial regression for between-sentence priming in the Brown corpus*

| Predictor | $\chi^2$ | df | Probability |
|---|---|---|---|
| Adaptation | 49657 | 25 | $p < 0.0001$ |
| Short-before-long | 918087 | 5 | $p < 0.0001$ |
| Interaction | 86643 | 25 | $p < 0.0001$ |

Table 6

*Summary of the log-likelihood $\chi^2$ statistics for the predictors in the multinomial regression for between-sentence priming in the WSJ corpus*

| Model | para: (3-a) | non-para: (3-b) | non-para: (3-c) | para: (3-d) | (3-a)−(3-b) | (3-c)−(3-d) |
|---|---|---|---|---|---|---|
| Baseline | −0.34 | −0.48 | −0.62 | −0.74 | 0.14 | 0.12 |
| Within | −0.30 | −0.51 | −0.53 | −0.30 | 0.20 | −0.20 |
| Copy | −0.52 | −0.83 | −0.27 | −0.31 | 0.32 | 0.04 |

Table 7

*Mean log surprisal values for items based on Frazier et al. (2000)*

| Model | para: (3-a) | non-para: (3-b) | non-para: (3-c) | para: (3-d) | (3-a)−(3-b) | (3-c)−(3-d) |
|---|---|---|---|---|---|---|
| Baseline | −33.47 | −32.37 | −32.37 | −31.27 | −1.10 | −1.10 |
| Within | −33.28 | −31.67 | −31.70 | −29.92 | −1.61 | −1.78 |
| Copy | −16.18 | −27.22 | −26.91 | −15.87 | 11.04 | −11.04 |

Table 8

*Mean log probability values for the best parse for items based on Frazier et al. (2000)*

| Model | para: (3-a) | non-para: (3-b) | non-para: (3-c) | para: (3-d) | (3-a)−(3-b) | (3-c)−(3-d) |
|---|---|---|---|---|---|---|
| Baseline | −33.47 | −32.37 | −32.37 | −31.27 | −1.10 | −1.10 |
| Within | −33.28 | −31.67 | −31.70 | −29.92 | −1.61 | −1.78 |
| Copy | −16.18 | −27.22 | −26.91 | −15.87 | 11.04 | −11.04 |
| Decay | −39.27 | −38.14 | −38.02 | −36.86 | −1.13 | −1.16 |

Table 9

*Mean log probability estimates for the best parse for items based on Frazier et al (2000)*

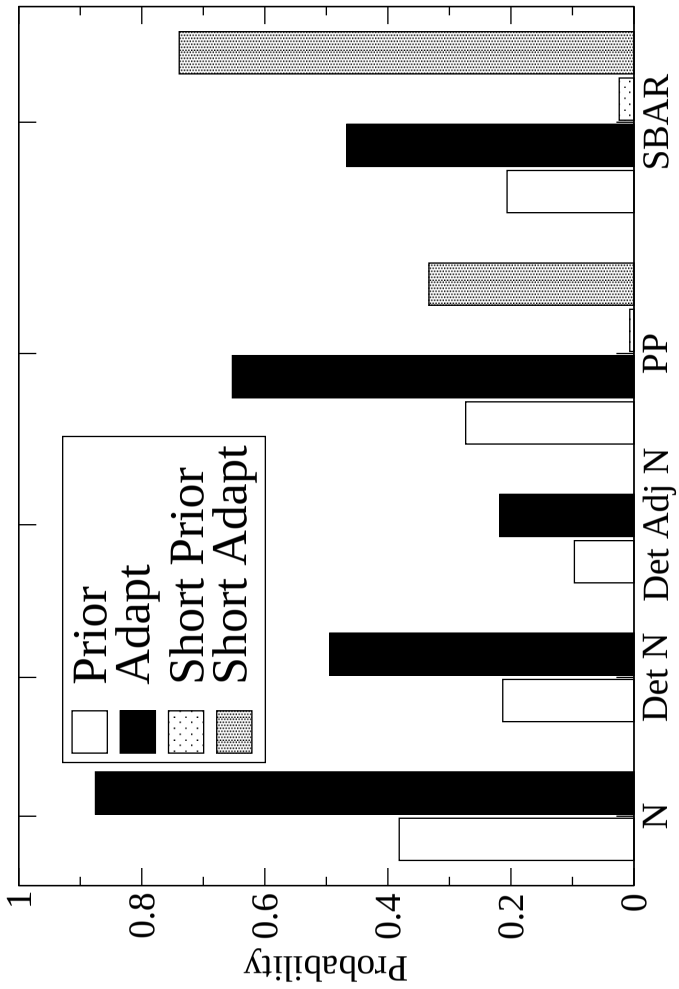
| Model | para: (5-a) | non-para: (5-b) | non-para: (5-c) | para: (5-d) | (5-a)−(5-b) | (5-c)−(5-d) |
|---|---|---|---|---|---|---|
| Baseline | −41.87 | −43.00 | −42.76 | −43.94 | 1.14 | 1.18 |
| Copy | −41.85 | −42.99 | −42.74 | −43.92 | 1.14 | 1.18 |
| Decay | −40.82 | −42.98 | −42.57 | −43.46 | 2.16 | 0.89 |

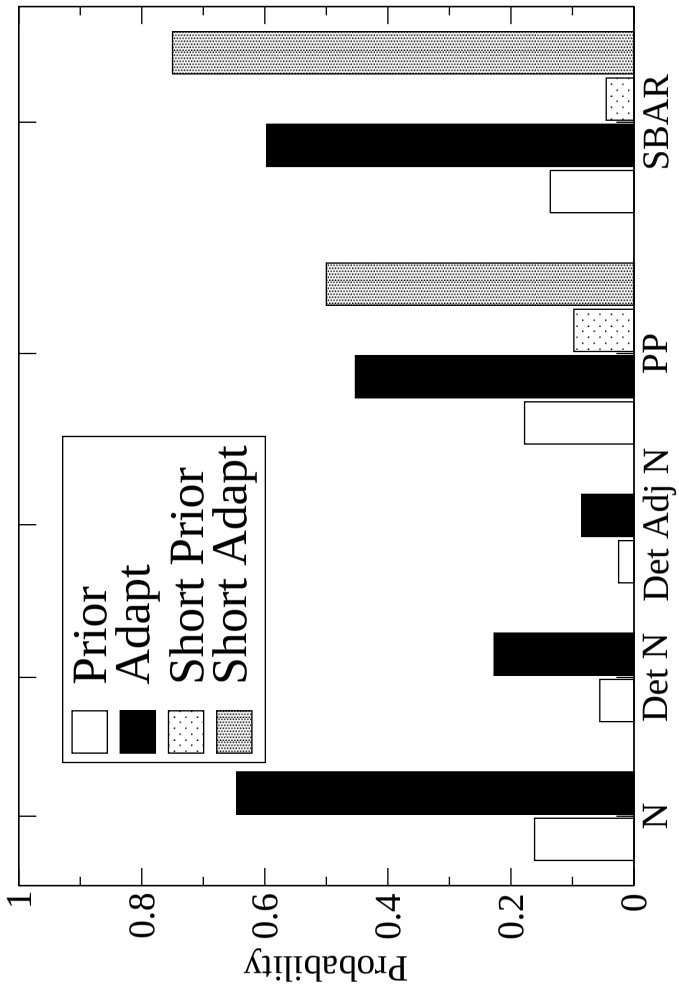
Table 10

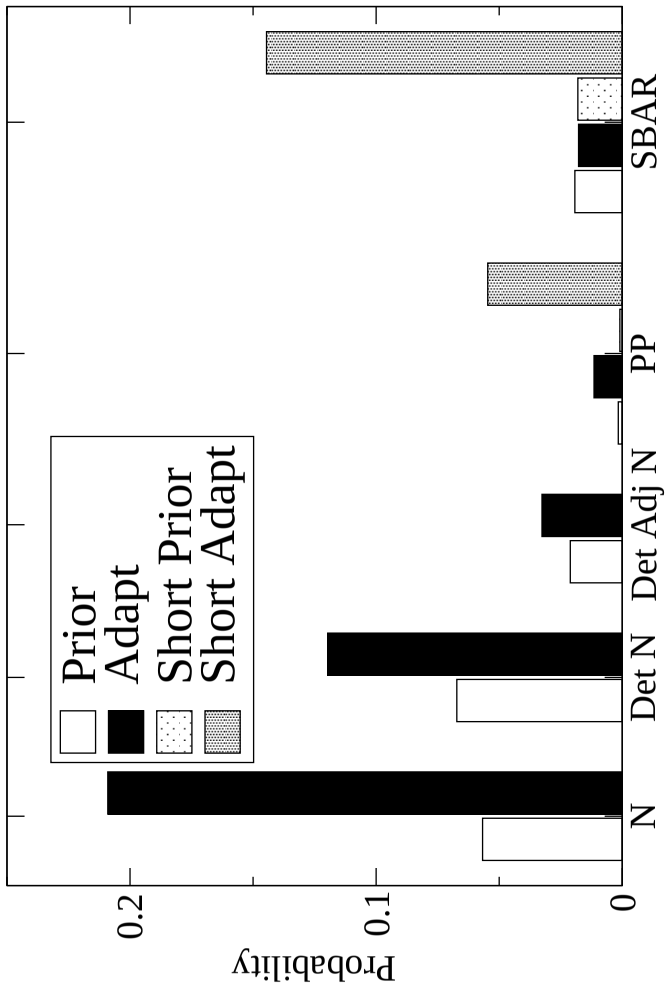
*Mean log probability estimates for the Knöferle and Crocker (2006) items*

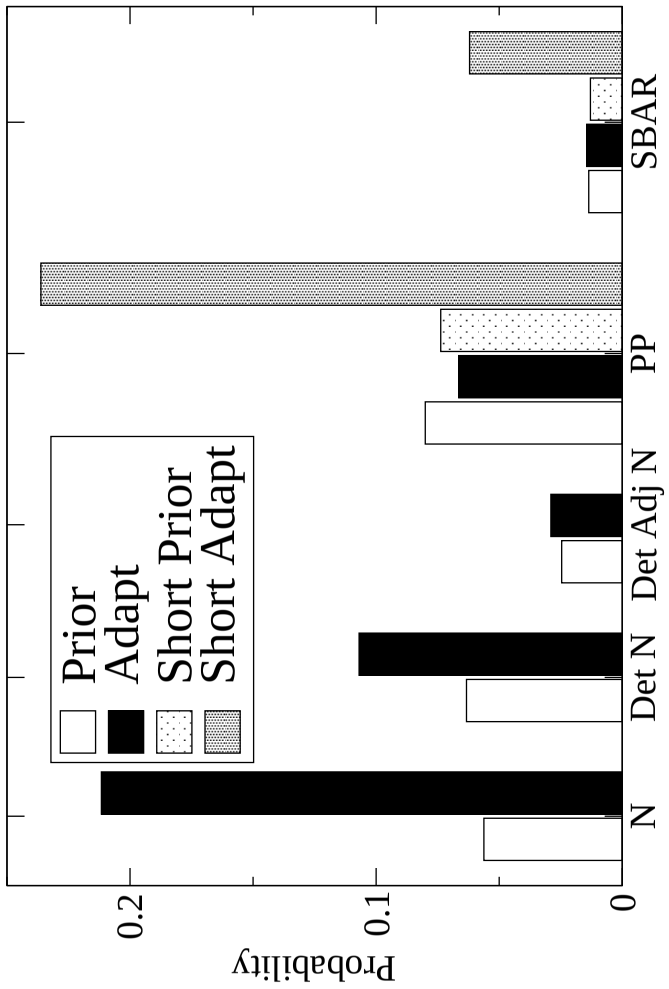
**Figure Captions**

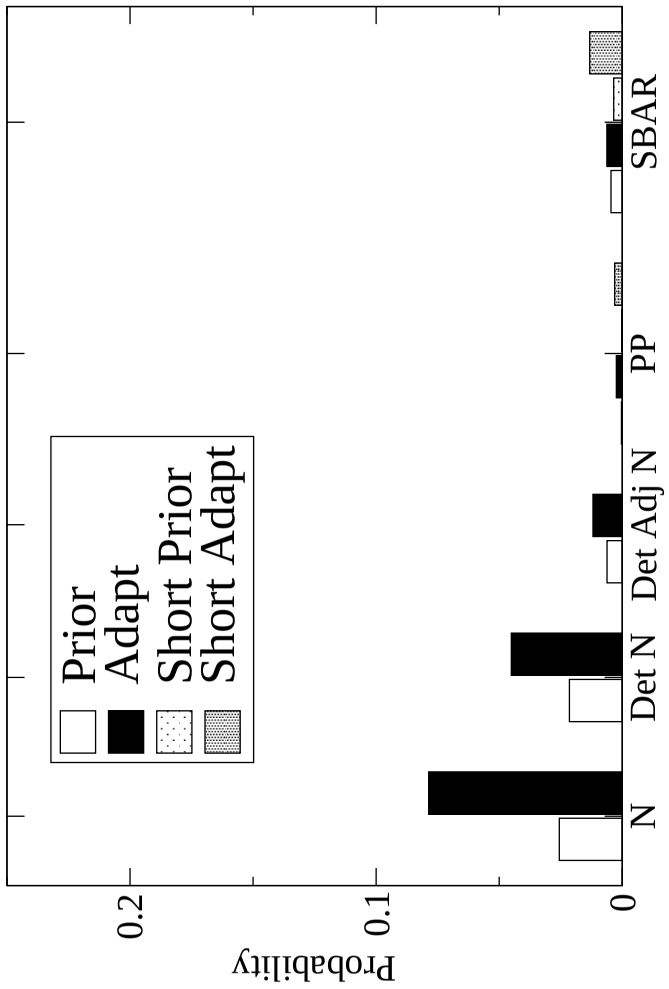
A Probabilistic Model of Parallelism, Figure 1

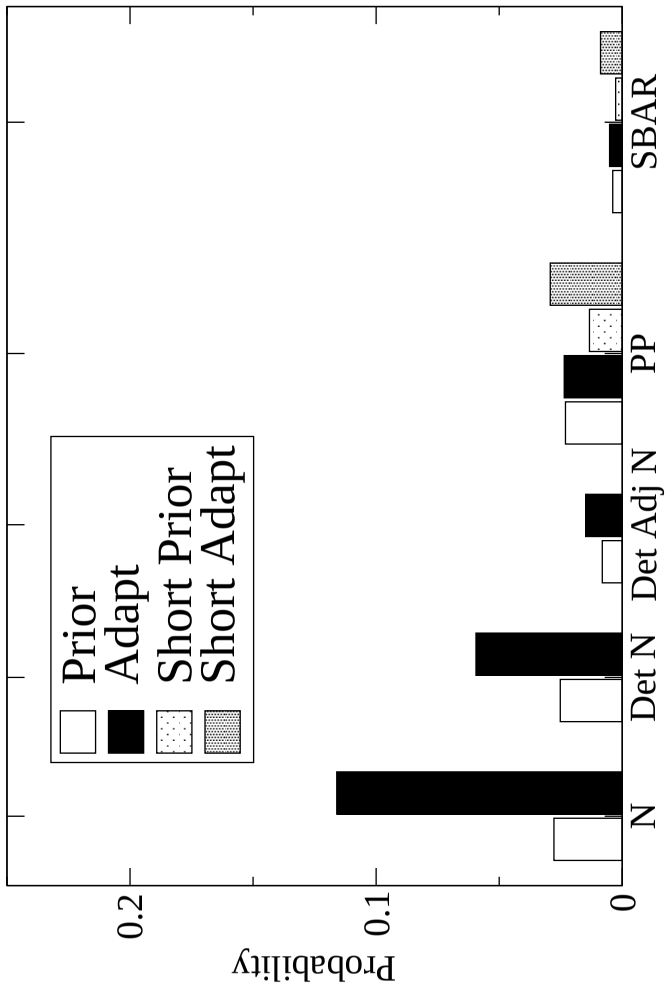


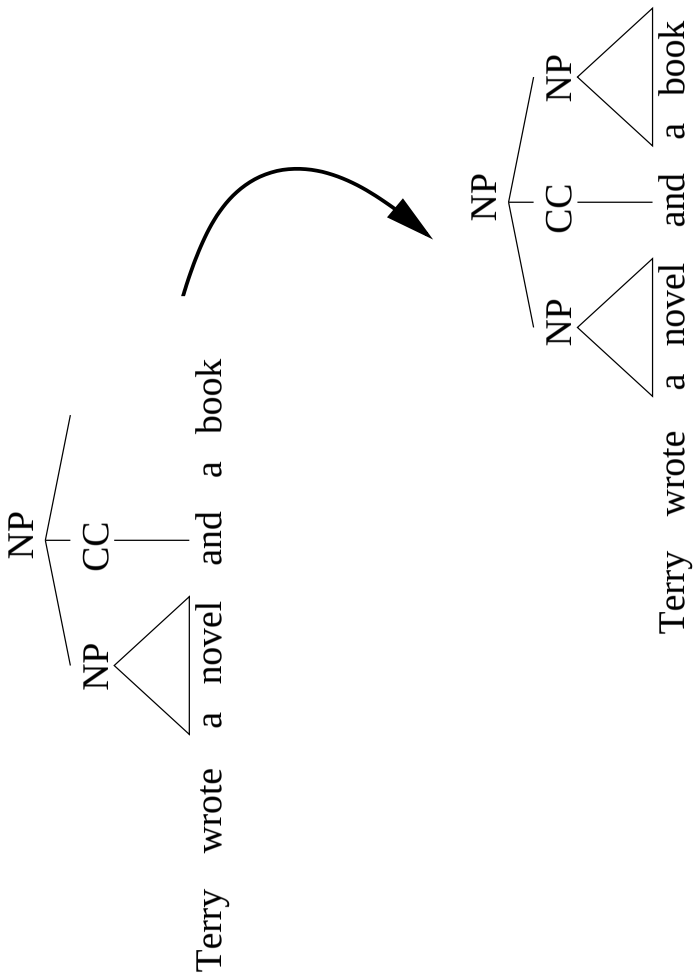

Probability
0
0.2
0.4
0.6
0.8
1
N
Det N
Det Adj N
PP
SBAR
Prior
Adapt
Short Prior
Short Adapt

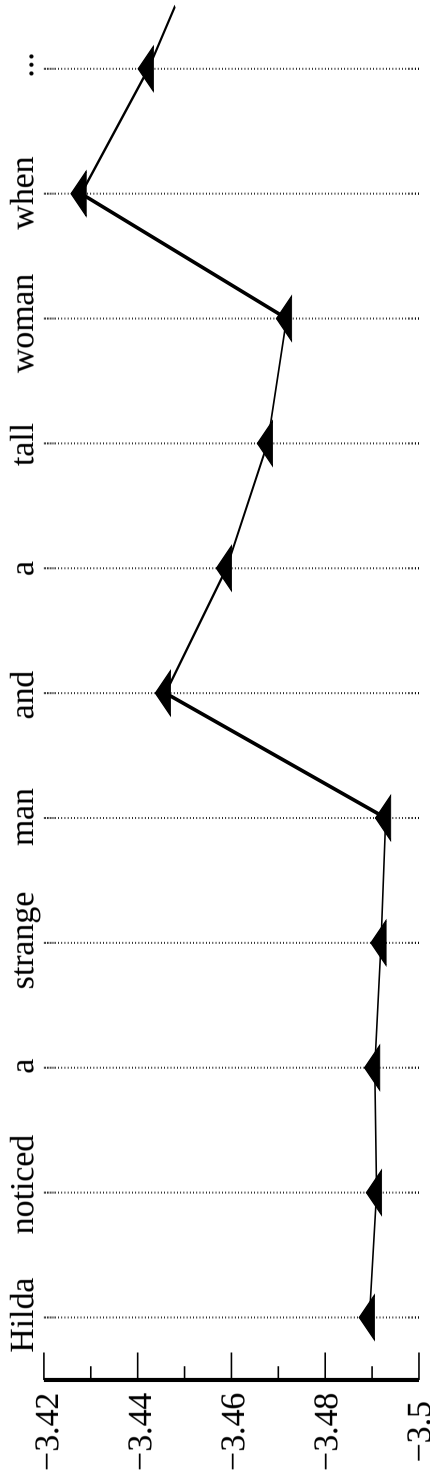

Probability
0
0.1
0.2
N
Det N
Det Adj N
PP
SBAR
Prior
Adapt
Short Prior
Short Adapt

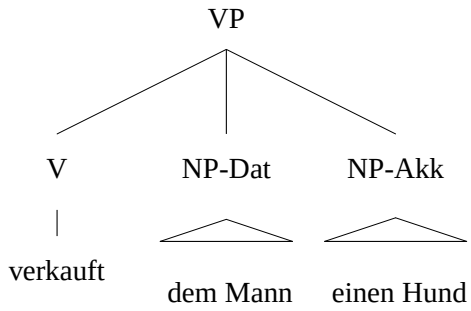

Probability
0
0.1
0.2
N
Det N
Det Adj N
PP
SBAR
Prior
Adapt
Short Prior
Short Adapt


Probability
0
0.1
0.2
N
Det N
Det Adj N
PP
SBAR
Prior
Adapt
Short Prior
Short Adapt

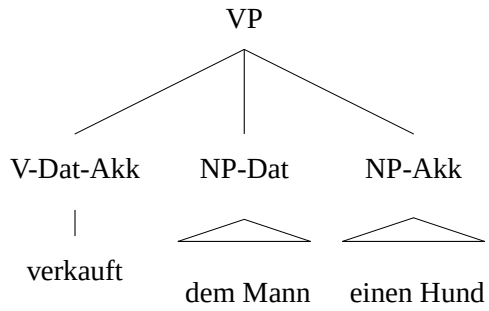

Probability
0
0.1
0.2
N
Det N
Det Adj N
PP
SBAR
Prior
Adapt
Short Prior
Short Adapt


NP
CC
NP
Terry wrote a novel and a book
NP
NP
CC
NP
Terry wrote a novel and a book

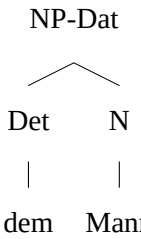

Hilda
noticed
a
strange
man
and
a
tall
woman
when
...
−3.42
−3.44
−3.46
−3.48
−3.5



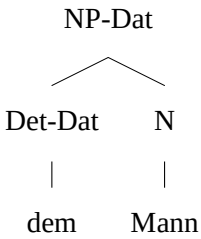

(a) A verb phrase (‘sold the man a dog’) as it appears in the corpus



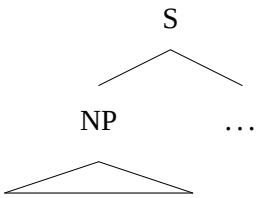

(b) Transforming the verb to include valence information




(a) A dative noun phrase (‘the man’) as it appears in the corpus

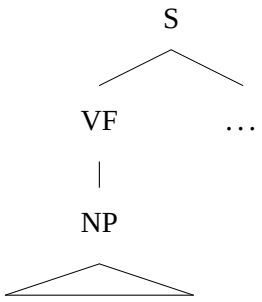



(b) A transformed accusative noun phrase




Nächste Woche

(a) A sentence with a constituent in the Vorfeld (‘next week’) as it appears in the corpus




Nächste Woche

(b) Transforming the sentence to overtly include a Vorfeld (VF) production