Hierarchic syntax improves reading time prediction

Marten van Schijndel **William Schuler**
Department of Linguistics
The Ohio State University
{vanschm,schuler}@ling.osu.edu

Abstract

Previous work has debated whether humans make use of hierarchic syntax when processing language (Frank and Bod, 2011; Fossum and Levy, 2012). This paper uses an eye-tracking corpus to demonstrate that hierarchic syntax significantly improves reading time prediction over a strong *n*-gram baseline. This study shows that an interpolated 5-gram baseline can be made stronger by combining *n*-gram statistics over entire eye-tracking regions rather than simply using the last *n*-gram in each region, but basic hierarchic syntactic measures are still able to achieve significant improvements over this improved baseline.

1 Introduction

In NLP, a concern exists that models of hierarchic syntax may be increasingly used exclusively to compensate for *n*-gram sparsity (Lease et al., 2006). In the context of psycholinguistic modeling, Frank and Bod (2011) find that hierarchic measures of syntactic processing are not as good at predicting reading times as sequential part-of-speech-based models of processing.[1] Fossum and Levy (2012) follow up on this finding and show that, when better *n*-gram information is present in the models, measures of hierarchic syntactic processing cost (PCFG surprisal; Hale, 2001; Levy, 2008) are as good at predicting reading times as the sequential models presented by Frank and Bod.

[1] Frank and Bod (2011) find that hierarchic measures significantly improve the descriptive linguistic accuracy of models but that such measures are unable to improve upon a strong linear baseline when predicting reading times.

The present study builds on this finding by showing that cumulative *n*-gram probabilities significantly improve an *n*-gram baseline to better capture sequential frequency statistics. Further, this study shows that measures of hierarchic structural frequencies (as captured by PCFG surprisal) significantly improve reading time predictions over that improved sequential baseline.

First, this work defines a stronger *n*-gram baseline than that used in previous studies by replacing a bigram baseline computed from 101 million words with an interpolated 5-gram baseline computed over 2.96 billion words. Second, while previous work has used *n*-grams from the end of each eye-movement region to model reading times in that region, this paper finds that such models can be significantly improved by combining *n*-gram statistics over the entire region (Section 3). Even when this improved baseline is combined with a standard *n*-gram baseline, this paper demonstrates that PCFG surprisal is a significant predictor of reading times (Section 4). This paper also applies region accumulation to total surprisal and finds that it is not significantly better than non-accumulated total surprisal. In fact, cumulative surprisal is shown not to be a significant predictor of reading times at all when a cumulative *n*-gram factor is included in the baseline. Finally, this paper compares two different models of hierarchic syntax: the Penn Treebank (PTB) representation (Marcus et al., 1993) and the psycholinguistically-motivated Nguyen et al. (2012) Generalized Categorial Grammar (GCG). Each model of syntax is shown to provide orthogonal improvements to reading time predictions (Section 5).

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1597–1605,
Denver, Colorado, May 31 – June 5, 2015.

Factors	Durations	
	R_{w4}^{w4}	R_{w5}^{w6}
Bigram	P(w_4|w_3)	P(w_6|w_5)
Cumu-Bigram	P(w_4|w_3)	P(w_6|w_5)·P(w_5|w_4)

Table 1: Bigram factors and their predictions of reading times in example eye-tracking regions. w_i represents word i. R_{wi}^{wj} represents the region from w_i to w_j (inclusive).

Bigram: The [1] red apple that the [2] girl ate ...

Cumu-Bigram: The [1] red apple that the [2] girl ate

X: bigram targets X: bigram conditions

Table 2: Influences on bigram factor predictions of reading times on *girl* following fixation on *red*.

2 Modeling

This study fits models to reading times from the Dundee corpus (Kennedy et al., 2003), which consists of eye-tracking data from 10 subjects who read 2388 sentences of news text from the newspaper, *The Independent*. Prior to using this corpus for evaluations, the first and last fixation of each sentence and of each line are filtered out to avoid potentially confounding wrap-up effects. Additionally, all fixations after saccades (eye movements) over more than 4 words are removed to avoid confounds with eye-tracker track-loss.

All evaluations are done with linear mixed effects models using lme4 (version 1.1-7; Bates et al., 2014).[2] There are two dependent reading time variables of interest in this study: first pass durations and go-past durations. During reading, a person's eye can jump over multiple words each time it moves, this study refers to that span of words as a *region*. First pass durations measure elapsed time until a person's eye leaves a given region. Go-past durations measure elapsed time until a person's eye moves further in the text. For example, in the fixation sequence: word 4, word 6, word 3, word 7, the first region would be from word 4 to word 6 and the second region would be from word 6 to word 7. The first pass duration for the first region would consist of the time fixated on word 6 before leaving the region for word 3, while the go-past duration would consist of the duration from the fixation of word 6 until the fixation of word 7. Separate models are fit to each centered dependent variable.

There are a number of independent variables in all evaluations in this study: sentence position (sentpos), word length (wlen), region length in words (rlen), whether the previous word was fixated (prevfix), and basic 5-gram log probability of the current word given the preceding context (5-gram). All independent predictors are centered and scaled before being added to each model. The 5-gram probabilities are interpolated 5-grams computed over the Gigaword 4.0 corpus (Graff and Cieri, 2003) using KenLM (Heafield et al., 2013). Gigaword 4.0 consists of around 2.96 billion words from around 4 million English newswire documents, which provides appropriate *n*-gram statistics since the Dundee corpus is also English news text.

Each mixed effects model contains random intercepts for subject and word, and random by-subject slopes for all fixed effects. Since the following evaluations use ablative testing to determine whether a fixed effect significantly improves the fit of a model compared to a model without that fixed effect, all models in a given evaluation include random slopes for all fixed effects used in that evaluation, even if the fixed effect is absent from that particular model.

3 A Cumulative *N*-gram Predictor

Since *n*-gram frequencies can have such a dramatic impact on the contribution of hierarchic syntax, this study tests whether *n*-gram factors can be improved. Models include a measure of *n*-gram frequencies to capture the rarity of observed sequences. Readers fixate longer on less predictable lexemes than on more predictable lexemes, but the predictability of a lexeme depends on the preceding context. Therefore, it is common for psycholinguistic models to include a measure of *n*-gram predictability for each fixated word conditioned on its context, but unless probabilities for words between fixations are also included, the probabilities used in this calculation are

[2]The models are fit using both the default *bobyqa* and the gradient *nlminb* algorithms to work around convergence issues.

Model	First Pass		Go-Past	
	Log-Likelihood	AIC	Log-Likelihood	AIC
Baseline	-1212399	2424868	-1261582	2523234
Base+*N*-gram	$-1212396^{\dagger}$	2424864	-1261577^{*}	2523226
Base+Cumu-*N*-gram	-1212392^{*}	2424856	-1261576^{*}	2523224
Base+Both	-1212387^{*}	2424848	-1261570^{*}	2523214

Baseline random slopes: sentpos, wlen, rlen, prevfix, 5-gram, cumu-5-gram
Baseline fixed effects: sentpos, wlen, rlen, prevfix

Table 3: Goodness of fit of *N*-gram models to reading times.[3] Significance testing was done between each model and the models in the section above it. Significance for Base+Both applies to improvement over each of the *n*-gram models. † $p < .05$ * $p < .01$

not probabilities of complete word sequences and may miss words that are parafovially previewed or simply inferred.

For example, in Table 1, the standard bigram factor (top line) predicts that the reading time of the region that ends with word 6 depends on word 5, but the probability of word 5 given its context is never included in the model, so an improbable transition between words 4 and 5 would not be caught. This might allow another factor to inappropriately receive credit for an extra long fixation on word 6. Instead, a better model would include the probabilities of every word in the sequence since that is the information that will need to be processed by the reader. Using log-probabilities, a *cumulative n*-gram factor can be created simply by summing the log probabilities over each region (comparable to the last line of Table 1). The cumulative *n*-gram predictor is able to account for the frequency of the entire lexical sequence and so should provide a better reading time predictor than the standard fixation-only *n*-gram predictor (see Table 2 for an example).

For this initial evaluation (Table 3), the baseline omits the fixed *n*-gram factor. Instead, a model is constructed without any fixed effects for *n*-gram. Then, the same model is fit to reading times after adding just a fixed effect for *n*-gram and after adding just a fixed effect for cumulative *n*-gram. Finally, a model is fit with both the cumulative and non-cumulative *n*-gram factors as fixed effects.[4] Significance between the models is determined using likelihood ratio testing.[5]

Table 3 shows that both *n*-gram factors significantly improve the fit of the model and the final line shows that each factor provides a significant orthogonal improvement. Both *n*-gram factors will therefore be included as fixed effects and as by-subject random slopes in the baselines of the remaining evaluations in this study.

4 Hierarchic Syntax Predictors

This section tests the main hypothesis of this study: that hierarchic syntactic processing is a significant contributor to reading times. For the purposes of this evaluation, total PCFG surprisal (Hale, 2001; Levy, 2008; Roark et al., 2009) will be used as a measure of hierarchic syntactic processing. Specifically, PCFG surprisal will be calculated using the van Schijndel et al. (2013a) incremental parser trained on sections 02-21 of the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993) using 5 iterations of split-merge (Petrov et al., 2006) and a beam width of 5000.

[3] Log-likelihood values are rounded to the nearest whole number, which is why the difference between Base and Base+Both can be larger than the cumulative difference between Base and the other two models.

[4] To ensure effects are not driven by individual subject differences, by-subject random slopes for both predictors of interest are included in the baseline. This practice is repeated throughout this study.

[5] Twice the log-likelihood difference of two nested models can be approximated by a χ^2 distribution with degrees of freedom equal to the difference in degrees of freedom of the models in question. The probability of obtaining a given log-likelihood difference D between the two models is therefore analogous to $P(2 \cdot D)$ under the corresponding χ^2 distribution.

Factors	Durations	
	R_{w4}^{w4}	R_{w5}^{w6}
surp	$-\log \mathrm{P}(w_4 \mid T_3)$	$-\log \mathrm{P}(w_6 \mid T_5)$
cumusurp	$-\log \mathrm{P}(w_4 \mid T_3)$	$\sum_{i=5}^{6} -\log \mathrm{P}(w_i \mid T_{i-1})$

Table 4: PCFG surprisal factors and their predictions of reading times in example eye-tracking regions. w_i represents word i. T_i represents the set of trees that can span from w_1 to w_i. R_{wi}^{wj} represents the region from w_i to w_j (inclusive).

4.1 Surprisal

PCFG surprisal (Hale, 2001; Levy, 2008) is a measure of incremental hierarchic syntactic processing. It reflects the information gained by observing a given word in a given context. In PCFG surprisal calculations, *context* is usually taken to refer to the preceding words in the sentence and their underlying syntactic structure. The PCFG surprisal $S(w_i)$ of a word at position i may be calculated as:

$$S(w_i) = \sum_{t \in T_{i-1}} -\log \mathrm{P}(w_i \mid t) \qquad (1)$$

where T_i represents the set of syntactic structures that can span from w_1 to w_i. PCFG surprisal in psycholinguistic models captures the influence of incremental hierarchic context when processing a given word.

For space considerations, in Table 4, the summation over T_{i-1} is notationally implicit:

$$S(w_i) = -\log \mathrm{P}(w_i \mid T_{i-1}) \qquad (2)$$

4.2 Evaluation

As in the previous section, a baseline model is fit to reading times without a fixed effect for surprisal, then surprisal is added as a fixed effect and significance of the fixed effect is determined using a likelihood ratio test with the baseline. The results (Table 5) show that PCFG surprisal is a significant predictor of both first pass and go-past durations even over a strong baseline including both types of *n*-gram factors.

The preceding section showed that applying region accumulation to an *n*-gram factor improves a model's fit to reading times. Previous work suggests region accumulation might improve the fit of syntactic factors to reading times (van Schijndel and Schuler, 2013; van Schijndel et al., 2013b), but the baselines in those studies only included unigram and bigram statistics and did not apply region accumulation to the *n*-gram models. It does make intuitive sense that region accumulation would help improve the fit of total PCFG surprisal for the same reason accumulating *n*-grams helps. For an example, see Table 4. A non-cumulative total PCFG surprisal factor (top line) would predict that duration of region R_{w5}^{w6} depends on T_5 (the set of trees that can span from w_1 to w_5), but the probability of generating the prefix of T_5 is never fully calculated by this factor. As with cumulative *n*-grams, cumulative PCFG surprisal of a region can be calculated by simply summing the PCFG surprisal of each word in the region.

When tested, however, the present work does not find any improvement from region accumulation of PCFG surprisal when stronger *n*-gram factors are also included (Table 5, Row 2), suggesting that the improvement in previous studies may have been due to latent *n*-gram information captured by cumulative PCFG surprisal. This finding is interesting because it suggests non-local hierarchic structure does not significantly influence reading times. The next section explores this hypothesis further by testing the fit of a hierarchic syntactic formalism whose strength lies in modeling long-distance dependencies.

5 Grammar Formalism Comparison

So far, this study has tried to allay previous concerns that models of hierarchic syntax may just be accounting for the sparsity of *n*-gram statistics (Charniak et al., 2006; Frank and Bod, 2011). This section investigates whether a representation of hierarchic syntax that preserves long-distance dependencies can improve reading time predictions over a hierarchic representation based on the Penn Treebank which discards long-distance dependencies. This evaluation compares total PCFG surprisal as calculated by the original Penn Treebank grammar to total PCFG surprisal calculated by the Nguyen et al. (2012) Generalized Categorial Grammar (GCG).

5.1 GCG

A GCG has a category set C, which consists of a set of primitive category types U, typically labeled

Model	First Pass		Go-Past	
	Log-Likelihood	AIC	Log-Likelihood	AIC
Baseline	-1212260	2424627	-1261488	2523084
Base+Surp	-1212253^{*}	2424617	-1261481^{*}	2523072
Base+CumuSurp	-1212259	2424627	-1261487	2523085
Base+Both	-1212253^{*}	2424619	-1261481^{*}	2523073

Baseline random slopes: sentpos, wlen, rlen, prevfix, 5-gram, cumu-5-gram, surp, cumusurp
Baseline fixed effects: sentpos, wlen, rlen, prevfix, 5-gram, cumu-5-gram

Table 5: Goodness of fit of hierarchic syntax models to reading times. Significance testing was done between each model and the models in the section above it. Significance for Base+Both applies only to improvement over the CumuSurp model. * $p < .01$

with the part of speech of the head of a category (e.g. **V**, **N**, **A**, etc., for phrases or clauses headed by verbs, nouns, adjectives, etc.), followed by one or more unsatisfied dependencies, each consisting of an operator (**-a** and **-b** for adjacent argument dependencies preceding and following a head, **-c** and **-d** for adjacent conjunct dependencies preceding and following a head, **-g** for filler-gap dependencies, **-r** for relative pronoun dependencies, and some others), followed by a dependent category type. For example, the category for a transitive verb would be **V-aN-bN**, since it is headed by a verb and has unsatisfied dependencies to satisfied noun-headed categories preceding and following it (for the subject and direct object noun phrase, respectively).

As in other categorial grammars, inference rules for local argument attachment apply functors of category c-**a**d or c-**b**d to initial or final arguments of category d:

$$d \;\; c\text{-}\mathbf{a}d \Rightarrow \; c \qquad \text{(Aa)}$$

$$c\text{-}\mathbf{b}d \;\; d \Rightarrow \; c \qquad \text{(Ab)}$$

However, the Nguyen et al. (2012) GCG uses distinguished inference rules for modifier attachment, which allows modifier categories to be consolidated with categories for modifiers in other contexts (preverbal, post-verbal, etc.), and with certain predicative categories. This allows derivations in the training corpus involving different modifier types to also be consolidated, which increases the power of the extracted statistics. Inference rules for modifier attachment apply initial or final modifiers of category u-**a**d to modificands of category c, for $u \in U$ and $c, d \in C$:

$$u\text{-}\mathbf{a}d \;\; c \Rightarrow c \qquad \text{(Ma)}$$

$$c \;\; u\text{-}\mathbf{a}d \Rightarrow c \qquad \text{(Mb)}$$

The Nguyen et al. (2012) GCG also uses distinguished inference rules to introduce, propagate, and bind missing non-local arguments, similar to the gap or slash rules of Generalized Phrase Structure Grammar (Gazdar et al., 1985) and Head-driven Phrase Structure Grammar (Pollard and Sag, 1994). Inference rules for gap attachment hypothesize gaps as initial arguments, final arguments, or modifiers, for $c, d \in C$:

$$c\text{-}\mathbf{a}d \Rightarrow c\text{-}\mathbf{g}d \qquad \text{(Ga)}$$

$$c\text{-}\mathbf{b}d \Rightarrow c\text{-}\mathbf{g}d \qquad \text{(Gb)}$$

$$c \Rightarrow c\text{-}\mathbf{g}d \qquad \text{(Gc)}$$

Non-local arguments, using non-local operator and argument category $\psi \in \{\text{-}\mathbf{g}, \text{-}\mathbf{h}, \text{-}\mathbf{i}, \text{-}\mathbf{r}\} \times C$, are then propagated to the consequent from all possible combinations of antecedents. For each rule $d \;\; e \Rightarrow c \in \{\text{Aa–b,Ma–b}\}$:

$$d \;\; e\psi \Rightarrow c\psi \qquad \text{(Ac–d,Mc–d)}$$

$$d\psi \;\; e \Rightarrow c\psi \qquad \text{(Ae–f,Me–f)}$$

$$d\psi \;\; e\psi \Rightarrow c\psi \qquad \text{(Ag–h,Mg–h)}$$

In order to consolidate relative and interrogative pronouns in different pied-piping contexts into just two reusable categories, this grammar uses distinguished inference rules for relative and interrogative pronouns as well as tough constructions (e.g. *this*

Model	First Pass		Go-Past	
	Log-Likelihood	AIC	Log-Likelihood	AIC
Baseline	-1212242	2424592	-1261474	2523055
Base+PTB	-1212239^{*}	2424587	-1261468^{*}	2523047
Base+GCG	$-1212239^{\dagger}$	2424589	-1261470^{*}	2523050
Base+Both	$-1212235^{\dagger}$	2424583	-1261465^{*}	2523043

Baseline random slopes: sentpos, wlen, rlen, prevfix, 5-gram, cumu-5-gram, surp-GCG, surp-PTB
Baseline fixed effects: sentpos, wlen, rlen, prevfix, 5-gram, cumu-5-gram

Table 6: Goodness of fit of models with differing syntactic calculations to reading times. Significance testing was done between each model and the models in the section above it. Base+Both first pass significance applies to improvement over PTB ($p < .05$) and to improvement over GCG ($p < .01$), Base+Both go-past significance applies to improvement over each independent model. † $p < .05$ * $p < .01$

bread is easy to cut), which introduce clauses with gap dependencies, for $c, d, e \in C$, $\psi \in \{\textbf{-g}\} \times C$:

$$d\textbf{-i}e\ c\textbf{-g}d \Rightarrow c\textbf{-i}e \quad \text{(Fa)}$$

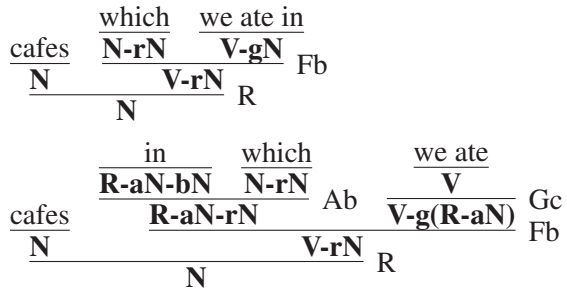
$$d\textbf{-r}e\ c\textbf{-g}d \Rightarrow c\textbf{-r}e \quad \text{(Fb)}$$

$$c\textbf{-b}(d\psi)\ d\psi \Rightarrow c \quad \text{(Fc)}$$

Also, inference rules for relative pronoun attachment apply pronominal relative clauses of category $c\textbf{-r}d$ to modificands of category e:

$$e\ \ c\textbf{-r}d \Rightarrow e \quad \text{(R)}$$

Because of its richer set of language-specific inference rules, the GCG grammar annotated by Nguyen et al. (2012) does not require different categories for words like *which* in different pied-piping contexts:



5.2 Evaluation

Following van Schijndel et al. (2013b), the GCG calculation of PCFG surprisal comes from a GCG-reannotated version of the Penn Treebank whose grammar rules have undergone 3 iterations of the split-merge algorithm (Petrov et al., 2006). A k-best beam with a width of 5000 is used in order to be comparable to the PTB calculation.

Significance testing is done as in the preceding evaluations: a baseline model is fit to reading times, each PCFG surprisal factor is added independently to the baseline, and both PCFG surprisal factors are added concurrently to the baseline. Each model is compared to the next simpler models using likelihood ratio tests.

The results (Table 6) show that GCG PCFG surprisal is a significant predictor of reading times even in the presence of the stronger n-gram baseline. Moreover, both PTB and GCG PCFG surprisal significantly improve reading time predictions even when the other PCFG surprisal measure is also included. This suggests that each is contributing something the other is not. Since the GCG grammar is derived from an automatically reannotated version of the Penn Treebank, there may be errors in the GCG annotation which cause errors in the estimates of underlying GCG structure. Since the PTB grammar is manually annotated by experts, the PTB grammar may be receiving credit for correct structural prediction in cases where GCG's estimates are incorrect. However, it seems likely that GCG may be providing a better fit in cases of long-distance dependencies because such relations are omitted from the PTB grammar.

A follow-up evaluation (not shown here) using the experimental design from Section 4 but using GCG PCFG surprisal rather than PTB PCFG surprisal revealed that cumulative PCFG surprisal is still not a significant predictor when calculated using GCG. The failure of cumulative PCFG surprisal to improve over basic GCG PCFG surprisal could be expected

Predictor	First Pass		Go-Past	
	coef	t value	coef	t value
sentpos	−2.47	−3.59	−2.82	−3.38
wlen	25.90	8.67	28.98	9.97
prevfix	−30.16	−7.81	−37.42	−11.49
n-gram	−2.39	−1.81	−6.70	−3.36
cumu-*n*-gram	−14.69	−7.36	−11.68	−5.01
rlen	−5.67	−1.31	−12.51	−2.59
surp-GCG	4.97	2.87	5.74	2.73
surp-PTB	4.20	3.23	4.85	3.29

Table 7: Fixed effect predictor coefficients for Base+PTB+GCG model.

since a strength of GCG is in enabling non-local decisions on a local basis (by propagating non-local decisions into the category labels), so any non-local advantage cumulative PCFG surprisal might confer is already compressed into the GCG categories.

The results of this evaluation suggest that reading times are mostly affected by local hierarchic structure, but the fact that GCG PCFG surprisal is able to provide a significant fit even in the presence of the PTB PCFG surprisal predictor suggests that some non-local information affects reading times. In particular, while this evaluation showed that accumulated syntactic context is not generally a good predictor of reading times, some or all of the non-local information contained in the GCG categories is used by readers and so influences reading time durations over the local structural information reflected in the PTB PCFG surprisal measure.

6 Discussion

The finding that the hierarchic grammars orthogonally improve reading time predictions suggests that hierarchic structural information has a significant influence on reading times. Since both the PTB and GCG calculations of surprisal contain sequential information (e.g., of part-of-speech tags), if the effect in this study was driven by purely sequential information as suggested by Frank and Bod (2011), one might expect either the PTB or the GCG calculations of surprisal (but not both) to be a significant predictor of reading times.

Instead, the present set of results support recent claims made by van Schijndel et al. (2014) that non-local subcategorization decisions are made early during processing and so would have a strong influence on the reading time measures used in the present study. Such decisions would have to be conditioned on hierarchic structural information not present in either PTB PCFG surprisal or the sequential structure models of Frank and Bod (2011).

Further, predictability has been shown to affect word duration during speech production (Jurafsky et al., 2001; Aylett and Turk, 2006), and Demberg et al. (2012) found that hierarchic structure significantly improves over *n*-gram computations of predictability in that domain as well. Together, these findings suggest that hierarchic structure is not only a convenient descriptive tool for linguists, but that such structure is deeply rooted in the human language processor and is used during online language processing.

Previous work has made a distinction between lexical surprisal, syntactic surprisal, and total surprisal (Demberg and Keller, 2008; Roark et al., 2009). Given a prefix derivation of the structure of the context, syntactic surprisal measures the information obtained by generating the structure that will enable the attachment of a newly observed lexical item. Lexical surprisal conveys the amount of information obtained by attaching the particular lexical observation to the new syntactic structure. Total surprisal is the original formulation of surprisal and is the composition of the other two types of surprisal (the information gained by generating a structure for the current lexical observation and attaching the observation to that structure). Fossum and Levy (2012) show that, with a non-cumulative bigram baseline, this distinction is not significant when predicting

reading times, so the present study simply uses total surprisal. It may be interesting in future work to see if the distinction between surprisal types becomes more or less useful as the sequential baseline improves.

The finding that cumulative *n*-gram information is useful in predicting reading times bears some resemblance to the finding that the spillover effect of a word is proportional to its logarithmic probability given the context (Smith and Levy, 2013). However, the spillover effect studied by Smith and Levy (2013) is one of a given fixation on the following fixation. The cumulative *n*-grams, in contrast, permit finer predictability of a word given the unfixated intervening context. The two measures are similar in that they both permit better modeling of the predictability of a word given its context, but the spillover measure could also be easily conceived as continued spillover processing from the preceding fixation, while cumulative *n*-grams reflect the predictability of the entire region between one fixation and the next. Further, cumulative *n*-grams could conceivably also capture processing of parafovial preview obtained during the previous fixation. Since the cumulative *n*-gram measure improves the computation of predictability of a word, it could also provide a better measure of the spillover effect a given word will have. Future work could investigate this by using cumulative *n*-grams both to compute the predictability of the current word and to predict the spillover effect from the preceding fixation. The present work suggests that doing so would provide even better reading time predictors.

7 Conclusion

First, this work suggests that the standard accounting for *n*-gram frequencies needs to change in psycholinguistic studies. Currently, the standard procedure is to use *n*-gram statistics only from the end of an eye-tracking region. This standard calculates the influence of the final word in each region given the lexical context, but that context is never accounted for in regions greater than one word in length. Instead, psycholinguistic models need to additionally account for the probability of the *context* given its own preceding context to provide a coherent model of the probability of the observed lexical sequence.

This work also shows that, even with good cumulative and non-cumulative estimates of the frequency effects generated by a given lexical sequence, measures of hierarchic structure provide a significant improvement to reading time predictions. Further, even in the presence of both a strong *n*-gram baseline and a linguistically accurate measure of hierarchic structure (PTB with 5 iterations of split-merge), a linguistically-motivated model of hierarchic structure is a significant predictor of reading times. As data coverage grows, some may worry that models of syntax will be superseded by better *n*-gram models. This study suggests that hierarchic syntax retains its value even in a world of big data.

Acknowledgements

Thanks to Stefan Frank for interesting discussion and helpful feedback on an earlier draft of this paper and to the anonymous reviewers for their comments. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1343012. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Matthew Aylett and Alice Turk. 2006. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *Journal of the acoustical society of America*, 119(5):3048–3059.

Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker, 2014. *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7.

Eugene Charniak, Mark Johnson, Micha Elsner, Joseph Austerweil, David Ellis, Isaac Haxton, Catherine Hill, R. Shrivaths, Jeremy Moore, Michael Pozar, and Theresa Vu. 2006. Multilevel coarse-to-fine pcfg parsing. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 168–175.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Vera Demberg, Asad B. Sayeed, Philip J. Gorinski, and Nikolaos Engonopoulos. 2012. Syntactic surprisal

affects spoken word duration in conversational contexts. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 356–367.

Victoria Fossum and Roger Levy. 2012. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of CMCL 2012*. Association for Computational Linguistics.

Stefan Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*.

Gerald Gazdar, Ewan Klein, Geoffrey Pullum, and Ivan Sag. 1985. *Generalized Phrase Structure Grammar*. Harvard University Press, Cambridge, MA.

David Graff and Christopher Cieri, 2003. *English Gigaword LDC2003T05*.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics*, pages 159–166, Pittsburgh, PA.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August.

Daniel Jurafsky, Alan Bell, Michelle Gregory, and William D. Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In Joan Bybee and Paul Hopper, editors, *Frequency and the emergence of linguistic structure*, pages 229–254. John Benjamins, Amsterdam.

Alan Kennedy, James Pynte, and Robin Hill. 2003. The Dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Matthew Lease, Eugene Charniak, Mark Johnson, and David McClosky. 2006. A look at parsing and its applications. In *Proceedings of AAAI*.

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

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Luan Nguyen, Marten van Schijndel, and William Schuler. 2012. Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*, pages 2125–2140, Mumbai, India.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL'06)*.

Carl Pollard and Ivan Sag. 1994. *Head-driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. *Proceedings of the 2009 Conference on Empirical Methods in Natural Langauge Processing*, pages 324–333.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

Marten van Schijndel and William Schuler. 2013. An analysis of frequency- and recency-based processing costs. In *Proceedings of NAACL-HLT 2013*. Association for Computational Linguistics.

Marten van Schijndel, Andy Exley, and William Schuler. 2013a. A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540.

Marten van Schijndel, Luan Nguyen, and William Schuler. 2013b. An analysis of memory-based processing costs using incremental deep syntactic dependency parsing. In *Proc. of CMCL 2013*. Association for Computational Linguistics.

Marten van Schijndel, William Schuler, and Peter W Culicover. 2014. Frequency effects in the processing of unbounded dependencies. In *Proc. of CogSci 2014*. Cognitive Science Society.