Compound effect of probabilistic disambiguation and memory retrievals on sentence processing: Evidence from an eye-tracking corpus

Umesh Patil (umesh.patil@gmail.com)
Department of Linguistics, Karl-Liebknecht Str. 24–25
14476 Potsdam, Germany

Shravan Vasishth (vasishth@uni-potsdam.de)
Department of Linguistics, Karl-Liebknecht Str. 24–25
14476 Potsdam, Germany

Reinhold Kliegl (kliegl@uni-potsdam.de)
Department of Psychology, Karl-Liebknecht Str. 24–25
14476 Potsdam, Germany

Abstract

We evaluate the predictions of surprisal and cue-based theory of sentence processing using an eye-tracking corpus, the Potsdam Sentence Corpus. Surprisal is a measure of processing complexity based on a probabilistic grammar and is computed in terms of the total probability of structural options that have been disconfirmed at each input word. The cue-based theory characterizes processing difficulty in terms of working memory costs that derive from decay and interference arising during content-based retrieval requests of previously processed material (e.g., to incrementally build the sentence structure). We show that both surprisal and cue-based parsing independently explain difficulty in sentences processing and interestingly, they have an over-additive effect on processing when combined together.

Keywords: Sentence processing; eye-tracking; cue-based theory; surprisal; memory retrievals

Introduction

Research in psycholinguistics provides much evidence for probabilistic disambiguation in human language processing at various levels including lexical, syntactic and semantic processing (Jurafsky, 1996, 2003). More frequent words and structures are easier to comprehend than less frequent ones. Surprisal (Hale, 2001) is a proposal which characterizes processing difficulty in terms of the amount of work done in probabilistically disconfirming sentence continuations as a consequence of the information supplied by the current word. Consider, for example, the famous garden path sentence in (1). It has been observed that English speakers hearing this sentence have great difficulty at "fell". Hale (2001) demonstrates using probabilistic context-free grammar that the difficulty occurs because at "fell" the parser has to disconfirm alternatives that together comprise a great amount of the probability mass.

(1) The horse raced past the barn *fell*.

Recent research in computational models of sentence comprehension has shown that surprisal is a significant predictor of eye movements while reading individual sentences and text (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Demberg & Keller, 2008). However, surprisal is likely to furnish only part of the explanation (Levy, 2008). As Lewis (1996) and Gibson (2000) argue, sometimes people take longer to process words that they need to connect to other words processed earlier. Resolving these linguistic relations seems to impose more processing effort even when the constructions are frequent or unsurprising. Grodner and Gibson (2005) provide evidence using self-paced reading study which involved reading sentences like (2) below. They observed monotonically increasing reading time at the verb "supervised" as a function of its distance from the subject "nurse".

(2) a. The nurse *supervised* the ...
b. The nurse from the clinic *supervised* the ...
c. The nurse who was from the clinic *supervised* the ...

This difference between surprisal and integration cost was addressed by Demberg and Keller (2008), who compared the predictions of surprisal with Gibson's (2000) Dependency Locality Theory (DLT), a theory of integration difficulty. They found that DLT's predictions played a limited role in explaining processing difficulty. DLT was a significant predictor only for reading times at nouns and verbs. Here we show that surprisal and retrieval costs unequivocally play a role in determining processing difficulty. More interestingly, we observed a significant interaction of surprisal and memory retrievals, suggesting that a simple additive model of surprisal and retrieval processes will not suffice.

We compared surprisal's predictions to the cue-based retrieval model of (Lewis & Vasishth, 2005) (LV05 henceforth) using the Potsdam Sentence Corpus (PSC) of German (Kliegl, Nuthmann, & Engbert, 2006). The

cue-based retrieval theory characterizes processing difficulty in terms of working memory costs that derive from decay and interference arising during content-based retrieval requests of previously processed material, e.g., to complete dependencies, or to incrementally build structure.

We implemented cue-based retrieval models for sentences from the PSC, closely following the approach taken by LV05 and generated predictions for retrieval cost at each word. We also computed surprisal's predictions using a probabilistic phrase-structure parser. The main findings are that (1) retrieval cost furnishes better models of eye-fixation measures than models based on baseline predictors such as unigram and bigram frequency, word length, Cloze predictability plus surprisal, and (2) surprisal and retrieval cost show a significant interaction in predicting reading times.

Surprisal

Surprisal offers a theoretical reason why a particular word in a sentence should be easier or more difficult to comprehend on the basis of underlying probabilistic grammatical knowledge of the language. The idea of surprisal is to model processing difficulty as a logarithmic function of the probability mass eliminated by the most recently added word. This number is a measure of the information value of the word just seen, as rated by the grammar's probability model; it is nonnegative and unbounded. More formally, the surprisal of the n^{th} word (w_n) in a sentence is defined as the log-ratio of the prefix probability before seeing the word, compared to the prefix probability after seeing it. The prefix probability at word w_n is defined as the total probability of all grammatical analyses that derive the prefix string $w = w_1 \cdots w_n$ which is initial part of the bigger string wv. For grammar G and a set of derivations D the prefix probability α_n at word w_n can be expressed as:

$$prefix\_probability(w, G) = \sum_{d \in D(G, wv)} probability(d) \; = \; \alpha_n$$

Then, the surprisal at w_n is:

$$surprisal(w_n) = \log_2(\frac{\alpha_{n-1}}{\alpha_n})$$

Intuitively, surprisal and hence the difficulty of processing increases when a parser is required to build some low-probability structure.

Cue-based theory

The cue-based theory of sentence processing is derived from the application of independently motivated principles of memory and cognitive skills to the specialized task of sentence parsing. As a result, sentence processing emerges as a series of skilled associative memory retrievals modulated by similarity-based interference and fluctuating activation. The corresponding parsing model is implemented in the cognitive architecture ACT-R (Anderson et al., 2005) which formalizes the cognitive principles mentioned above in terms of the following set of equations:

1. The base activation (B_i) of chunk i, where t_j is the time since the j^{th} retrieval of the item, d is the decay parameter, and the summation is over all n retrievals, is
$$B_i = ln(\sum_{j=1}^{n} t_j^{-d})$$

2. Total activation (A_i) of a chunk i is defined as the summation of its base activation and strength of association. W_j is the amount of activation from the elements j in the goal buffer and S_{ji}s are the strengths of association from elements j to chunk i
$$A_i = B_i + \sum_j W_j S_{ji}$$

3. S_{ji} is defined in terms of fan_j which is the number of items associated with j
$$S_{ji} = S - ln(fan_j)$$

4. Retrieval latency of chunk i is defined in terms of A_i and F, a scaling constant
$$T_i = Fe^{-A_i}$$

The cue-based retrieval theory quantifies the processing difficulty at each word in terms of its *attachment time*, which is the sum of (i) the time required to retrieve the currently-built syntactic structure in order to attach the word into that structure, and (ii) a baseline cost of 100 milliseconds, which is the time required for the execution of the retrieval request and the subsequent attachment of the current word into the existing structure. See LV05 for details about data structures and the parsing algorithm used.

To summarize, the delay in retrieval of a prior syntactic element due to similarity based interference and fluctuating activation is assumed to induce difficulty in processing.

Experiment

The experiment involved a quantitative evaluation of the predictions of surprisal and cue-based theory using a corpus of eye movements during reading single sentences.

Methods

Data For the analyses in this paper, we selected 32 sentences from the Potsdam Sentence Corpus (PSC), which is an eye-tracking corpus consisting of fixation

durations recorded from 222 persons, each reading 144 German sentences (Kliegl, Nuthmann, & Engbert, 2006). These 32 sentences were selected in a way that enabled us to cover a wide range of syntactic structures.

For generating surprisal values for each word in these selected sentences we used a probabilistic context-free phrase-structure parser from Levy (2008), which is an implementation of Stolcke's Earley parser (Stolcke, 1995). We unlexicalized the parser to avoid overlap of surprisal's predictions with the word frequency effect.

We hand-crafted an ACT-R model for each selected sentence, closely following the approach taken by LV05. The model of each sentence was run for 30 simulations and a prediction of attachment time for every word was generated by averaging across all simulations. All ACT-R parameter values were kept the same as those used by LV05 except for activation noise. In LV05, five out of six simulations were carried out without switching on the activation noise. They also noted from preliminary experiments that adding activation noise did not change their results significantly. Since, one of ACT-R's standard assumptions is that there is always some noise added to the activation value of a chunk at each retrieval which permits modeling various kinds of memory errors, we set its value to 0.45 (this was one of the values used in Vasishth, Bruessow, Lewis, & Drenhaus, 2008).

Statistical Analyses The statistical analyses were carried out using linear mixed-effects models (Bates & Sarkar, 2007; Gelman & Hill, 2007) and the *Deviance Information Criterion* or DIC (Gelman & Hill, 2007, 524–527) was used to compare the relative goodness of fit between simpler and complex models. Linear models were fit for the following "early" and "late" eye movements measures:

SFD - fixation duration on a word during first pass if it is fixated only once

FFD - time spent on a word, provided that the word is fixated during the first pass

FPRT - the sum of all fixations on a word during the first pass

TRT - the sum of all fixations

FPSKIP - the probability of skipping the word during the first pass

We considered following baseline predictors in addition to surprisal and attachment time:

unigram - logarithm of token frequency of a word in Das Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts (DWDS) (Geyken, 2007; Kliegl, Geyken, Hanneforth, & Würzner, 2006)

bigram - logarithm of the conditional likelihood of a word given its left neighbor in DWDS (also called transitional probability)

word length - number of characters in conventional spelling

predictability - empirical predictability as measured in a Cloze task with human subjects (Taylor, 1953; Ehrlich & Rayner, 1981; Kliegl, Grabner, Rolfs, & Engbert, 2004)

Sentences and participants were treated as partially crossed random factors; that is, we estimated the variances associated with differences between participants and differences between sentences, in addition to residual variance of the dependent measures. For the analysis of FPSKIP (coded as a binary response for each word: 1 signified that a skipping occurred at a word, and 0 that it did not), we used a generalized linear mixed-effects model with a binomial link function (Bates & Sarkar, 2007; Gelman & Hill, 2007).

For each reading time analysis reported below, reading times more than three standard deviations away from the mean were removed before the analyses, excluding at most 1.7% of the data. Attachment time and all dependent measures except FPSKIP were log transformed. Word length, surprisal and attachment time were centered in order to render the intercept of the statistical models easier to interpret.

In the initial analyses, as expected, we found collinearity among the baseline predictors. Since collinearity can inflate the estimates of coefficients' standard errors leading to unreliable results, and can also lead to uninterpretable coefficient values, removal of collinearity between predictors was crucial before fitting the linear models for different fixation measures. For removing collinearity, we incrementally regressed each of these predictors against one or more baseline predictors and used residuals of the regressions as the predictors in the subsequent linear models. This was done in the following three steps:

1. Regression of unigram frequency against word length-
   uni.res = residuals (unigram ~ length)

2. Regression of bigram frequency against word length and residual unigram values obtained from step 1-
   bi.res = residuals (bigram ~ length + uni.res)

3. Regression of predictability against word length, residual unigram and bigram obtained from step 1 & 2-
   pred.res = residuals (predictability ~ length + uni.res + bi.res)

As a result, we had four baseline predictors — length, uni.res, bi.res, pred.res — which were completely non-collinear.

Table 1: Linear model coefficients, standard errors and t-values for surprisal, attachment time and interaction of attachment time and surprisal. An absolute t-value of 2 or greater indicates statistical signicance at $\alpha = 0.05$.

	Coef	SE	t-value
SFD			
surprisal	0.021722	0.001195	18
att. time	0.084338	0.013722	6
att. time:surprisal	0.048706	0.009518	5
FFD			
surprisal	0.018304	0.001032	18
att. time	0.062361	0.012361	5
att. time:surprisal	0.039307	0.008327	5
FPRT			
surprisal	0.021520	0.001217	18
att. time	0.056154	0.014221	4
att. time:surprisal	0.050750	0.009743	5
TRT			
surprisal	0.028558	0.001389	21
att. time	0.058249	0.016197	4
att. time:surprisal	0.055988	0.011128	5

Table 2: Linear model coefficients, standard errors and t-values for baseline predictors for TRT.

	Coef	SE	t-value
TRT			
length	0.031052	0.000949	33
uni.res	-0.023228	0.002322	-10
bi.res	-0.011984	0.000879	-14
pred.res	-0.006162	0.002752	-2

Table 3: Linear model coefficients, standard error, z-scores and p-values with FPSKIP as the dependent measure.

	Coef	SE	z-score	p-value
att. time	-0.51588	0.09401	-5.5	<0.001
surprisal	-0.18235	0.01000	-18.2	<0.001
att. time:surp	-0.12521	0.08067	-1.6	0.121

Table 4: Deviance Information Criterion values for simpler model (baseline predictors + surprisal) vs. more complex model (simpler model + attachment time).

	Simpler model	Complex model
SFD	8624.7	8576.5
FFD	9908.0	9873.1
FPRT	22606.0	22581.9
TRT	30695.5	30674.6
FPSKIP	36140.8	36111.5

Results & Discussion

The results of the mixed-effects models are summarized in tables 1 to 3. We observed significant main effects of both surprisal and attachment cost across ”early” as well as ”late” measures and also on FPSKIP. The coefficient for FPSKIP is negative reflecting the fact that the probability of fixating a word increases with increase in surprisal and retrieval cost. These results illustrate that surprisal as well as retrieval cost can account for variance in eye-tracking measures independent of baseline predictors (such as unigram and bigram frequency, word length, Cloze predictability, etc.). For comparison, coefficients of baseline predictors for TRT are listed in table 2; similar coefficient values were obtained for other reading time measures.

The interaction of attachment time and surprisal is significant for all measures except for FPSKIP (though even in this case the coefficient has the expected sign), which indicates that there is a disproportionate increase in reading difficulty when both surprisal and retrieval cost are high.

Table 4 compares the DIC values for simpler models (baseline predictors + surprisal) and complex models (baseline predictors + surprisal + attachment time). For all dependent measures the predictive error (DIC value) was lower in the more complex model that included attachment time, which means that the complex models should be preferred to the simpler ones.

Retrieval cost, surprisal and their interaction show effects on ”early” as well as ”late” measures. This suggests that structure-building and retrieval processes start very soon after lexical access begins.

Implications for eye movement models Besides the contribution to psycholinguistic theories, this work can contribute towards extending models of eye movement control such as E-Z Reader (Pollatsek, Reichle, & Rayner, 2006) and SWIFT (Engbert, Nuthmann, Richter, & Kliegl, 2005) which despite being the two most fully developed models of eye movements, do not incorporate any theory of language processing. The latest version of E-Z Reader (Reichle, Warren, & McConnell, 2009) makes an attempt in this direction by augmenting

the model with a post-lexical integration stage, named I. This stage is assumed to reflect all of the post-lexical processing like linking the word into a syntactic structure, generating a context-appropriate semantic representation, and incorporating its meaning into a discourse model. However, the amount of time to complete I, t(I), is independent of the language processing demands at that word; instead t(I) is sampled from a gamma distribution having a mean of 25 msec and standard deviation of 0.22. Models of sentence processing like the two evaluated here or, preferably, a systematic combination of them would offer a more realistic way of computing t(I). A similar approach of incorporating post-lexical processes can be taken in other eye movement models depending on the particular architecture of each model.

Conclusions

This work evaluated the combined contribution of two theories of sentence processing, viz., surprisal and cue-based retrieval theory. The two approaches capture different aspects of sentence processing, namely instantaneous probabilistic disambiguation and processing constraints due to memory retrievals. It was shown that when effects of these theories were combined together to predict eye movements measures, they emerged as significant predictors even when word length, n-gram frequency and Cloze predictability were taken into account. Moreover, they showed an over-additive effect on several eye movements measures. This needs to be taken into account in future models of sentence processing that integrate surprisal and retrieval costs. Also, models of eye movement could benefit from this work. Although the size of the evaluation corpus is small (total 32 sentences and 222 participants) and models of cue-base parsing were hand-crafted, this work serves as a first step towards developing a broad coverage model of sentence processing that combines the two processes – probabilistic disambiguation and memory retrieval – in a principled way.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2005). An integrated theory of mind. *Psychological Review*.

Bates, D., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes (R package version 0.9975-11) [Computer software].

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, *2*(1), 1-12.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *108*(2), 193-210.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*, 641-655.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). Swift: A dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777–813.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum (Ed.), *Idioms and Collocations: Corpus-based Linguistic, Lexicographic Studies.* London: Continuum Press.

Gibson, E. (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium.* Cambridge, MA: MIT Press.

Grodner, D., & Gibson, E. (2005). Some consequences of the serial nature of linguistic input. *Cognitive Science*, *29*(2), 261290.

Hale, J. (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* (pp. 1–8). Pittsburgh, PA.

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

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics.* MIT Press.

Kliegl, R., Geyken, A., Hanneforth, T., & Würzner, K. (2006). *Corpus matters: A comparison of German DWDS and CELEX lexical and sublexical frequency norms for the prediction of reading fixations.* Unpublished manuscript.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency and predictability effects of words on eye movements in reading. In *Eye movements and information processing during reading* (Vol. 16, p. 262-284). East Sussex: Psychology Press. (Special Issue)

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, *135*, 12-35.

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

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, *25*, 93115.

Lewis, R. L., & Vasishth, S. (2005). An activation-based

model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 1-45.

Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the e-z reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology*, *52*, 1 – 56.

Reichle, E., Warren, T., & McConnell, K. (2009). Using e-z reader to model effects of higher level language processing on eye-movements during reading. *Psychonomic Bulletin & Review*, *16*(1), 1– 21.

Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, *21*, 165–202.

Taylor, W. (1953). Cloze procedure: a new tool for measuring readability. *Journalism Quarterly*, *30*, 415–433.

Vasishth, S., Bruessow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*(4).