Surprising parser actions and reading difficulty

Marisa Ferrara Boston, John Hale
Michigan State University
USA
{mferrara,jthale}@msu.edu

Reinhold Kliegl, Shravan Vasishth
Potsdam University
Germany
{kliegl,vasishth}@uni-potsdam.de

Abstract

An incremental dependency parser's probability model is entered as a predictor in a linear mixed-effects model of German readers' eye-fixation durations. This dependency-based predictor improves a baseline that takes into account word length, *n*-gram probability, and Cloze predictability that are typically applied in models of human reading. This improvement obtains even when the dependency parser explores a tiny fraction of its search space, as suggested by narrow-beam accounts of human sentence processing such as Garden Path theory.

1 Introduction

A growing body of work in cognitive science characterizes human readers as some kind of probabilistic parser (Jurafsky, 1996; Crocker and Brants, 2000; Chater and Manning, 2006). This view gains support when specific aspects of these programs match up well with measurable properties of humans engaged in sentence comprehension.

One way to connect theory to data in this manner uses a parser's probability model to work out the *surprisal* or log-probability of the next word. Hale (2001) suggests this quantity as an index of psycholinguistic difficulty. When the transition from previous word to current word is low-probability, from the parser's perspective, the surprisal is high and the psycholinguistic claim is that behavioral measures should register increased cognitive difficulty. In other words, rare parser actions are cognitively costly. This basic notion has proved remarkably applicable across sentence types and languages (Park and Brew, 2006; Demberg and Keller, 2007; Levy, 2008).

The present work uses the time spent looking at a word during reading as an empirical measure of sentence processing difficulty. From the theoretical side, we calculate word-by-word surprisal predictions from a family of incremental dependency parsers for German based on Nivre (2004); these parsers differ only in the size k of the beam used in the search for analyses of longer and longer sentence-initial substrings. We find that predictions derived even from very narrow-beamed parsers improve a baseline eye-fixation duration model. The fact that any member of this parser family derives a useful predictor shows that at least some syntactic properties are reflected in readers' eye fixation durations. From a cognitive perspective, the utility of small k parsers for modeling comprehension difficulty lends credence to the view that the human processor is a single-path analyzer (Frazier and Fodor, 1978).

2 Parsing costs and theories of reading difficulty

The length of time that a reader's eyes spend fixated on a particular word in a sentence is known to be affected by a variety of word-level factors such as length in characters, n-gram frequency and empirical predictability (Ehrlich and Rayner, 1981; Kliegl et al., 2004). This last factor is the one measured when human readers are asked to guess the next word given a left-context string.

Any role for parser-derived syntactic factors

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 5–8,
Columbus, Ohio, USA, June 2008.

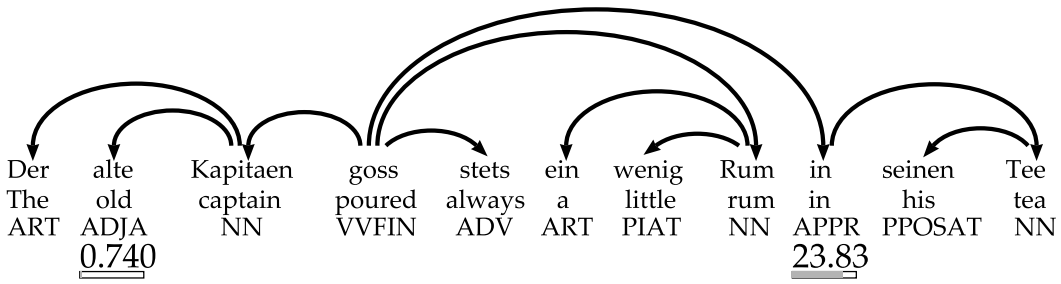

Figure 1: Dependency structure of a PSC sentence.

would have to go beyond these word-level influences. Our methodology imposes this requirement by fitting a kind of regression known as a linear mixed-effects model to the total reading times associated with each sentence-medial word in the Potsdam Sentence Corpus (PSC) (Kliegl et al., 2006). The PSC records the eye-movements of 272 native speakers as they read 144 German sentences.

3 The Parsing Model

The parser's outputs define a relation on word pairs (Tesnière, 1959; Hays, 1964). The structural description in Figure 1 is an example output that depicts this dependency relation using arcs. The word near the arrowhead is the *dependent*, the other word its *head* (or governor).

These outputs are built up by monotonically adding to an initially-empty set of dependency relations as analysis proceeds from left to right. To arrive at Figure 1 the Nivre parser passes through a number of intermediate states that aggregate four data structures, detailed below in Table 1.

σ	A stack of already-parsed unreduced words.
τ	An ordered input list of words.
h	A function from dependent words to heads.
d	A function from dependent words to arc types.

Table 1: Parser configuration.

The stack σ holds words that could eventually be connected by new arcs, while τ lists unparsed words. **h** and **d** are where the current set of dependency arcs reside. There are only four possible transitions from configuration to configuration. `Left-Arc` and `Right-Arc` transitions create dependency relations between the top elements in σ and τ, while `Shift` and `Reduce` transitions manipulate σ.

When more than one transition is applicable, the parser decides between them by consulting a probability model derived from the Negra and Tiger newspaper corpora (Skut et al., 1997; König and Lezius, 2003). This model is called Stack3 because it considers only the parts-of-speech of the top three elements of σ along with the top element of τ. On the PSC this model achieves 87.9% precision and 79.5% recall for unlabeled dependencies. Most of the attachments it gets wrong (Table 2) represent alternative readings that would require semantic guidance to rule out.

Error type	Amount
Noun attachment	4.2%
Prepositional Phrase attachment	3.0%
Conjunction	1.9%
Adverb ambiguity	1.8%
Other	1.1%
Total error	**12.1**%

Table 2: Parser errors by category.

To compare "serial" human sentence processing models against "parallel" models, our implementation does beam search in the space of Nivre-configurations. The number of configurations maintained at any point is a changeable parameter k.

3.1 Surprisal

In Figure 1 the thermometer beneath the German preposition "in" graphically indicates a high surprisal prediction derived from the dependency parser. Greater cognitive effort, reflected in reading time, should be observed on "in" as com-

pared to "alte." The difficulty prediction at "in" ultimately follows from the frequency of verbs taking prepositional complements that follow nominal complements in the training data. Equation 1 expresses the general theory: the surprisal of a word, on a language model, is the logarithm of the prefix probability eliminated in the transition from one word to the next.

$$\text{surprisal}(n) = \log_2 \left(\frac{\alpha_{n-1}}{\alpha_n} \right) \quad (1)$$

The prefix-probability α_n of an initial substring is the total probability of all grammatical analyses that derive $w = w_1 ... w_n$ as a left-prefix (Equation 2).

$$\alpha_n \quad = \sum_{d \in \mathcal{D}(G, wv)} \text{Prob}(d) \quad (2)$$

In a complete parser, every member of $\mathcal{D}$ is in correspondence with a state transition sequence. In the beam-search approximation, only the top *k* configurations are retained from prefix to prefix, which amounts to choosing a subset of $\mathcal{D}$.

4 Study

The study addresses whether surprisal is a significant predictor of reading difficulty and, if it is, whether the beam-size parameter *k* affects the usefulness of the calculated surprisal values in accounting for reading difficulty.

Using total reading time as a dependent measure, we fit a baseline linear mixed-effects model (Equation 3) that takes into account word-level predictors log frequency (*lf*), log bigram frequency (*bi*), word length (*len*), and human predictability given the left context (*pr*).

$$\log(TRT) = \quad (3)$$
$$5.4 - 0.02lf - 0.01bi - 0.59len^{-1} - 0.02pr$$

All of the word-level predictors were statistically significant at the α level 0.05.

Beyond this baseline, we fitted ten other linear mixed-effects models. To the inventory of word-level predictors, each of the ten regressions uniquely added the surprisal predictions calculated from a parser that retains at most *k*=1...9,100 analyses at each prefix. We evaluated the change in relative quality of fit due to surprisal with the *Deviance Information Criterion* (DIC) discussed in Spiegelhalter et al. (2002). Whereas the more commonly applied Akaike Information Criterion (1973) requires the number of estimated parameters to be determined exactly, the DIC facilitates the evaluation of mixed-effects models by relaxing this requirement. When comparing two models, if one of the models has a lower DIC value, this means that the model fit has improved.

4.1 Results and Discussion

Table 3 shows that the linear mixed-effects model of German reading difficulty improves when surprisal values from the dependency parser are used as predictors in addition to the word-level predictors. The coefficients on the baseline predictors remained unchanged (Equation 3) when any of the parser-based predictors was added.

Table 3 also suggests the returns to be had in accounting for reading time are greatest when the beam is limited to a handful of parses. Indeed, a parser that handles a few analyses at a time (*k*=1,2,3) is just as valuable as one that spends far greater memory resources (*k*=100). This observation is consistent with Brants and Crocker's (2000) observation that accuracy can be maintained even when restricted to 1% of the memory required for exhaustive parsing. The role of small *k* dependency parsers in determining the quality of statistical fit challenges the assumption that cognitive functions are global optima. Perhaps human parsing is boundedly rational in the sense of the bound imposed by Stack3 (Simon, 1955).

5 Conclusion

This study demonstrates that surprisal calculated with a dependency parser is a significant predictor of reading times, an empirical measure of cognitive difficulty. Surprisal is a significant predictor even when examined alongside the more commonly used predictors, word length, predictability, and n-gram frequency. The viability of parsers that consider just a small number of analyses at each increment is consistent with conceptions of the human comprehender that incorporate that restriction.

Model	Coefficient	Std. Error	t value	DIC
Baseline	-	-	-	144511.1
k=1	0.033691	0.002285	15	143964.9
k=2	0.038573	0.002510	15	143946.2
k=3	0.037320	0.002693	14	143990.4
k=4	0.041035	0.002853	14	143975.7
k=5	0.048692	0.002953	16	143910.9
k=6	0.046580	0.003063	15	143951.6
k=7	0.045008	0.003118	14	143974.4
k=8	0.042039	0.003165	13	144006.4
k=9	0.040657	0.003225	13	144023.9
k=100	0.029467	0.003878	8	144125.4

Table 3: Coefficients and standard errors from the multiple regressions using different versions of surprisal (baseline predictors' coefficients are not shown for space reasons). t values > 2 are statistically significant at $\alpha = 0.05$. The table also shows DIC values for the baseline model (Equation 3) and the models with baseline predictors plus surprisal.

References

H. Akaike. 1973. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Caski, editors, *2nd International Symposium on Information Theory*, pages 267–281, Budapest, Hungary.

T. Brants and M. Crocker. 2000. Probabilistic parsing and psychological plausibility. In *Proceedings of COLING 2000: The 18th International Conference on Computational Linguistics*.

N. Chater and C. Manning. 2006. Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10:287–291.

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

V. Demberg and F. Keller. 2007. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Manuscript, University of Edinburgh.

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

L. Frazier and J. D. Fodor. 1978. The sausage machine: a new two-stage parsing model. *Cognition*, 6:291–325.

J. Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of 2*$^{\text{nd}}$ *NAACL*, pages 1–8. Carnegie Mellon University.

D.G. Hays. 1964. Dependency theory: A formalism and some observations. *Language*, 40:511–525.

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

R. Kliegl, E. Grabner, M. Rolfs, and R. Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16:262–284.

R. Kliegl, A. Nuthmann, and R. Engbert. 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.

E. König and W. Lezius. 2003. The TIGER language - a description language for syntax graphs, Formal definition. Technical report, IMS, Universität Stuttgart, Germany.

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

J. Nivre. 2004. Incrementality in deterministic dependency parsing. In *Incremental Parsing: Bringing Engineering and Cognition Together*, Barcelona, Spain. Association for Computational Linguistics.

J. Park and C. Brew. 2006. A finite-state model of human sentence processing. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, pages 49–56, Sydney, Australia.

H. Simon. 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118.

W. Skut, B. Krenn, T. Brants, and H. Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97*, Washington, DC.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. 2002. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, 64(B):583–639.

L. Tesnière. 1959. *Eléments de syntaxe structurale*. Editions Klincksiek, Paris.