

Integration Costs on Auxiliaries? – a self-paced reading study using WebExp

Vera Demberg (vera@coli.uni-saarland.de)

Cluster of Excellence, Saarland University,
Campus C7.4, 66123 Saarbrücken, Germany

Abstract

In their evaluation of the integration cost component of Dependency Locality Theory on the Dundee Corpus, Demberg and Keller (2008) found no significant main effect of DLT integration cost on reading times, but suggested that this might be due to auxiliaries incurring some of the full verb’s integration cost and thus facilitating processing of the verb. This hypothesis however, has to date not been tested. The present paper fills this gap by reporting an experiment on subject vs. object relative clauses including auxiliaries, as well as by testing Demberg and Keller’s hypothesis directly on the Dundee Corpus.

A further contribution of this paper is methodological: we replicate experimental results on the subject vs. object relative clause asymmetry in a self-paced-reading experiment run remotely on the web using WebExp.

Keywords: Dependency Locality Theory, Relative Clause, Auxiliary, WebExp, Dundee Corpus, Self-paced reading, Eye-tracking

Introduction

Dependency Locality Theory (DLT), proposed by (Gibson, 1998; Gibson & Pearlmutter, 2000) is a theory of sentence processing which has received quite a lot of attention in the field of psycholinguistics, and has been argued to explain a range of phenomena including the SRC/ORC processing difficulty asymmetry, difficulty of centre embeddings, cases of processing breakdown, filler-gap dependencies, heavy NP shift and extraposition.

(Demberg & Keller, 2008) evaluated the integration cost component of DLT on an eye-tracked corpus of newspaper articles (Dundee Corpus; Kennedy and Pynte, 2005), and found that verbs which were preceded by nouns were read more slowly than verbs which were preceded by both auxiliaries and nouns. Demberg and Keller thus hypothesized that integration costs might not be incurred at the main verb (as predicted by DLT), but at the auxiliary, at which it should thus be possible to observe an integration cost effect.

They did however not test whether such an integration cost effect could indeed be detected on the auxiliaries. This paper fills this gap through two studies that test Demberg and Keller’s hypothesis: a self-paced reading experiment of matched and controlled subject and object relative clauses containing auxiliaries, as well as a corpus study analysing the auxiliaries across various syntactic constructions in the Dundee Corpus for an integration cost effect.

Background and Related Work

Dependency Locality Theory

An important component for quantifying processing difficulty in DLT is the so-called “integration cost”. Integration cost

(IC) measures the distance between a head its dependent in terms of new discourse referents (DR) that occur inbetween them. Figure 1 shows the dependencies for a subject relative clause (SRC) and an object relative clause (ORC). In the example, discourse referents are marked as either 0 (no new discourse referent) or 1 (new discourse referent). The dependency edge between *reporter* and the main clause verb *admitted* is annotated with “+2” to express that two discourse referents occur between these words (namely, *attacked* and *senator*). The interesting case when comparing the subject and object relative clause is the embedded verb *attacked*: the integration cost is 1 in the SRC case (just the cost of constructing the discourse referent), while the it is 3 in the case of the object relative clause. There is a cost of 1 for constructing the discourse referent, plus a cost of 2 for integrating the relative pronoun *who* at the trace **t**, at which point two discourse referents (*senator* and *attacked*) intervene. There are also integrations of *senator* and *attacked* and the trace and *attacked*, but no new discourse referents occur between them, so these integrations are cost-free.

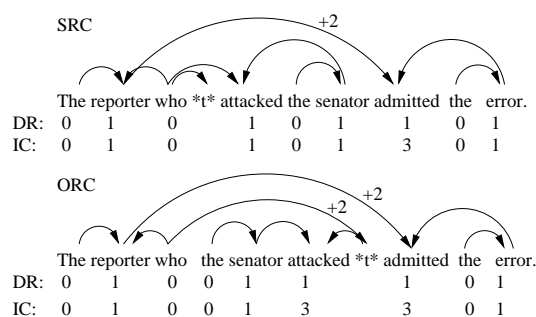


Figure 1: Dependency Locality Theory Integration Cost.

Auxiliaries in DLT

Most previous experimental studies on locality effects do not contain auxiliaries. An exception is Experiment 4 from (Warren & Gibson, 2002), which compares self-paced reading times of an object relative clause with a full or pronominal embedded subject NP vs. a complementizer clause with full or pronominal subject NP, see example (1).

- (1) a. Relative clause: The woman who **you/the boy had accidentally pushed** off the sidewalk got upset and decided to report the incident to the policeman standing nearby.
- b. Complement clause: The woman knew that **you/the boy had accidentally pushed** the girl

but gave him/you a long lecture anyway.

Warren and Gibson found that the reading times were longest in the full NP in ORC condition, which is also the one in which highest integration costs are expected. On the auxiliary, the difference between the ORC with full NP and the ORC with pronoun is significant. While this would be in line with an integration effect already on the ORC auxiliary, Warren and Gibson also point out the possibility that the longer reading times on the auxiliary might be a spill-over effect.

The present study seeks to directly investigate whether integration costs can be measured on auxiliaries.

Result from Demberg and Keller, 2008

Demberg and Keller’s (2008) evaluation of Dependency Locality Theory on the Dundee Corpus showed that there was no general positive correlation between DLT integration costs and reading times. They however looked at integration cost at verbs in more detail, and found that verbs which integrate an auxiliary and a nominal dependent exhibit a reduced estimated reading time compared to verbs that only integrate a nominal dependent, while there seemed to be an overall effect of increased reading time at verbs with more nominal dependents, see Table 1.

Demberg and Keller therefore suggested that the relevant integration cost might not be incurred at the main verb, but at the auxiliary itself, which might integrate nominal dependents and thus incur a non-zero integration cost (DLT assume that auxiliaries are cost-free). When the auxiliary would then

Table 1: First pass durations for verbs (with non-zero integration cost) in the Dundee Corpus: coefficients for verbs grouped by verbal dependents ($N > 20$) and their significance levels for a model fitted on residual reading times (with respect to a model including other predictors known to influence reading times). Abbreviations in the table refer to part of speech tags used by the Penn Treebank annotation: AUX: auxiliary, PRP: personal pronoun, NN: singular or mass noun, NNP: proper noun, singular, RP: particle, MD: modal, NNS: plural noun, RB: adverb, AUXG: auxiliary present participle, TO: preposition to, JJ: adjective, VBP: non-third person singular present verb. Table from Demberg and Keller (2008).

Dependents	Coeff	Signif	N
NNP-AUX-AUX	-62.41	**	21
NNS-AUX-AUX	-35.65	*	57
NNS-MD-AUX	-30.75	**	110
PRP-AUX-PRP-AUX	-29.72	***	184
NN-MD-AUX	-25.35	**	153
PRP-AUX	-22.64	***	700
PRP-AUX-RB	-21.75	*	133
AUXG	-20.26	*	121
NNP-AUX	-19.05	**	301
TO-PRP	-16.97	***	723
NNP	12.01	**	1372
NN-RB	22.26	*	127
NN-NNS	76.43	***	25
PRP-MD-PRP-MD-JJ	105.4	*	65

be integrated with the main verb, it would facilitate integration (hence the negative coefficient), as the main work of the integration of the nominal dependents has already happened at the auxiliary. They also point out that this explanation is compatible with syntactic theories such as Head-driven Phrase Structure Grammar (Pollard & Sag, 1994), which assume that auxiliaries inherit the subcategorization frame of the main verb, and that dependents are unified (integrated) into the subcategorization frame at the auxiliary. Demberg and Keller did however not test this hypothesis in their study, so the contribution of this paper is to fill this gap and test both in a controlled experiment and on the Dundee corpus whether the hypothesis that the verb’s integration cost can be measured on the auxiliary is true.

Experiment: Auxiliaries in Relative Clauses

As a first experiment, we chose to use a strictly controlled experimental setting in which we compare the processing of subject vs. object relative clauses including auxiliaries. The processing difference in subject vs. object relative clauses is well-established: Object relative clauses (as in (2-b)) are more difficult to process than subject relative clauses (1a) (King & Just, 1991), with increased reading times on the ORC embedded verb as opposed to the SRC embedded verb (Staub, 2010). Dependency Locality Theory (DLT; Gibson, 2000) accounts for this effect in terms of long-distance dependencies, see the explanation of this case in Section on Dependency Locality Theory.

We created 24 subject and object relative clauses with auxiliaries preceding the embedded verb, based on the experimental items from (Staub, 2010), see (2).

- (2) a. The mathematician who [AUX had] [v visited] [NP the chairman] found a solution to the problem.
 b. The mathematician who [NP the chairman] [AUX had] [v visited] found a solution to the problem.

Data Collection

We ran a self-paced reading experiment with 126 participants online, using WebExp, www.webexp.info (Keller, 1999), an experimental software that carries out psychological experiments over the internet. Keller, Gunasekharan, Mayo, and Corley (2009) demonstrate that response times collected with WebExp are sufficiently accurate to conduct reaction time experiments over the internet. Experiment 2 from (Keller et al., 2009) replicates results from a lab-based phrase-by-phrase self-paced reading experiment using the WebExp software.

Participants were recruited using Amazon Mechanical-Turk, which we used to create HITs linking to the WebExp experiment. In order to encourage participants to complete the whole WebExp experiment, the HIT also contained a field that required participants to fill in a password which was provided on the last screen of the WebExp experiment. We restricted the HITs to workers who were based in the USA, and

have a HIT approval rate $\geq 80\%$. In the instructions, we additionally required workers to only participate if they were native speakers of English. We successfully collected data from 126 participants¹ (approx. 60 per condition). Following recommendations in (Keller et al., 2009), we only allowed workers with a Windows or Linux operating system.

The experiment was programmed as a word-by-word self-paced reading experiment. Due to a limitation of WebExp, each sentence within a set of sentences to be randomized has to contain the same number of words. In addition to inserting auxiliaries, we therefore edited the items from (Staub, 2010) to conform to this format by adapting the length of the region following the relative clause.

Mixed-effects modelling

In order to test whether our manipulation of relative clause type has an effect on the reading times on the auxiliaries, we use linear mixed effects models from the R lme4 package (R. Baayen, 2008; R. H. Baayen, Davidson, & Bates, 2008). This type of model can be thought of as a generalization of linear regression that allows the inclusion of random factors as well as fixed factors. We treat subjects and items as a random factors, which means that our models contain an intercept term for each subject and each item, representing the individual differences among the subjects and differences between our items. Furthermore, we include random slopes under both subject and item for our predictor (relative clause type), essentially accounting for idiosyncrasies of a participant or item with respect to the predictor, such that only the part of the variance that is common to all participants and all items is attributed to the main effect for our predictor.

We excluded as outliers any reading times shorter than 100msec or longer than 1000msec.

Results

We found a significant effect of relative clause type on the auxiliary (AUX), the embedded verb (VB) and the determiner (DT), but not on the noun (NN) or the sentence’s main verb (VBM), see Figure 2.

Auxiliaries and verbs were read significantly faster in the subject relative clause condition than in the object relative clause condition. We furthermore found that SRC determiners were read more slowly than ORC determiners; a similar effect was found on the noun region in early reading time measures in Staub (2010), see Figure 3. Differences in terms of the location of the effect may be due to differences in self-paced reading vs. eye-tracking.

The faster reading times on the verb of the subject relative clause are in line with DLT integration cost, while the large and significant effect on the auxiliary seems to support also

¹There were some problems with WebExp, which sometimes failed to correctly transfer the collected data to our server. This resulted in loss of about 30% of our data. This problem had been observed by others and reported earlier; it appears to be independent of the operative system and browser used by workers. Workers whose data failed to transfer were paid normal rates regardless.

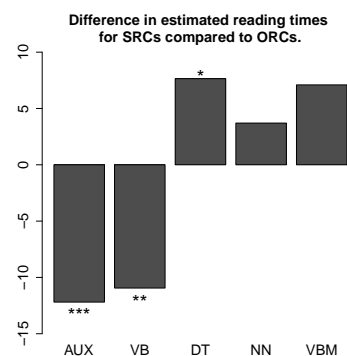


Figure 2: Coefficients for relative clause condition shown as a bar plot, a negative value at the auxiliary means that the auxiliary was read faster in the subject relative clause condition than in the object relative clause condition.

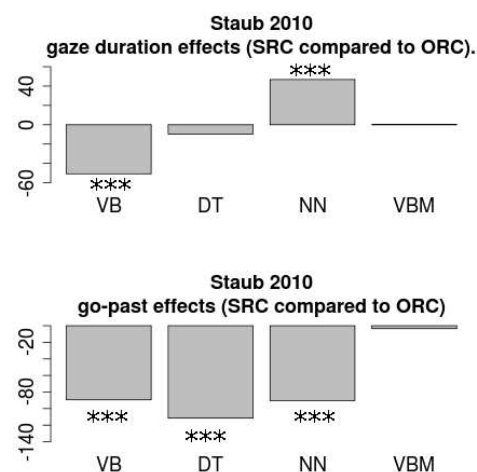


Figure 3: Experimental results from Staub, 2010 using the same materials without auxiliaries in an eye-tracking study.

an account of an integration cost effect being observable at the auxiliary. There is however a confound in such subject vs. object relative clause stimuli: the regions preceding the auxiliary (or the verb, in other studies) differ in that the SRC auxiliary is preceded by the relative pronoun, while the ORC pronoun is preceded by a noun.

We therefore ran a second analysis which takes into account spill-over effects. We approximate the spill-over effect that a word can cause by calculating with mixed-effects models the effect of one word’s log frequency on the following word’s reading times². In order to obtain an accurate estimate of typical spill-over effects, we ran this estimation on all words in our experiment, including fillers. used residual reading times for each of the words in our relative clause ma-

²While it would be possible to approximate the spill-over effect given not only the previous word but also the word before that, note that for our stimuli, only the previous word can plausibly explain the longer reading times on the ORC auxiliary. (Word *aux* – 2 in the ORC is shorter and more frequent than word *aux* – 2 in the SRC.)

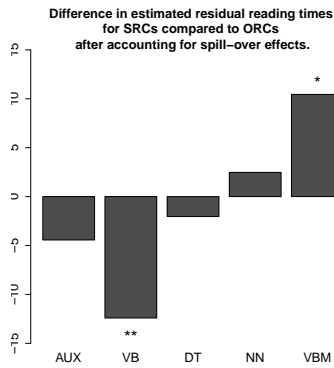


Figure 4: When taking spill-over effects into account, the effect on the auxiliary disappeared.

materials as a response variable and re-ran the earlier mixed-effects analysis on these residual reading times. The results then looked rather different: the facilitation effect on SRC auxiliaries, and difficulty effect on SRC determiners disappears when we account for spill-over effect, see Figure 4.

We furthermore find a significant effect of relative clause type on the reading times of the main verb, indicating that the main verb in the subject relative clause condition was read more slowly than the main verb in the object relative clause condition. This effect is puzzling as previous studies on self-paced reading rather seem to suggest a tendency for longer reading times on the ORC main verb, while reading times in Staub’s eye-tracking experiment were virtually identical.

Discussion

We find that increased reading times on auxiliaries can be explained in terms of spill-over effects, thus not supporting the hypothesis of (Demberg & Keller, 2008), who suggested that integration costs might occur at auxiliaries and facilitate integration at the verb. Instead, our findings support the original predictions of Dependency Locality Theory (Gibson, 2000). This result calls for a closer investigation of the Dundee corpus data to directly examine whether any integration cost effect is associated with auxiliaries in that data, see our Experiment 2.

In terms of experimental methodology, we provide evidence for the validity of self-paced reading using WebExp, by replicating the established relative clause asymmetry result on the embedded verb. Initial evidence that timing using WebExp is sufficiently accurate for self-paced reading studies is presented in (Keller et al., 2009). To the best of our knowledge, the present results are the first ones for word-by-word SPR using WebExp, as previous studies used much larger regions.

Corpus Study: Auxiliaries in Dundee Corpus

Data

For our data analysis, we used the Dundee Corpus (Kennedy & Pynte, 2005), an English language eye-tracking corpus

based on text from *The Independent* newspaper. The texts contain about 51,000 words and were read by 10 native speakers of English. The text was presented on a computer screen, five lines at a time at a line length of 80 characters.

Since the corpus data is not syntactically annotated, we parsed the entire corpus with the Stanford parser (Klein & Manning, 2003; De Marneffe, MacCartney, & Manning, 2006), which generates both a phrase structure parse and a dependency representation. We calculated DLT integration cost based on the top-ranked dependency output for each sentence. We evaluated our integration cost implementation using a short text that had been hand-annotated with integration cost values (Wu, Bachrach, Cardenas, & Schuler, 2010). This evaluation gives us an estimate of how well our automatic annotation tool performs, and also enables us to evaluate our new implementation based on the Stanford parser with an older implementation (Demberg & Keller, 2008) which was based on the MiniPar parser (Lin, 1998), see Table 2.

Table 2: Evaluation on a text of 770 words, manually annotated with integration costs.

Parser	% correct IC	correlation (Kendall) to manual annotation
MINIPAR	83%	$\tau = 0.77, p \ll 0.001$
Stanford	89%	$\tau = 0.84, p \ll 0.001$

We then automatically aligned each auxiliary with the automatically determined integration cost calculated for its governing main verb, in order to measure whether any effect of increased integration cost at the verb might be measurable on the auxiliary. In order to decrease noise in the data set, we excluded any cases in which further discourse referents occurred between the auxiliary and the verb, or where no verb could be found within a window of three words after the auxiliary. We furthermore excluded contractions (e.g. *we’ll*).

Methods

We analysed the data using linear mixed effects models. Because the corpus data are not as closely controlled as the experimental data from the first study, and because the methodology differs (eye-tracking here vs. self-paced reading in experiment 1), we run mixed-effects models with a larger range of different predictors, including the length of a word in characters `WordLength`, its log frequency `WordFreq`, a flag indicating whether the previous word was fixated `PrevFix`, the frequency of the preceding word to account for spill-over effects `PrevFreq`, forward and backward transitional probabilities `ForwTransProb` and `BackwTransProb`, the word number within the sentence `WordNo`, the fixation landing position in relation to word length `LandPos`, the launch distance of the saccade `LaunchDist`, the surprisal³ at the word `Surprisal` as well as the verb’s integration cost `IntegCost`.

³Surprisal was calculated using the Roark parser (Roark, Bachrach, Cardenas, & Pallier, 2009).

As response variables, we use four different measures of reading times: first fixation duration (duration of the first fixation, if any on the first pass through the sentence from left to right), first pass duration (sum of duration of fixations during first pass reading on a word before leaving the word), total duration (sum of the durations of all fixations on a word) and go past times (time spent between the first fixation in first pass reading on a word and first leaving it to the right).

We only analysed auxiliaries which had received at least one fixation. Before fitting LME models, we applied outlier removal: we computed the mean reading time (over all items and participants), and then removed all data points which deviated more than two standard deviations from the mean. Outliers can affect the results of analyses on the Dundee corpus, as (Roland, Mauner, O'Meara, & Yun, 2012) show. Furthermore, this way of trimming the data also reduces the long tail of the reading time distribution, resulting in a distribution that is closer to normal. This left us with 1257 data points.

Results

None of our models support the hypothesis that higher integration costs at the verb increase reading times at the auxiliary preceding the verb. In none of the reading time models did the verb's integration cost come out as a significant predictor of reading times on the auxiliary, see for example the best fitting regression model for first pass times in Table 3. When we add the verb's integration cost as a predictor, model fit is not significantly improved ($p \approx 0.75$), and we even get a negative coefficient ($\beta = -0.33, t = -0.293$), so there is not much reason to believe at this point that the failure of finding a positive significant effect would simply be due to an insufficient number of data points.

Table 3: Final model for first pass times on auxiliaries, showing that reading times are longer when word length increases, and shorter when the previous word was fixated or was a highly frequent word.

Predictor	Coeff	Std. Err	t value	Signif
(Intercept)	230.52	9.44	24.40	***
WordLength	5.46	2.40	2.27	*
PrevFreq	-5.18	1.50	-3.45	**
PrevFix	-25.19	4.04	-6.23	***
LandPos	-11.98	15.02	-0.79	
Surprisal	2.00	1.33	1.50	
WLen:LandP	-9.02	6.13	-1.47	

To more closely inspect the relationship between DLT integration cost and reading times and understand where the negative coefficient comes from, we ran a generalized additive model with a spline ($k=30$) for the verb's integration cost as a predictor and the auxiliaries' first pass times as the response variable. As can be seen in Figure 5, there is indeed a negative trend for increasing integration cost, at least for auxiliaries with the most common integration cost values one to

three. (An integration cost value of zero can only occur if the main verb of the sentence is a copula.)

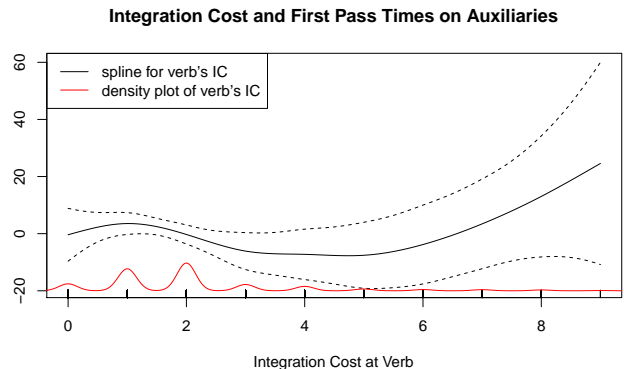


Figure 5: Spline plot for the verb's integration cost fitting the first pass times on the auxiliary.

Overall Discussion

Neither the results from the experiment nor the results from the corpus study support the hypothesis suggested in (Demberg & Keller, 2008), that integration may already happen at the auxiliary and costs of such an integration would be measurable in reading times. The results from the corpus study are in line with the results from the closely controlled relative clause experiment. While the experiment compared auxiliaries in a specific syntactic construction, the corpus study complements the first experiment in that it includes auxiliaries from many different syntactic constructions. It also shows that the result can be replicated for contextualized reading in more naturalistic conditions.

In both studies, we took care to account for spill-over effects from previous words, and such effects indeed turned out to be important in both studies: in the relative clause study, the interpretation of results changes completely when taking into account spill-over effects, and in the corpus study, the variables capturing overspill effects significantly improve model fit.

Getting back to the fundamental question underlying these studies, these results lead to the following hypotheses about integration cost at auxiliaries:

- Auxiliaries help integration at the verb in a way which is not directly reflected in their reading times.
- Auxiliaries do in fact not help integration at the verb, and DLT integration cost, despite showing stable effects in experimental settings, does hence not have much general explanatory power for data of properly contextualized utterances such as those occurring in a corpus.
- The integration cost estimate using automatic parsing tools is not accurate enough (in particular due to shortcomings in dealing with traces).

The last concern is unlikely to be valid, however, as it only applies to the corpus study, and doesn't explain failure to find an effect in the first experiment. Also, we have taken great care to re-implement integration cost based on a state-of-the-art parser, and have used heuristics to account for traces.

Question a) could be addressed experimentally by manipulating constructions with respect to the presence of an auxiliary (the effects of spill-over could be diminished by using an adverb before the verb such that the word immediately before the verb is always the same).

Conclusions

The contribution of this paper is two-fold. We find that increased reading times on auxiliaries in subject vs. object relative clauses can be explained in terms of spill-over effects, thus do not support the hypothesis of (Demberg & Keller, 2008), who suggested that integration costs might occur at auxiliaries and facilitate integration at the verb. Instead, our findings support the original predictions of Dependency Locality Theory (Gibson, 2000). These results are further supported by a corpus study of auxiliaries from the Dundee Corpus. We can not find any significant effect of the verb's integration cost on reading times at the auxiliary in any of our reading time measures. Given our evidence about the lack of any detectable integration cost effect on auxiliaries, we can no longer explain away the lack of an overall positive effect of integration costs on the verbs from the Dundee corpus as being due to the presence of auxiliaries.

Our second contribution is methodological: the self-paced reading study (Experiment 1) provides evidence for the validity of word-by-word self-paced reading via WebExp, by replicating the established relative clause asymmetry result on the embedded verb. While initial evidence that timing using WebExp is sufficiently accurate for self-paced reading studies is presented in (Keller et al., 2009), the present results are, to the best of our knowledge, the first ones for word-by-word self-paced-reading using WebExp.

Acknowledgments

I want to thank Silas Weinbach for his contributions to this work in programming the WebExp and Mechanical Turk experiment, as well as Jesus Calvillo and Asad Sayeed for their contributions to the new integration cost component based on the Stanford parser.

References

- Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge Univ Pr.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*.
- De Marneffe, M., MacCartney, B., & Manning, C. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of Irec* (Vol. 6, pp. 449–454).
- Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109, 193–210.
- Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68, 1–76.
- Gibson, E. (2000). Dependency locality theory: A distance-dated theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 95–126). Cambridge, MA: MIT Press.
- Gibson, E., & Pearlmutter, N. (2000). Distinguishing serial and parallel parsing. *Journal of Psycholinguistic Research*, 29(2), 231–240.
- Keller, F. (1999). *Webexp: A java toolbox for web-based psychological experiments*. University of Edinburgh, Human Communication Research Centre.
- Keller, F., Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing accuracy of web experiments: A case study using the webexp software package. *Behavior Research Methods*, 41(1), 1–12.
- Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, 45, 153–168.
- King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30, 580–602.
- Klein, D., & Manning, C. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting on association for computational linguistics-volume 1* (pp. 423–430).
- Lin, D. (1998). An information-theoretic definition of similarity. In J. W. Shavlik (Ed.), *Proceedings of the 15th international conference on machine learning* (pp. 296–304). Madison, WI: Morgan Kaufmann.
- Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.
- Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009, August). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 324–333). Singapore.
- Roland, D., Mauner, G., OMeara, C., & Yun, H. (2012). Discourse expectations and relative clause processing. *Journal of Memory and Language*.
- Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, 116, 71–86.
- Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, 85(1), 79–112.
- Wu, S., Bachrach, A., Cardenas, C., & Schuler, W. (2010, July). Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1189–1198). Uppsala, Sweden.