Eye-tracking Evidence for Integration Cost Effects in Corpus Data

Vera Demberg (v.demberg@sms.ed.ac.uk) and
Frank Keller (keller@inf.ed.ac.uk)
School of Informatics, 2 Buccleuch Place
Edinburgh EH8 9LW, UK

Abstract

We tested the predictions of Dependency Locality Theory (DLT), a theory of linguistic processing complexity, against reading time data extracted from a large eye-tracking corpus. DLT predicts differences in processing complexity for subject and non-subject relative clauses. We found elevated reading times on two distinct regions of these relative clauses, in line with the complexity effects predicted by DLT. We also found that transitional probability has an effect on reading time in these two regions, independent of the DLT effect. We argue our approach provides an important new way of testing sentence processing theories by evaluating them against reading data obtained from an eye-tracking corpus of naturally occurring text.

Keywords: sentence processing, processing complexity, eye-tracking, linguistic corpora, relative clauses

Introduction

Research on human sentence processing has traditionally focused on syntactic ambiguity, based on the observation that certain locally ambiguous constructions pose difficulty for the human sentence processor. Such difficulty manifests itself typically in the form of increased processing time (e.g., elevated reading times on the disambiguating region).

While disambiguation is an important source of difficulty in human sentence processing, such difficulty can also arise in unambiguous sentences. A classic example are relative clauses, which have been investigated extensively in the literature on syntactic processing difficulty. Experimental results show that English subject relative clauses (SRCs) as in (1-a) are easier to process than non-subject relative clauses (NSRCs) as in (1-b). Experimentally, this difficulty is evidenced by the fact that reading times for the region R1 in the SRC are lower than reading times for the corresponding region R3 in the NSRC (King & Just, 1991).

(1) a. The reporter who $[\text{attacked}]_{R1}$ the senator admitted the error.
b. The reporter who $[\text{the}]_{R2}$ senator $[\text{attacked}]_{R3}$ admitted the error.

Findings such as these have motivated processing theories that do not rely on ambiguity resolution, but instead capture the complexity involved in computing the syntactic dependencies between the words in a sentence. The most prominent such theory is Dependency Locality Theory (DLT), proposed by Gibson (1998, 2000). DLT not only captures a the SRC/NSRC asymmetry, but also accounts for a wide range of other complexity results, including processing overload phenomena such as center embedding and cross-serial dependencies.

While DLT has been validated against a large range of experimental results, it has not been shown that it can also successfully model complexity phenomena in naturally occurring text. It is possible that complexity effects observed in carefully controlled lab experiments are rare or absent in naturalistic data such as those found in corpora. The present paper aims to test DLT's predictions on the Dundee Corpus, a large corpus of newspaper text for which the eye-movement record of 10 participants is available. From this corpus, a range of eye-tracking measures can be computed, but the results hold for naturalistic, contextualized text, rather than for isolated example sentences manually constructed by psycholinguists.

In what follows, we will present two studies on the Dundee Corpus that test DLT's predictions for relative clauses, for two different regions of analysis. We compare our results against a baseline model that does not compute processing complexity directly, but that instead relies on the transitional probability between words.

Background

Dependency Locality Theory

According to Gibson's (1998; 2000) Dependency Locality Theory, processing complexity is associated with the cost of the computational resources consumed by the processor. Two distinct cost components can be distinguished: the (i) *integration cost* associated with integrating new input into the structures already built at a given stage in the computation, and (ii) the *memory cost* involved in the storage of parts of the input that may be used in parsing later parts of an input. Here, we will focus on integration cost, as "reasonable first approximations of comprehensions times can be obtained from the integrations costs alone, as long as the linguistic memory storage used is not excessive at these integration points" (Gibson, 1998, 19f). Integration cost is defined as follows:

(2) **Linguistic Integration Cost**
The integration cost associated with integrating a new input head h_2 with a head h_1 that is part of the current structure for the input consists of two parts: (1) a cost dependent on the complexity of the integration (e.g. constructing a new discourse referent); plus (2) a distance-based cost: a monotone increasing function I(n) energy units (EUs) of the number of new discourse referents that have been processed since h_1 was last highly activated. For simplicity, it is assumed that I(n) = n EUs. (Gibson, 1998, 12f)

According to this definition, integration cost is dependent on two factors. First, the type of element to be integrated matters: new discourse referents (e.g., indefinite NPs) are assumed to involve a higher integration cost than old/established discourse referents, identified by pronominals. Second, integration cost is sensitive to the distance between the head being integrated and the head it attaches to, where distance is calculated in terms of intervening discourse referents.

As an example, consider the subject vs. non-subject RC example in (1). At the embedded verb region in the SRC (Region R1), two integrations take place: the gap generated by the relative pronoun *who* needs to be integrated with the verb. The cost for this is I(0), as zero new discourse referents have been processed since the gap was encountered. In addition, the embedded verb *attacked* needs to be integrated with its preceding subject, an integration which crosses one new discourse referent (the embedded verb itself), leading to a cost of I(1). The total cost at region R1 is therefore I(0) + I(1).

In the NSRC (Region R3), the integration cost is I(2) for the integration of the gap generated by the relative pronoun, as two new discourse referents (*the senator* and *attacked*) intervene between the gap and the embedded verb. In addition, the integration of the verb with its subject *the senator* consumes I(1) energy units, as one new discourse referent has been processed, viz., *attacked* itself. The total cost for R3 in the NSRC is therefore I(2) + I(1). So overall, DLT predicts that R3 is more difficult to process than R1.

It is also interesting to consider the DLT predictions for another region, viz., the word immediately following the relative pronoun. In the SRC case, this region is again R1, the verb *attacked*, with a cost of I(0) + I(1). In the NSRC case, however, a noun phrase follows the relative pronoun, and the relevant region is R2, the word *the*, which causes an integration cost of I(0), as no new discourse referents have been processed since *the* was encountered. Hence DLT predicts that R1 is more difficult to process than R2.

The following summarizes the DLT predictions for SRCs and NSRCs (see Gibson 1998, 20f):

(3) The reporter who attacked the senator admitted
– I(0) I(0) I(0)+I(1) I(0) I(0)+I(1) I(3)
the error.
I(0) I(0)+I(1)

(4) The reporter who the senator attacked admitted the
– I(0) I(0) I(0) I(0) I(1)+I(2) I(3) I(0)
error.
I(0)+I(1)

In what follows, we will compare reading times for SRCs and NSRCs in an eye-tracking corpus for the embedded verb region (R1 vs. R3) and for the post-relative pronoun region (R1 vs. R2). We will also measure reading times on the relative pronoun; here, DLT does not predict any differences in processing difficulty.

Transitional Probability

It is well-known that reading times in eye-tracking data are influenced not only by high-level, syntactic variables but also by a number of low-level variables that have to do with the physiology of reading (see McDonald & Shillcock 2003b for a review). These variables include word frequency (more frequent words are read faster), word length (shorter words are read faster) and the landing position of the eye on the word. Together with variation between readers, these variables account for a sizable proportion of the variance in the eye-movement record.

Recently, it has also been shown that information about the sequential context of a word can influence reading times. In particular, McDonald & Shillcock (2003b) present data extracted from an eye-tracking corpus (a smaller corpus than the Dundee corpus used here) that show that forward and backward transitional probabilities are predictive of first fixation and gaze durations: the higher the transitional probability, the shorter the fixation time. By *forward transitional probability* McDonald & Shillcock (2003b) refer to the conditional probability of a word given the previous word $P(w_n|w_{n-1})$, while the *backward transitional probability* is the conditional probability of a word given the next word $P(w_n|w_{n+1})$. These corpus results are backed up by results demonstrating the role of forward transitional probabilities in controlled reading experiments (McDonald & Shillcock 2003a; but see Frisson et al. 2006, who equate transitional probability and Cloze predictability).

Given these findings, transitional probability provides a potential alternative explanation for reading time effects in corpus data. For example, in (1), the difference between R1 and R3 could be simply due to an effect of forward transitional probability: if $P(attacked|who)$ is larger than $P(attacked|senator)$, then we predict that R1 is read more quickly than R3, which is the same prediction that the DLT makes. We will therefore include forward transitional probability in the corpus analyses presented below.

Experiment 1: Embedded Verb Region

The aim of this experiment was to test a key prediction of DLT, viz., that subject and non-subject relative clauses differ in the amount of difficulty encountered in the verb region (regions R1 and R3 in (1)).

Method

Data For our data analysis, we used the Dundee Corpus (Kennedy et al., 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 a state-of-the-art parser (Charniak, 2000). We checked parsing reliability for our target construction (relative clauses) on the 23rd section of the Wall Street Journal and found recall to be 96% and precision to be 92%. In the Dundee Corpus, we found a total of 434 relative clauses headed by *who*, *which*, or *that*. Since each of the items was read by the 10 subjects, this provides us with 4340 data points in total. However, we excluded some of the data points according to the criteria described in the following section.

Selection Criteria From the 4340 relative clauses, we automatically extracted the embedded verb (the verb heading the relative clause). In relative clauses with auxiliaries or modals, we extracted the main verb of the relative clause, because this is where integration cost occurs. In the case of predicative constructions, we extracted the inflected form of the predicative verb *be*.

We excluded all the data points where the critical region (the embedded verb) was the first or last word of the line, and also all cases where the verb was followed by a any kind of punctuation. This eliminates wrap-up effects that occur at line breaks or at the end of sentences. Furthermore, we excluded all data points that were in a region of four or more adjacent

Pronoun	SRC	NSRC	Proportion of NSRC
that	150	18	10.7%
which	86	39	31.7%
who	137	4	2.8%
Total	373	61	14%

Table 1: Frequency of relative clause types in the Dundee eye-tracking corpus.

words that had not been fixated, since such regions were either not read by the participant or subject to data loss due to tracking errors. We computed the reading times for regions R1 and R3 for each item and each subject (a total of 3007 data points).

Independent Variables Each data point was associated with eight variables. These were the identity of the relative pronoun (*who*, *which*, or *that*), the type of the relative clause (SRC or NSRC), word length, the logarithm of the word frequency (estimated from the British National Corpus, BNC), the word's part of speech (POS), the logarithm of the forward transitional probability ($P(w_n|w_{n-1})$, where w_n is the verb; also estimated from the BNC), the word landing position, and the subject ID. The following POS tags occurred: AUX, MD, VB, VBP, VBN, VBG, VBD and VBZ (the Penn Treebank POS tag set was used, see Marcus et al. 1993).

There are a number of well-known correlations between the independent variables: short words are usually more frequent than long words, the fixation landing position depends on word length, the transitional probability and the frequency of a word are positively correlated. As Table 1 shows, the relative clause types were furthermore distributed differently for the three pronouns, and thus partially correlate with RC types.

Dependent Variables Each word in the data set is associated with the following eye-tracking measures: first fixation duration, total fixation duration, and a binary value that marks whether a word was fixated or skipped.

The *first fixation duration* of a region is the time that was spent on the first fixation on that region before any word further to the right was fixated. First fixation duration is zero if the region was first skipped and then regressed to later. The *first pass duration* is similar to first fixation duration, the difference being that all fixations on a word that occurred before any word to the right was fixated are summed up. Finally, the *total fixation duration* is the sum of the durations of all fixation on a region.

Each of these measures was taken as the dependent variable in a separate regression analysis. Because there is a fundamental difference between fixated and skipped words (i.e., it is not easy to justify why a skipped word would be interpretable on a linear scale (its reading time is 0) and comparable to fixated words), we performed linear regressions on the reading times for the fixated verbs (1886 verbs for first fixation durations, 2220 verbs for total fixation durations), and a separate logistic regression (with dependent variable fixated vs. skipped) for whole set of 3007 verbs.

Regression Procedure For each of the continuous dependent variables (total time, first fixation, first pass), we built separate linear mixed effect models that included the eight independent variables (Pinheiro & Bates, 2000) (aka hierarchical linear regression models, Richter (2006)). To make sure effects were stable across different modeling techniques, we ran both a linear mixed effect model that included SUBJ (the subject) as a random effect and also performed separate regressions for each of the 10 subjects and tested whether the coefficients for these models were reliably different from zero using a t-test (as suggested in Lorch & Myers 1990, method 3). Minimal models were obtained by entering all of the independent variables and all possible binary interactions between them into the model and then simplifying the model by comparing Akaike Information Criterion values. (The AIC is a measure that optimizes model fit by taking into account the amount of variance explained as well as the number of degrees of freedom.)

For the binary dependent variable (skipping), we ran a logistic regression model, using the same methods as for the linear regression.

Results

Linear Regression for Fixated Words We fitted a mixed effects regression model as specified above to the data. The results show a significant main effect of relative clause type ($p < 0.001$) for R1 and R3: SRC verbs were read more quickly than NSRC verbs (see Table 2). We also found a significant interaction between RC type and word frequency. The word frequency effect by itself is well known: frequent words are read faster than infrequent ones. The interaction between word frequency and relative clause type reflects the fact that in our data, the frequency effect was more pronounced in non-subject relative clauses than in subject relative clauses (hence the positive coefficient for the interaction, which weakens the frequency effect). The POS tag was no significant predictor for reading time on this region, presumably because its contribution is already explained by length and frequency and their interaction.

We also found effects for word length (longer words take longer to read), and transitional probability (words with high transitional probability are read faster than words with low transitional probability). This effect occurred in addition to the RC type effect, which means that longer reading times on the non-subject relative clause verbs cannot simply be explained by a lower predictability of the word, but suggests that the linguistic structure makes a distinct contribution. Two more interactions were significant: the interaction between word length and landing position on the word, as well as an interaction between word frequency and word length (short words are typically more frequent than longer ones).

Our model explains a reasonably large proportion of the variance in the data, the value for adjusted R-squared (which also takes into account the number of degrees of freedom) is 15.6%.

The findings for first fixation duration and first pass duration are almost identical. The main difference between those early measures and total reading times is that transitional probability and word landing position do not come out as significant predictors for first fixation and first pass times.

Logistic Regression for Skipped Words Skipping probabilities are almost identical for subject and non-subject relative clauses: they amount to about 36% for first pass skipping (i.e., the word is skipped before a word to the right is fixated)

Predictor	Coeff.	Sign.
(Intercept)	263.42	***
RC type-SRC	-177.04	***
Length	21.47	***
Word landing position	6.39	
Logarithmic frequency	-11.66	**
Transitional probability	-24.73	***
Length:landing position	-2.94	***
Log. freq:length	2.65	***
RC type:log. freq	18.65	***

$^{**}p < 0.01$, $^{***}p < 0.001$

Table 2: Regression coefficients and their significance levels for a minimal model of total reading time for the embedded verb region.

and 25% for total skipping (i.e., the word is never fixated). We ran a logistic regression for first pass skipping probabilities. The significant predictors for word skipping were transitional probability, word frequency, and word length.

Discussion

Our results provide evidence for DLT, which predicts that verbs in SRCs are processed more quickly than verbs in NSRCs, due to lower integration costs. In addition, we also find a significant effect of forward transitional probability in this region. Since the inclusion of the transitional probability factor into the model did not cause the RC type effect to disappear, we conclude that these factors explain different proportions of the variance in reading times, and that the two effects are largely independent (the correlation coefficient between transitional probability and RT type predictors is only -0.073).

As expected, a significant proportion of the data is also explained by low-level factors such as length, frequency, and fixation landing position and their interactions. As a single predictor, RT type accounts for about 3% of the variance, and RT type together with its interaction with frequency accounts for 10.5% of the variance. On the other hand, transitional forward probability explains 7.8% of the variance by itself. The low-level effects length, word landing position, word frequency and their interactions account for 14.4% of the variance. All of these numbers refer to regressions with subject as an error term.

Experiment 2: Relative Pronoun Regions

The aim of this experiment was to test a second prediction of DLT with respect to the processing complexity of relative clauses: SRCs should incur a higher processing cost than NSRCs on the word following the relative pronouns (regions R1 and R2 in (1)). In addition to comparing reading times on R1 and R2, we also tested for effects on the relative pronouns (where DLT predicts an SRC/NSRC difference, see (3) and (4)), and on the second word following the relative pronoun, where spillover effects from R1 and R2 can be expected.

Method

Data and Procedure We used the same relative clause data from the Dundee Corpus as in the first experiment. Also the regression technique was the same.

Selection Criteria The relative pronoun and the first two words immediately following it were extracted from the 4340 relative clauses. As in the first experiment, all data points where the critical region (any of the relative pronoun or the two following words) was located the beginning or end of the line when presented on the screen were removed from the data set, as well as all critical regions that included words with any kind of punctuation. Again, we excluded all data points that were in a region of four or more adjacent words that had not been fixated, and all pronouns that had auxiliaries attached to them (e.g. *that'll*, *who'd*). We computed the reading times on for regions R1 and R2 and the relative pronoun for each item and each subject (3067 data points in total).

Independent Variables The relative pronoun had the following variables associated with it: pronoun identity (*who*, *that*, *which*), subject ID, word length, fixation landing position, logarithm of the word frequency, logarithm of the transitional probability, and RC type. The first and second word following the relative pronoun were each associated with the following variables: word length, logarithm of the word frequency, POS tag of the word, transitional probability, and landing position. Furthermore, the information from the relative pronoun and from the other word in the critical region are also entered into the regression. In the tables, any variables that refer to the first word are marked '.1' while all those that refer to the second word are marked '.2'.

In the critical region, POS tag and RC type are strongly associated: the words that follow the pronoun in the non-subject RC are always noun phrases, while SRCs begin with verb phrases. Thus, the length and frequency distributions of the words in R1 and R2 are quite different: The first word of the NSRC is often a short and frequent determiner or personal pronoun, whereas SRCs begin with auxiliaries, modals or main verbs (see Table 4). For a list of the POS that occur for both RC types, see Table 3. Also, the POS tags of the first and second word of the RC depend on each other, since they are often part of the same constituent (NP or VP respectively).

Dependent Variables Again, each word in the critical region is associated with the following measures: first fixation duration, first pass duration, total fixation duration, and a binary value that marks whether a word was fixated or skipped. Each of these measures was taken as the dependent variable in a separate regression analysis.

Results

Relative Pronoun We calculated a minimal model (according to the AIC measure) that explains $\approx 7\%$ of the variance. The best predictors for reading time in this model are RC type ($p = 0.04$), fixation landing position, transitional probability from the previous word to the pronoun, transitional probability from the pronoun to the next word, and pronoun identity. Furthermore, we found interactions between fixation landing position and pronoun identity (which also coincides with word length), as well as between pronoun identity and transitional probability.

In a single predictor analysis, relative pronouns were read more quickly in the SRC condition than in the NSRC condition ($p < 0.001$), but this effect was more extreme for the relative pronouns *which* and *who* than for *that*, which is read fast in the NSRC condition as well. A possible explanation for this effect is that the word sequence *that NP* is more frequent than *which/who NP* due to the ambiguity of *that*. We found no general effect for RC type in first fixation and first pass measures in the pronoun region, but also the same effect of faster reading of *that* in the NSRC condition (although pronoun frequency and transitional probabilities were included as independent variables in the model).

Skipping Skipping of the relative pronoun is more frequent in the SRC condition than in the NSRC condition: first pass skipping probability was 60% for SRCs but only 45% for NSRCs. A similar contrast was found in total skipping, which was 46% for SRCs and 33% for NSRCs. We investigated a number of hypotheses to explain this early skipping effect:

1. Relative pronouns have different distributions for SRCs and NSRCs: *who* typically occurs with SRCs, and may be skipped more often as it is shorter than the other pronoun. We would then expect pronoun type to be a good predictor for skipping probability.

2. In SRCs, the first word after the relative pronoun is on average longer than the first word of an NSRC. Low level perceptual processes might thus cause saccades to the longer word directly, skipping the relative pronoun. We would then expect the length of the next word to be a good predictor for skipping.

3. SRCs and NSRCs might differ in predictability from the word before the relative pronoun. The more predictable the relative pronoun is, the more probable it is to be skipped. We would therefore expect the pronoun's transitional probability to be a good predictor for skipping.

Our data support hypothesis 2: For both regression methods, skipping is significantly predicted by the length of the first word of the relative clause: The longer that word, the higher the probability of the relative pronoun to be skipped. Transitional probability was not a significant predictor, and pronoun identity was significant according to method 3 from Lorch & Myers (1990), but not according to the mixed effects method.

However, RC type persists as a significant predictor ($p = 0.01$) for skipping even under this alternative explanation. This indicates that low level processes involving word length cannot fully explain the skipping of relative pronouns, and that the effect of RC type should be a topic for future research.

Post-Relative Pronoun The significant predictors for total reading times for the first word after the relative pronoun are frequency and length of that word, as well as the landing position, especially in interaction with word length. We also found that word length and frequency of the following word were significant predictors, as well as RC type and the word's POS tag (see Table 3).

POS tag of the first word and RC type were entered as an interaction into the regression, because the POS tags form two exclusive sets with respect to their RC type. We found that after accounting for the variance that is due to frequency and length effects, the critical region was generally read more quickly in the NSRC condition than in the SRC condition, see the coefficients in Table 3.[1]

Predictor	Coeff.	Sign.
(Intercept)	190.73	**
Landing position.1	9.95	*
Logarithmic frequency.1	-0.02	
Length.1	30.63	***
Logarithmic frequency.2	-2.55	
Length.2	-2.92	
Log. freq.1:length.1	-1.44	.
Landing pos.1:length.1	-3.20	***
POS.1-DT:RC type-NSRC	4.97	
POS.1-EXAUX:RC type-NSRC	-50.50	
POS.1-JJ:RC type-NSRC	28.03	
POS.1-NNP:RC type-NSRC	-86.99	**
POS.1-NNPPOS:RC type-NSRC	-4.69	
POS.1-NNS:RC type-NSRC	67.16	**
POS.1-PRP:RC type-NSRC	29.21	
POS.1-PRP$:RC type-NSRC	121.07	*
POS.1-AUX:RC type-SRC	20.54	
POS.1-MD:RC type-SRC	14.34	
POS.1-RB:RC type-SRC	40.83	*
POS.1-VB:RC type-SRC	1.60	
POS.1-VBD:RC type-SRC	17.29	.
POS.1-VBN:RC type-SRC	-44.40	
POS.1-VBP:RC type-SRC	21.94	.

$.p < 0.10$, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Table 3: Regression coefficients and their significance levels for a minimal model of total reading time for the first word following the relative pronoun.

For first fixation times, only two of our independent variables were found to be significant predictors: word frequency ($p < 0.01$) and RC type ($p < 0.0001$, reading times for SRCs are again longer). Together with the inter-subject random effect, these two predictors explain 11% of the variance in first fixation reading times. The hypothesis that the RC type effect be due to differences in word length is not confirmed by the regression model, as length in not a significant predictor for first fixation times.

For the second word after the relative pronoun, we did not find any significant correlation with relative clause type. We found that 16.5% of the variance for total reading times is explained by a model that includes word length ($p < 0.0001$), word frequency, transitional probability (all $p < 0.01$) and the interaction between transitional probability and word length, and frequency and word length (both $p < 0.0001$).

[1] When removing the variable for the POS of the first word from the regression equation, model fit is a little lower. Highly significant factors in the model ($p < 0.001$) are RC type (longer reading times for subject RCs), transitional probability, frequency and length of the first word, as well as the interactions between RC type and transitional probability, RC type and frequency, frequency and word length, and landing position and length. Typical factors like frequency and transitional probability do not come up in the regression that involves POS tags, because their contribution to the variance is already explained by the word's POS (e.g., determiners are shorter and more frequent than adjectives).

	SRC	NSRC	Sign.
Transitional probability.1	-3.07	-2.90	.
Logarithmic frequency.1	10.60	11.79	***
Length.1	4.51	4.12	**

$.p < 0.10$, $**p < 0.01$, $***p < 0.001$

Table 4: Differences in transitional probability, frequency and word length and their significance levels for the first word after the relative pronoun with respect to RC type.

Skipping For skipping probabilities on the first and second words after the relative pronoun, we find frequency and length to be the significant predictors: shorter and more frequent words (which occur frequently in the NSRC condition, see Table 4) were skipped more often, and skipping was also highly dependent on whether the previous word had been skipped.

Regressions to the first word are more probable in NSRCs than in SRCs (although the difference does not reach significance level). We found regressions to mainly depend on the word's frequency, earlier skipping and the predictability of the following word. If the following word had a low predictability, regressions are more probable.

Discussion

We found increased reading times on the word directly following the relative pronoun for SRCs compared to NSRCs. This is consistent with the predictions of DLT, which assumes an increased integration cost for SRCs on this region. There was no spillover effect on the following region (the second word of the relative clause). We also tested for effects on the relative pronoun itself, and found that this region is read faster in SRCs than in NSRCs. Also, relative pronouns are skipped more often for SRCs. This is a new effect that is not readily predicted by DLT. However, a tentative explanation maybe that the word following the relative pronoun is on average longer for SRCs than for NSRCs. This might explain the greater tendency to skip the pronoun, perhaps because of parafoveal preview of the next word.

Transitional probability was a significant predictor in the region following the relative pronoun only when POS-tags were not entered into the regression. On the spillover region (the second word following the relative pronoun) and on the relative pronoun itself, we found effects of RC type and transitional probability. Overall, these findings indicated that transitional probability cannot serve as an alternative (but as an additional) explanation of the DLT effect we found.

In this context, is interesting to note that Hale's (2001) surprisal model makes opposite predictions for the word following the relative pronoun: in NSRCs, this region should be read more slowly, because the probability encountering a noun phrase following the relative pronoun is smaller than that of encountering a verb.

Conclusions

In this paper, we tested a theory of processing complexity, Gibson's (1998; 2000) Dependency Locality Theory (DLT), against reading time data extracted from a large eye-tracking corpus. We were able to show that DLT correctly predicts differences in processing complexity for subject and non-subject relative clauses. The complexity effect manifests itself in two distinct regions in the relative clause, leading to elevated reading times in these regions, as predicted by DLT. We also showed that transitional probability (McDonald & Shillcock, 2003b) has an effect on reading time in these regions, independent of the DLT effect.

To our knowledge, this is the first time a theory of sentence processing has been tested on data from an eye-tracking corpus. While we have only dealt with one construction (relative clauses) and one theory (DLT), we believe that our corpus-based approach constitutes an important new methodology for evaluating models of sentence processing, and we plan to evaluate other models (e.g. surprisal, Hale 2001). Such models are currently tested exclusively on data obtained for isolated, manually constructed sentences in controlled lab experiments. The validity of the models can be enhanced considerable if we are able to show that they scale up to model reading data from an eye-tracking corpus of naturally occurring text.

Acknowledgments

This research was supported by EPSRC grant EP/C546830/1 'Prediction in Human Parsing'. We are grateful to Roger Levy for numerous suggestions and comments regarding this work.

References

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, (pp. 132–139), Seattle, WA.

Frisson, S., Rayner, K., & Pickering, M. J. (2006). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 862–877.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, *68*, 1–76.

Gibson, E. (2000). Dependency locality theory: A distance-dased 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.

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

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.

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 149–157.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*, 313–330.

McDonald, S. A., & Shillcock, R. C. (2003a). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*, 648–652.

McDonald, S. A., & Shillcock, R. C. (2003b). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, *43*, 1735–1751.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes*, *41*, 221–250.