Gerold Schneider

Do Non-native Speakers Read Differently? Predicting Reading Times with Surprisal and Language Models of Native and Non-native Eye Tracking Data

Abstract: Theories of entrenchment and usage-based models have revolutionized cognitive linguistics and are also spearheading the paradigm shift in linguistics from theory-driven to empirical research. Entrenched, formulaic sequences are easier to process for native speakers, but more difficult to learn for L2 learners. We investigate the correlation between reading times as manifested in eye tracking corpora and text-derived measures of formulaicity, e.g., surprisal, word frequency, and a discourse-related pragmatic feature, predict reading times of L1 and L2 readers, and assess the differences. We use freely available corpora, such as GECO, which contains eye tracking based reading times by several native and non-native speakers.

We address the following RQs:

1) Which features correlate to and are predictive of reading times?
2) Are the features and their weights similar for L1 and L2 readers?
3) What is the role of individual variation?
4) Can L2 reading times be predicted as well as L1 times?
5) Does a comparison of L1 and L2 reading times reveal to us which constructions are particularly taxing for L2 readers?

We establish a ranking of features and show that surprisal is a less important feature for language learners, supporting recent findings that they can profit less form context due to less exposure and lack of routinization. Individual variation is strong and unsystematic, and learners can be predicted less well, partly also because slower readers (among them many L2) have lower model fit and can be seen as less efficient at the reading task. We finally zoom in on zones that are particularly taxing for learners, and observe that they find unusual word order, rare words and constructions and idioms harder to process. Our predicted reading times are quite accurate, the error is smaller than individual variation. This means that our models are suitable for cognitive and didactive purposes.

https://doi.org/10.1515/9783111017433-008

1 Introduction

This study quantitatively assesses reading times, aims to gauge the relative importance of the factors involved in a linear regression model, and it describes differences between reading times and factors when comparing native speakers (L1) and language learners (L2). Particularly, the influence of surprisal and of discourse-related factors are assessed, and where the ranking of features differs between L1 and L2 readers. The predictions of our model are quite accurate, on average the prediction error is much smaller than individual variation.

Quantitative methods and statistical models have revolutionized linguistics and especially cognitive linguistics (Glynn/Fischer 2010; Newman/Rice 2010; Janda 2013; Divjak/Levshina/Klavan 2016). Correlations between frequencies or frequency-derived measures and mental processing have been reported in numerous studies. For example, Ellis/Frey/Jalkanen (2009) show that there are strong correlations between collocation strength and word recognition. Wulff (2008) discusses that for the detection of collocations a frequency-based approach performs better than a similarity-based approach, indicating that frequency and semantics are intricately related. The Firthian hypothesis which says that a word is largely defined by the frequency of its context has given rise to models of distributional semantics (Sahlgren 2006).

Frequency and expectation can be used as a measure of what is easier to process and what is more expected. Rayner/Duffy (1986) have shown that the probability of words has an influence on the recognition of words if they are in isolation. The probability of words in their context is also related to recognition speed. Concerning frequent word sequences, which often grow into formulaic sequences, Conklin/Schmitt (2012) confirm:

> Virtually every study, using a variety of research methodologies, shows that formulaic language holds a processing advantage over nonformulaic language for native speakers. [. . .] The crucial role of frequency in processing clearly applies not only to individual words but also to formulaic sequences. It appears that frequency of exposure is a key aspect of learning formulaic sequences. (56)

If frequency and frequency-derived measures such as expectations (Shannon 1951) are predictive of human processing load, then models using these factors should be able to predict them. Reaction times have been widely accepted as measures of processing time (Grön 1996; MacWhinney 2001; Norman/Shah/Turkstra 2019). The reaction time in language reception in the form of reading texts is reading time (RT). RT is a psycholinguistic reaction time for integrating the read material. Smith/Levy (2013) show that the probability of a word in its context, so-called surprisal, closely correlates with reading time, and Schneider (in press)

predicts reading time using cognitive measures like surprisal. Frank (in press) summarizes the approach of predicting reading times with surprisal as follows:

> if linguistic prediction is probabilistic (i.e., statistical), it can be formalized and quantified using concepts from information theory. The most successful of these information-theoretic measures is surprisal – the negative logarithm of word's occurrence probability given the (linguistic) context. (3)

Simple, parsimonious, but reliable language models for native speakers can be built in this way.

Conklin/Schmitt (2012) also refer to language learning. Entrenched, formulaic sequences are more difficult to learn for L2 learners (Schneider/Gilquin 2016). Frequency of exposure plays a key role in language learning. But in L2 (second language) research, the question of how much reading times correlate with, or can be predicted by, language models have been less well investigated. While Pawley/Syder (1983) point out that language learners, due to their lack of exposure, have serious restrictions of building up nativelike routinization and thus intuition, Gries/Wulff (2005) show that language learners, too, are aware of the constructions in the language that they learn. Language learners partly base their knowledge of L2 on the constructions of their L1 (first language) and adapt them to make a choice on what to utter. The arising transfer can both be a help or a source of error and increased processing time, as the research tradition of second-language acquisition (SLA) has well documented (e.g., Saville-Troike/Barto 2016).

Frequency of exposure also plays a major role in grammaticalization and language change. "Frequency is not just a result of grammaticisation, it is also a primary contributor to the process" (Bybee 2007: 337). This insight is on the one hand the cornerstone of construction grammar (Goldberg 2006; Hilpert 2019). On the other hand, the lower frequency with which language learners have been exposed to constructions and sequences also leads to the expectation that rare constructions and idiomatic sequences may be harder to process for learners, both in language production and reception.

The choices which speakers, listeners and readers have to make, involve complex mental processes (Larsen-Freeman 1997; Larsen-Freeman/Cameron 2008). Well-studied instances of speaker decisions are alternations such as the dative shift (Bresnan et al. 2007; Bresnan/Nikitina 2009), for which logistic regression models can predict the outcome with high accuracy. But alternations are only one of the many choices that people have to make when they use language, and they mainly relate to language production. Decisions are required at every word, both to utter and to integrate it during reading, or at least every word sequence, due to routinization. In unexpected contexts, decisions are harder and

take more time. In the context of SLA, the few well-studied areas include e.g., Verb-Argument Constructions (Gries/Wulff 2005; Ellis 2013). A model of the interacting complex phenomena and the discourse is still largely absent, however. Ellis (2013) summarizes this lack of research on the topic as follows:

> Research to date has tended to look at each hypothesis by hypothesis, variable by variable, one at a time. But they interact. And what is really needed is a model of usage and its effects upon acquisition. We can measure these factors individually. But such counts are vague indicators of how the demands of human interaction affect the content and ongoing coadaptation of discourse. (8)

For language learners, more contexts and more words are unexpected, and as they are less skilled in routinization (Pawley/Syder 1983) – a general increase in reading time can be expected. Segalowitz/Segalowitz (1993) report longer reaction times and more variability in L2 than in L1 speakers, using a lexical decision task. Despite this early experiment, there is still relatively little research on L2: "there are as yet very few applications of reaction time methodologies in applied linguistics" (Racine 2014: 4).

In this study, we aim to contribute to these lacunae by using context-aware language models. In particular we use surprisal, and other context-based measures. An important pragmatic factor is recency in discourse: has an entity been introduced before, when was it mentioned last? Recent mentions are more present and more quickly accessible in speakers' and listeners' minds. An important syntactic measure is punctuation – explicit markers of clause and sentence boundaries also mark boundaries of processing units. A trivial but important factor to consider is word length – longer words take more time to read. We use the factors to predict reading times in psycholinguistic experiments obtained by measuring eye tracking (e.g., Conklin/Pellicer–Sánchez/Carrol 2018), and we compare L1 and L2 readers. The correlation between surprisal (Levy/Jaeger 2007) and reading times is generally accepted, but it is unclear how much it correlates with other factors, and what differences between native speakers and language learners are, and also the role of individual differences has not been studied sufficiently.

Specifically, we address the following research questions:

1) Which features correlate to and are predictive of reading times?
2) Are the features and their weights similar for L1 and L2 readers?
3) What is the role of individual variation? This question needs to be addressed because possible differences between L1 and L2 could be overshadowed by individual differences.
4) Can L2 reading times be predicted as well as L1 times?
5) Does a comparison of L1 and L2 reading times reveal to us which constructions are particularly taxing for L2 readers?

Our paper is structured as follows. We present a brief overview of previous research in comparison to our study in section 2, and data and methods in section 3. We present quantitative results in section 4, where we first assess correlations of reading times to our investigated features, and then use linear regression to predict reading times. In section 5, we present a qualitative study. Particularly, we discuss which linguistic phenomena are taxing for L2 readers, i.e., phenomena for which require considerably longer processing times.

2 Related Approaches

Eye Tracking data can be used as models of mental load and processing time to researchers. In this section, we give a brief review of related approaches.

2.1 Correlations between Reading Times, Surprisal and Other Factors

The correlation between surprisal (Levy/Jaeger 2007) and reading times has been confirmed by several studies in eye tracking experiments (Frank et al. 2013). Eye movement experiments have shown that surprisal correlates to reading times (Demberg/Keller 2008), but it is unclear how much it correlates with other factors. We first give a brief impression of the data compiled by Frank et al. (2013). This corpus contains individual sentences in isolation, a controlled setting in which discourse factors and semantics should not play a major role, so that only the local context influences processing. Surprisal can thus be expected to be particularly important. We measure the size of the correlation of surprisal, and compare it to other factors. Correlation strength is intuitive to interpret.

For the first 7724 words of the Frank Corpus (Frank et al. 2013), which includes 1931 words by four readers, Schneider (in press) observes a Pearson correlation of 0.25 between bigram surprisal and reaction time (RT) expressed in the variable RT Go-Past, which gives the total gaze time in milliseconds for each token, i.e., the milliseconds spent until finally leaving to further right). A correlation of about 0.25 may seem low; but when considering which other factors correlate, most correlate less strongly. Tab. 1, adapted from Schneider (in press) lists a selection of further variables.

The only factor with a similarly high correlation that Schneider (in press) found in the Frank corpus is word length in characters – longer words take longer to read.

Tab. 1: A selection of correlating factors of four reader in Frank et al. (2013).

RTrightbound correlated to:	Pearson Correlation	My Comments
Length of word in letters	0.256	Highly correlating factor
Bigram Surprisal	0.256	Equally high
Observed / Expected Collocation	0.012	Very low
Position of word in sentence	0.129	Longer sentences take longer to read
Sentence number	−0.071	No slowdown during reading progress

The influence of word frequency on RT has been investigated in many studies (e.g., Rayner/Duffy 1986). While psycholinguistic studies more typically obtain predictability by presenting sentence fragments to subjects (cloze tasks), we use surprisal calculated from large corpora, like Demberg/Keller (2008) and Shain (2019), in order to address the criticism by Ellis (2013) that interactions between variables and decisions of speakers or readers need to be considered. For the sake of parsimony, we use a simple surprisal model, and only those features which are most significant, as reported in previous research. According to Schneider (in press) the most significant features for L1 readers are: word length, presence of punctuation, distance to last previous occurrence of the same word, and surprisal. We use these four features for predicting RT of L2 readers in the current study.

2.2 Reading Times of Language Learners

Racine (2014) states that reaction-time research in applied linguistics is generally still rare. Also, in the area of L2 eye tracking, there are only few studies comparing reading times of L1 and L2 speakers, in particular Underwood/Schmitt/Galpin (2004), Siyanova–Chanturia/Conklin/Schmitt (2011), and Schilk (2017).

Underwood/Schmitt/Galpin (2004) focus on the processing of formulaic sequences. They report mean fixation times of 201 ms (at a standard deviation of 26 ms) for L1 readers, and 228 ms (at a standard deviation of 29 ms) for relatively advanced L2 readers. They conclude that the final word of formulaic sequences are fixated significantly less long by native speakers, indicating their routinization advantage and suggesting that they are more likely to store entire formulaic sequences as single units in the mental lexicon.

Siyanova–Chanturia/Conklin/Schmitt (2011) measure differences in the processing of idioms with figurative meanings. They conclude that idioms are read significantly faster by L1 readers, irrespective of whether they have compositional (literal) or non-compositional (figurative) meaning.

Schilk (2017) compares reading times of selected verb-object and adjective-noun collocations based on frequent learner errors (Nesselhauf 2005). He compares less advanced L2 speakers to more advanced L2 speakers and concludes that the less advanced speakers show significantly longer fixation times for both verb-object and adjective-noun combinations than the more advanced speakers.

While these three studies provide valuable insights and confirm the hypothesis that L2 readers process the selected phenomena more slowly, they cannot offer the broad overview considering all phenomena in their interrelated nature as Ellis (2013) proposes. In order to model this interrelated nature, natural language models can be used. We predict reading times with linear regression based on a variety of features including surprisal. Frank (in press) suggests to use recursive neural networks trained on reading times form L1, L2 or both types of readers. Our approach uses regression modelling instead, which is typically slightly less accurate but more parsimonious, as regression models easily allow us to asses factor weights, measure model fit and explain areas of prediction inaccuracy. Frank's proposal bears enormous promise, but currently "research on bilingual comprehension by neural networks is clearly still in its infancy" (in press: 14).

3 Related Approaches

In this section, we introduce our data and methods.

3.1 Data

There are several corpora that are annotated for reading time using eye tracking data. For our study, we have considered the following four sources:

1. **Reading times for model evaluation** (Frank et al. 2013). It contains 205 simple domain-independent sentences read by 43 participants. The motivation for the compilation was that "understanding newspaper or narrative texts requires vast amounts of extra-linguistic knowledge to which the models have no access . . . a more appropriate data set for model evaluation would consist of independent sentences that can be understood out of context" (Frank et al. 2013: 1185). The corpus also contains ten L2 readers.
2. **Ghent Eye tracking Corpus** (GECO; Cop et al. 2017). This is knowledge-dependent corpus, which entails that extra-linguistic knowledge influences reading time, but also offers the change to include discourse features. An

entire Agatha Christie novel is read out by a dozen of L1 and L2 speakers. The motivation for the collection of the corpus was: "this corpus has the potential to evaluate the generalizability of monolingual and bilingual language theories and models to the reading of long texts and narratives" (Cop et al. 2017: 602).

3. **Dundee Corpus** (Kennedy/Hill/Pynte 2003; Kennedy et al. 2013): 10 native English and 10 native French speakers read a text of 56,000 words. The corpus is not freely available.
4. **Provo Corpus** (Luke/Christianson 2018): 55 paragraphs, containing 2,800 words are read out by 84 native speakers of English. The corpus is available for free, but it has no L2 readers and was thus not suitable for our study.

Based on this comparison, we decided to use the Ghent Eye tracking Corpus (GECO; Cop et al. 2017) as our main corpus. Additionally, we also measure reading times in Reading Times for Model Evaluation (Frank et al. 2013). We restricted our investigation to the 12 L1 readers whose data is complete (some others have e.g., not read the entire novel), and to the 7 L2 readers who had less than 50% daily exposure to English. While discarding some L2 readers increases data sparseness, concentrating on the least exposed readers allows us to concentrate on prototypical L2 readers. According to Cop et al. (2017) readers with more than 50% daily exposure (Bilinguals L2) show no significant differences compared to native English speakers in terms of the performance in the tests which all participants had to take (Cop et al. 2017: 607, Tab. 1, last column) while the differences to speaker with less than 50 daily exposures are highly significant (Cop et al. 2017: 607, Tab. 1, second last column). These tests included the LexTALE test, spelling score, and lexical decision accuracy. There was no significant difference in text understanding between L2 and native English speakers, which indicates that both groups read the novel similarly carefully.

All L2 speakers in this study are native speakers of Dutch, which has the advantage that they are comparable in terms of L1 influence, but adds the serious limitation that just one L1 background is reflected in the data. It will not be possible to discern which areas of slowdown point to general learner-specific processing, and which are typically difficult for L1 Dutch speakers, due to inference or due the language-specific differences. Typologically, Dutch is a Germanic language like English, and the enormous influence of French on English should also not add major difficulties to Dutch speakers, who typically have a working knowledge of French.

3.2 Methods

We correlate reading times (RT) and to relevant factors, and then predict RT with these factors. We always use the total reading times, i.e., the total gaze duration, sometimes involving more than one gaze if the reader backtracks. As correlation measure, we use Pearson correlation. For the prediction of reading times, we use linear regression. In what follows, we list the predictors used in the model. These are surprisal, distance, word length and punctuation.

Surprisal (Levy/Jaeger 2007), our first feature, is generally defined as the probability of a word in its context, or p(word|context) in Bayesian terms. It is usually expressed as a logarithm to give an information-theoretic value, the surprise in bits for seeing a new word in the given context.[1] The detailed definitions can vary, we are using a simple operationalization: the probability of a word linearly combined with the probability of transition from the previous word: "the forward transitional probability $P(w_k|w_{k-1})$ is a simple form of surprisal" (Demberg/Keller 2008).

Our definition is thus:

$$bigram\ surprisal = log\frac{1}{p(w_k)} + log\frac{1}{p(w_k|w_{k-1})}$$

We have learned the probabilities from the British National Corpus (Aston/Burnard 1998). The probability of a word is simply its frequency divided by the corpus size. An example of a sentence with bigram surprisal is given in Fig. 1. We can see areas of low surprisal, for instance the pronoun *I*, which is generally frequent, and even more so at the beginning of a sentence, which explains why surprisal for *I* is lower in its first occurrence than after *when* later in the sentence. A further example is the word *to*, which is frequent and also in a common context in both occurrences here. The context *trying to make* is slightly less common than *what to do*, which leads to a slightly higher surprisal, an expectation that is also mirrored by a slightly higher reading time of *to* in *trying to make*. Surprisal is highest for the name *John Cavendish* – even the frequent name *John* is so infrequent that it cannot be predicted from the previous words, unless we have discourse-specific knowledge.

Surprisal allows us to measure chunking (Altenberg 1998) and the competition between the idiom and syntax principle (Sinclair 1991). Linguistic contexts dominated by the idiom principle have low surprisal, many chunks, are easy to

1 We cannot provide an introduction to Information Theory here, but let us look at a simple example: in order to express 8 equally likely words, 3 bits are needed, as 2^3 equals 8.

process, but contain little information. Linguistic contexts which make maximal use of syntactic creativity can compress a lot of information into few words, but this makes it very hard for readers or listeners to follow: surprisal is very high, the continuation of the utterance is hard to predict. Shannon's (1951) noisy channel easily breaks down when redundancy is too low. In spoken language this can lead to misunderstanding and uncertainty, while in written language it typically leads to longer reading times and backtracking. According to Levy/Jaeger (2007), successful communication needs to strike a balance between the two: surprisal should stay approximately constant. This is the principle of uniform information density (UID). "UID can be seen as minimizing comprehension difficulty" (Levy/Jaeger 2007: 850).

UID holds quite well in spoken language, while some compressed written genres (Biber/Conrad 2009), particularly the scientific genre, exhibit frequent areas of high surprisal (Schneider/Grigonyte 2018).

ID	Word	Reading Time (ms)	Surprisal
115	I	512	11.263337
116	was	279	16.195808
117	trying	225	12.843803
118	to	231	11.66212
119	make	184	15.584163
120	up	470	17.04785
121	my	0	15.787277
122	mind	277	19.215173
123	what	214	15.162532
124	to	130	11.381907
125	do	0	17.664406
126	when	468	13.040814
127	I	0	17.910504
128	ran	0	19.851737
129	across	3723	24.424627
130	John	555	26
131	Cavendish	717	26

Fig. 1: An example sentence from GECO with bigram surprisal.

Word probability is correlated to word recognition speed (Rayner/Duffy 1986), both in isolation and in context. Predictability of the context in psycholinguistics is often obtained by presenting sentence fragments to subjects. We use surprisal instead, which allows us to address all phenomena in their complexity and interaction (Ellis 2013). Eye movement experiments have shown that surprisal correlates to reading times (Demberg/Keller 2008). Correlations to EEG

activity has also been investigated (Frank et al. 2013). Shain (2019) shows that word frequency and predictability from the context are hard to separate and that predictability by means of surprisal is a better predictor. We thus do not include frequency separately.

Smith/Levy (2013) show that the effect of word predictability on reading time is logarithmic, across 6 orders of magnitude, from very rare to very frequent words. The logarithmic correlations entail that the modelling of surprisal as an information-theoretic value using a logarithmic scale is cognitively adequate.

Surprisal gives us a language model, albeit a simplified one. It is surface-oriented, in the sense that it uses extremely small amounts of context. In a real-world discourse, previously seen words are expected and are thus read fast. We address this shortcoming by measuring the distance to the most recent occurrence of the current word. This feature, **distance**, is our second feature. RT depends on word frequency, but this effect largely disappears after three repetitions (Rayner/Raney/Pollatsek 1995) in the discourse. Church (2000) observes that the probability for seeing a content word twice in a text is closer to (p(word) /2) than p(word) * p(word) which would be expected under the independence assumptions. Particularly the GECO material, where an entire novel is read, needs an integration of discourse-related features. The correlation in GECO between the logarithm of the distance and RT is 0.46 for L1 and 0.38 for L2 readers, values that are so high that we decided to include them. For previously unseen words, a default value of distance=10000 is given.

The third feature to be included is **word length**. This feature is trivially related to RT – longer words take longer to read. We measured a correlation between the logarithm of word length and RT is 0.64 for L1, and 0.55 for L2 readers. We expect this feature to dominate the feature weights.

As punctuation symbols are too small to measure fixations, and as they are often never fixed on, the slowdown caused by punctuation is not directly accessible in GECO. The full wordform in the data simply includes punctuation symbols. We have removed them and instead introduce a binary feature **punctuation**, which we set whenever a word is followed by a punctuation symbol. We do not distinguish between commas, full stops or other punctuation symbols.

3.3 Linear Regression and Step-wise Regression

Linear regression techniques and mixed models are frequently used in linguistics (see e.g., Winter 2013; Gries 2015; Speelman/Heylen/Geeraerts 2018; or Schneider/Lauber 2019). We are predicting the observed reading times (RT) for each word in the corpus.

We use multivariate models as many factors are involved. They comprise surprisal, distance to the last previous occurrence of the word, length of the word, presence of punctuation. They were the most significant features for predicting the reading time of L1 speakers in GECO in Schneider (in press). These factors in combination partly explain the observed reading times. A frequently used measure of the percentage of the data that is explained by the model is the R^2 metric. Molnar (2020) summarizes its function as follows:

> R-squared tells you how much of your variance can be explained by the linear model. R-squared ranges between 0 for models where the model does not explain the data at all and 1 for models that explain all of the variance in your data.

We also report the adjusted R^2, a version of R^2 which takes the number of factors used into consideration. This is important as a higher number of factors increases the likelihood of overfitting, and reduces the parsimony of the model. This adjustment is an operationalization of Occam's razor, a principle which states that if several theories explain a fact equally well the simpler explanation should be given preference.

The complex multifactorial nature of language in general, and reading time in particular, involves correlations between the many features. We employ model selection with stepwise regression in the form of step-down methods for ranking the weights of features. The step-down method for feature ranking, leave-one-out, (function `drop1`), is part of the R base package. Rodríguez (2020) describes stepwise regression as follows:

> The basic idea of the procedure is to start from a given model . . . and take a series of steps, by either deleting a term already in the model, or adding a term from a list of candidates for inclusion.

Stepwise regression leads to more reliable results than using the model with all significant features because interactions between the features are taken into account. An assessment of feature weights based on full models also has the problem that the standard regression function `aov()` in R uses the F-measure in such a way that it depends on the order of the tested features in the entered formula. A further standard function for linear regression models (function `lm` in R) report each factor level separately, which makes it difficult to assess the overall importance of a feature.

4 Quantitative Results

In this section, we present our quantitative results. In the next section, we then take a qualitative perspective.

First (section 4.1), we focus on individual differences. The differences are so big that we suggest to pool participants. In particular, we will use mean reading times. Schneider (in press) discusses the motivation for using means or also modes in more detail. For our current purpose, we intend to model typical L1 and L2 readers, abstracting away from individual differences. Then (section 4.2), we predict reading times with linear regression models.

4.1 Differences between Individuals and L1 vs L2

The individual differences between the participants are very pronounced. The first four readers in the Frank corpus (all L1) have a mean reading time of between 150 ms and 274 ms per word. In GECO, the differences are similar: the fastest L1 reader has a mean of 124 ms, the slowest 253 ms. The L1 mean in GECO is 199 ms, and the standard deviation is 43 ms. L2 readers are generally slower – their mean is 266 ms, at 34 ms standard deviation. The densities of reading speed are plotted in Fig. 2, with the means as dashed lines.

Schneider (in press) reports a Pearson correlation of 0.25 for RT and surprisal for the Reading Times for Model Evaluation corpus (Frank et al. 2013). We found considerably lower correlation in GECO: for L1, the correlation has a mean of 0.159, and for L2, the mean is 0.128. The difference between the Frank corpus and GECO partly stems from the fact that L2 readers have lower correlations, and partly from the fact that GECO is a coherent discourse so that other factors play an important role. It can also be observed that fast readers generally exhibit a stronger correlation, which may indicate that they manage better to concentrate on the important subtasks, such as predicting likely continuations. The same explanation can also be adduced for language learners, which have had much less exposure to language material (Pawley/Syder 1983). The density curve for the correlations of individuals to surprisal, split by L1 and L2, is given in Fig. 3, with the means of all readers added as a dashed line.

The correlation between RT and the correlation (between RT and surprisal) is –0.449 for L1, and –0.401 for L2 readers. The fact that there is such a meta-correlation means that fast readers correlate more strongly to surprisal. Efficient readers match the surprisal model considerably better, they probably manage to profit better from the context.

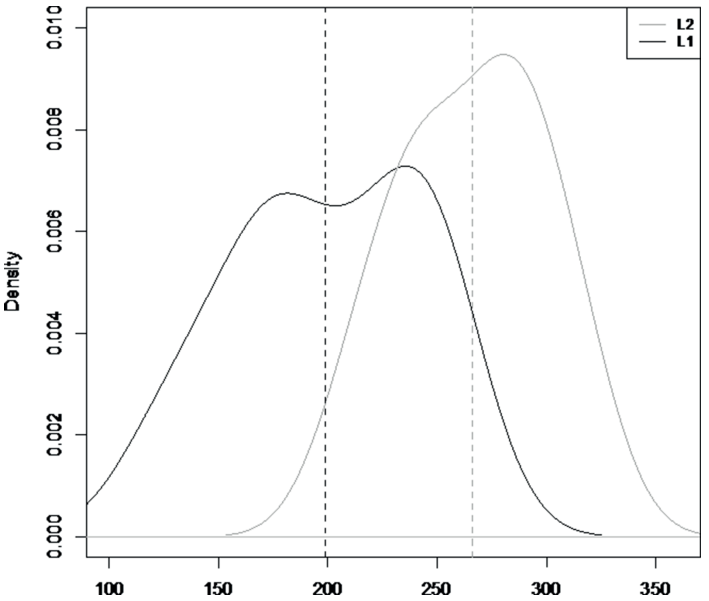

Fig. 2: Density curves of per-word reading time means per participants, split by L1 and L2.

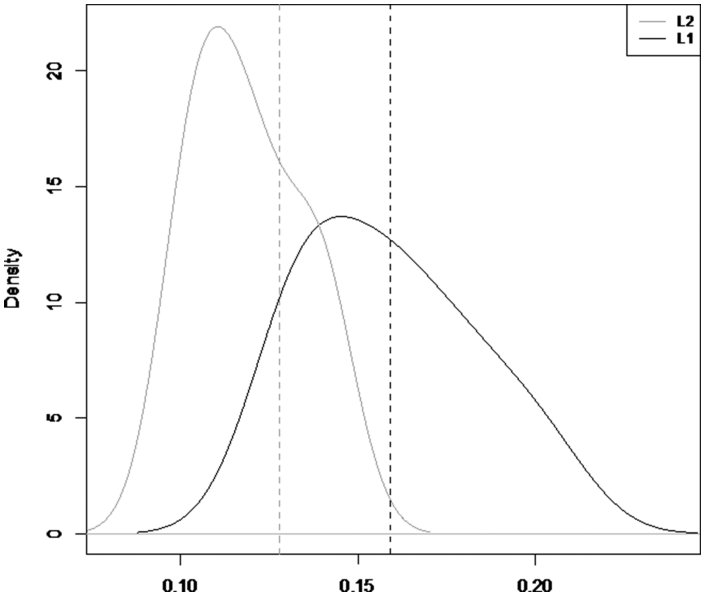

Fig. 3: Correlations between reading time and surprisal, by L1 and L2 individuals.

The very strong individual variation means that individual reading times are also not very strongly correlated: the reading time of a different reader is as good a predictor as surprisal – the mean of the RT correlations between the L1 readers in GECO is 0.150. As individual variation is so strong, we use the mean of the reading times of L1 and L2 as a smoother variable, henceforth RT means. The correlation of the RT means to surprisal is considerably higher: 0.35 for L1, and 0.25 for L2. RT means plotted against reading times, with trend lines for L1 and L2, are given in Fig. 4. The fact that the trend line is less steep for L2 also shows that L2 readers have lower correlation to surprisal, suggesting that they manage less to profit from the context.

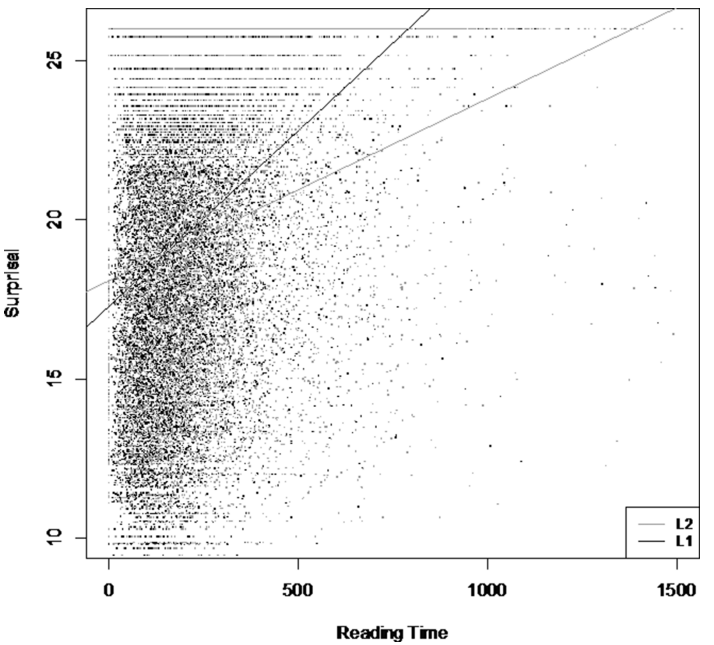

Fig. 4: Plot of surprisal against reading time, with trend lines for L1 and L2.

4.2 Predicting Reading Times with Regression Models

In order to assess the weights of the various factors, we use linear regression models to predict the reading times of L1 and L2 readers.

4.2.1 Individual Variation

In a first pilot model (trained on the Frank corpus), the individual readers (four L1 readers) are kept as a factor in order to assess the weight of the factor individuality. The factor weights are given in Fig. 5.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
SURPRISAL	**1**	**13818343**	**13818343**	**602.734**	**< 2e-16**	*******
LENGTH	**1**	**15508232**	**15508232**	**676.445**	**< 2e-16**	*******
tags	**38**	**4479610**	**117884**	**5.142**	**< 2e-16**	*******
subj_nr.f	**3**	**9748779**	**3249593**	**141.742**	**< 2e-16**	*******
word_pos	1	48112	48112	2.099	0.14748	
sent_nr	1	1284480	1284480	56.027	7.94e-14	***
prob	1	180147	180147	7.858	0.00507	**
SURPRISAL:LENGTH	1	30	30	0.001	0.97126	
Residuals	7676	175980675	22926			

Fig. 5: R output for L1 factor weights in Frank's corpus.

As can be seen in the Fig. when considering the F-score values in the second last column, word length (LENGTH) is the most important factor, surprisal (SURPRISAL) emerges as almost equally important, followed by the individual reader (subj_nr.f) and POS tag (tags). Further significant factors are the tagger confidence (prob), and the sentence number. We did not include these factors in our current study, though. The position of the word in the sentence (word_pos) is not a significant factor. The fact that tagger confidence is significant is an interesting psycholinguistic observation, which we will not pursue further as it is not an argument of the current paper, but words that are ambiguous for the tagger have longer RT, i.e., they need more processing effort.

The strong individual variation observed thus far could prompt one to use a mixed model approach in which the subject is a random effect. We used a mixed model with the lme4 package of R, which reported only a very small systematic effect by the reader: standard deviation of the random effect of the individual was 42.2, more than 7 times smaller than the residual (311.7). We thus decided to predict the much smoother RT means rather than individual reading times.

4.2.2 Prediction of L1 RT

We now present a model predicting L1 RT means, and then a different one for L2 RT means in Section 4.3.3 below. Both models are trained on GECO. The L1 model is given in Fig. 6.

```
> fitL1 = lm(ppMean1 ~ LENGTH + SURPRISAL + log(distance) +
PUNCTUATION, data=eyegecomBOTH)
> summary(fitL1)

Call:
lm(formula = ppMean1 ~ LENGTH + SURPRISAL + log(distance) +
PUNCTUATION,
    data = eyegecomBOTH)

Residuals:
    Min      1Q  Median      3Q     Max
-245.40  -54.82  -12.90   37.82 1335.31

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.2164     3.5401   1.474    0.141
LENGTH          27.5095     0.4238  64.907  < 2e-16 ***
SURPRISAL        1.4293     0.1915   7.465 8.95e-14 ***
log(distance)    4.0678     0.4128   9.854  < 2e-16 ***
PUNCTUATIONyes  40.0453     2.3420  17.099  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 87.45 on 11513 degrees of freedom
Multiple R-squared:  0.4753,     Adjusted R-squared:  0.4751
F-statistic:  2608 on 4 and 11513 DF,  p-value: < 2.2e-16
```

Fig. 6: R output for factor weights for L1 on GECO.

The quality of the model can be assessed in several ways. Fig. 6 shows that R^2 of the model is 0.475 (and adjusted R^2 is very similar as we have a simple model), which means that the model explains 47.5% of the variation in the data. The predictions of the model are 61.6 ms off on average (second column), which is 34% of the RT mean of 199 ms (last column). The average error is modelled on the calculation of the standard error: the squared difference between observed RT mean (O) and the model prediction (E) is calculated, which indicates the variance, and the square root of this expression delivers the standard deviation. The Z-score mean, i.e., standard deviation of our prediction divided by standard deviation of individual reading times, is 0.54, which means that our predictions are typically off by 54% of the standard deviation. The Z-scores mean is below 1, which indicates that individual variation is much larger, we can conclude that this model makes a reasonably accurate prediction of reading time. The Z-score of the best model is 0.51, which means that our predictions is off by 51% of the individual variation. In other words, our predicted RT is well within the expected individual variation, which means that the predictions of the model can be used fairly reliably for applications that aim to predict RT of a typical reader, the reader that our model predicts would be a totally unobtrusive test person.

The performance of this model is compared to simpler models, using stepwise regression and feature ablation, in Tab. 2. The dominant factor of word length (line 1) already makes linear predictions that are only off by about 35%. The best model (last line) is one percent better. Surprisal on its own is 45% off, word length and surprisal in combination is off by 34.6%. The increase in accuracy generally mirrors the ranking of factor weights.

Tab. 2: Prediction accuracy of linear regression models on L1.

QUALITY OFPREDICTION	√(O-E)^2=typical error in ms	mean(Z-score)= typical error/sd	relative offness=typical error/mean
Length (L)	63.30	0.5244	34.92%
Surprisal (S)	82.14	0.6805	45.32%
L+S	62.69	0.5194	34.59%
L+S+punctuation	62.13	0.5148	34.28%
L+S+punctuation+distance	61.65	0.5108	34.02%

4.2.3 Predictions of L2 RT

We now turn to the prediction of L2 RT. We use the same factors as in the last line of Tab. 2 (i.e., length + surprisal + punctuation + distance) to predict the reading times of the L2 readers in GECO. The L2 model is given in Fig. 7.

We can see that the model fit is much lower. R^2 is only 0.332. The lower model fit also explains why the T values are generally lower. The order of the factor weights is similar, but surprisal is slightly less significant.

The lower model fit also means that L2 readers are less systematic. Also, the accuracy of predicting L2 RT means is lower than the one for L1 RT means. Tab. 3 compares the quality of predictions. The typical error increase from 62 to 91 ms, the Z-score increases from 0.51 to 0.56. This is still below 1, which means that we also predict an unobtrusive L2 reader, but it is harder to predict RT of language learners.

In addition, word recognition seems to be more difficult for L2 readers than for L1 readers, possibly because more words are unknown or unfamiliar to L2 readers. The mechanical and trivial correlation between word length and reading time is 0.667 for L1, but 0.570 for L2 readers. As longer words are typically rarer and harder to learn (Graën/Alfter/Schneider 2020) a lower correlation between word length and RT is not necessarily expected.

```
> fitL2 = lm(ppMean2 ~ LENGTH + SURPRISAL + log(distance) +
PUNCTUATION, data=eyegecomBOTH)
> summary(fitL2)

Call:
lm(formula = ppMean2 ~ LENGTH + SURPRISAL + log(distance) +
PUNCTUATION,
    data = eyegecomBOTH)

Residuals:
    Min      1Q  Median      3Q     Max
-373.46  -80.55  -20.75   54.13 2029.28

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     46.1320     5.3545   8.616  < 2e-16 ***
LENGTH          33.1838     0.6411  51.764  < 2e-16 ***
SURPRISAL        0.9737     0.2896   3.362 0.000776 ***
log(distance)    4.1487     0.6244   6.645 3.18e-11 ***
PUNCTUATIONyes  23.1290     3.5424   6.529 6.89e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 132.3 on 11513 degrees of freedom
Multiple R-squared:  0.3324,     Adjusted R-squared:  0.3322
F-statistic:  1433 on 4 and 11513 DF,  p-value: < 2.2e-16
```

Fig. 7: Factor weights for L2 on GECO.

Tab. 3: Prediction accuracy of linear regression models on L2 compared to L1.

QUALITY OF PREDICTION	√(O-E)^2=typical error in ms	mean(Z-score)= typical error/sd	relative offness=typical error/mean
L1 : L+S+punctuation+distance	62.13	0.5148	34.02%
L2 : L+S+punctuation+distance	91.21	0.5635	38.84%

4.2.4 Model Analysis and Feature Order

In order to interpret the L1 and L2 models psycholinguistically, we assessed their feature weights. A model with so many features, particularly when dealing with a highly redundant system like Natural Language (MacWhinney/Bates 1989; Shannon 1951) leads to a range of strong interactions. While the feature significance p(|t|), and the t-value delivered by `lm()` in R, and also the F-measure from `aov()` provide useful hints for model selection and interpretation, they partly depend on the order in which the features appear in the equation. The leave-one-out method `drop1` is a step-wise regression approach and gives a more reliable

impression of relative feature weights. The R output for this is given in Fig. 8. The model for L1 is given at the top of the Fig., the one for L2 at the bottom.

```
> drop1(fitL1, test = "F")
Single term deletions

Model:
ppMean1 ~ LENGTH + SURPRISAL + log(distance) + PUNCTUATION
              Df Sum of Sq       RSS    AIC  F value    Pr(>F)
<none>                     88041895 103000
LENGTH         1  32217059 120258954 106590 4212.938 < 2.2e-16 ***
SURPRISAL      1    426115  88468011 103054   55.722  8.95e-14 ***
log(distance)  1    742605  88784500 103095   97.108 < 2.2e-16 ***
PUNCTUATION    1   2235771  90277666 103287  292.366 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> drop1(fitL2, test = "F")
Single term deletions

Model:
ppMean2 ~ LENGTH + SURPRISAL + log(distance) + PUNCTUATION
              Df Sum of Sq       RSS    AIC  F value    Pr(>F)
<none>                     201419529 112532
LENGTH         1  46878434 248297964 114940 2679.539 < 2.2e-16 ***
SURPRISAL      1    197751 201617281 112541   11.303 0.0007762 ***
log(distance)  1    772440 202191969 112574   44.152 3.175e-11 ***
PUNCTUATION    1    745828 202165357 112573   42.631 6.888e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 8: Step-wise regression with leave-one-out on L1 and L2 model.

The order of features suggested by our regression experiments (Fig. 7) is confirmed by looking at the *F*-measures:

Word length > punctuation ≥ distance to previous occurrence of same word > surprisal

In the comparison between L1 and L2 it can be observed that all F values of L2 are lower, as the model fit is much lower. L2 readers show much more variability and their reading times are harder to predict. In the comparison of the *F*-values we can observe that surprisal is indeed less important for L2 readers, again confirming the lack of routinization and expectation of the continuation of the text. L2 readers also seem to make less efficient use of punctuation symbols, which give clues to the syntactic structure. The discourse feature of the distance to the last previous occurrence of the same word, and the trivial feature of word length are also less important for L2 readers, but they keep more of their predictive power in comparison to other features.

4.2.5 Prediction of L2 RT by L1 RT

Finally, we consider a model in which we add L1 routinization experience and vocabulary knowledge to predict L2 reading time. We do so by adding L1 RT means as an independent variable. This model assesses how useful it is to know L1 RT to predict L2 RT, in comparison to other factors. If L1 and L2 readers had nearly identical reading behaviour, we would expect that L1 RT overshadows all other factors. The feature weights of the corresponding linear model are given in Fig. 9.

```
> fitb2 = lm(ppMean2 ~ LENGTH + log(distance) + ppMean1 + SURPRISAL
+ PUNCTUATION, data=eyegecomBOTH)

> drop1(fitb2, test="F")
Single term deletions

Model:
ppMean2 ~ LENGTH + log(distance) + ppMean1 + SURPRISAL + PUNCTUATION
              Df Sum of Sq       RSS    AIC   F value    Pr(>F)
<none>                     177940797 111107
LENGTH         1  11224825 189165622 111809  726.1976 < 2.2e-16 ***
log(distance)  1    186672 178127469 111117   12.0769 0.0005124 ***
ppMean1        1  23478732 201419529 112532 1518.9724 < 2.2e-16 ***
SURPRISAL      1     11521 177952318 111105    0.7453 0.3879729
PUNCTUATION    1      8157 177948954 111105    0.5277 0.4675837
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 9: Feature weights of a model predicting L2 reading times with L1 reading time as predicting variable.

RT means of the L1 readers (ppMean1) is the strongest predictor as its F-value is highest, but word length has almost equally strong weight. We can conclude that L2 readers read texts differently, but L1 reading times are still marginally the best predictor.

5 Qualitative Results

After we have seen that there are strong differences not only in the reading speed but also in the way L1 and L2 readers read a text, we will investigate which linguistic phenomena are treated differently by L1 and L2 readers, with the aim of finding out what L2 readers find particularly difficult. For this investigation, we have visualised the differences between the reading times using a heat map in MS Excel, and zoom in on areas of particularly strong differences. Strong differences are automatically marked by yellow to dark red highlighting. Words and zones with large differences between L1 and L2 readers stand out in strong colours.

In order to relate L2 RT means to L1 RT means we employ the overuse measure O/E or *Observed divided by Expected*. The expected value is the mean of L1 and L2. O/E(L2) is then:

$$O/E(L2) = \frac{O(L2)*2}{O(L1) + O(L2)}$$

In order to group the O/E value around 0 we display O/E-1 in the second column of the following visualisations. A value of 0 expresses equal reading time for L1 and L12, +1 means that L2 readers take longer than L1 (which happens if all L1 readers have no fixation and thus RT of 0 on a word), a negative number means that L2 readers are faster than L1. As we wanted to spot zones of reading difficulty for L2 readers in addition to individual words, we also calculate the mean over 5 words. This value is given in column 3 and can serve as an indication of relative reading difficulty. In the last column we list the total RT of the last 5 words, i.e., the absolute reading difficulty. All values represent the means across the readers of the L1 and L2 class, respectively. As L1 readers use about 200 ms per word (see Fig. 2 in Section 4.1 above), values above 1000 ms for 5 words are also indicative of an area where L2 readers experience a slow-down.

By reading the entire heat-map-enriched corpus vertically, we were able to identify several linguistic phenomena that L2 readers spent a lot of time on. These are:

- Fronting, i.e., non-canonical word order
- Zero-relative pronouns
- Rare vocabulary items
- Nominalisations
- Long attachments
- Rare constructions
- Unusual word meanings
- Complex preposition and phrasal verb constructions
- Idioms
- Irregular and strong verbs

In the following, we present screenshots of the heat-maps and identify the zones in which the L2 readers slowed down. Examples of non-canonical word order due to fronting are given in Figs. 10 and 11. The fronted object *what* in the sentence *That's just what I want* in Fig. 10, and the auxiliary-subject inversion triggered by *never* in *Never have I seen such a ghastly look on any man's face* in Fig. 11 cause a considerable slowdown in L2 readers compared to L1 readers.

CHECK	O/E - 1	across 5 OI	across 5 RT
that	0.406	-0.270	1125.308
s	0.406	0.123	948.538
just	0.042	0.310	996.615
what	0.809	1.661	2997.692
I	0.509	2.171	2868.385
want	0.778	2.544	4610.846

Fig. 10: Heat-map for "That's just what I want".

CHECK	O/E - 1	across 5 OI	across 5 RT
Never	0.275	0.629	742.231
have	0.205	0.956	824.000
I	0.651	1.036	702.846
seen	0.215	1.688	905.923
such	0.711	2.057	1093.846
a	0.046	1.828	998.231
ghastly	0.140	1.764	1096.692
look	0.351	1.463	1241.154
on	-0.149	1.100	1147.385
any	0.192	0.580	888.538
man	0.527	1.061	943.154
s	0.527	1.447	817.923
face	0.108	1.204	839.923

Fig. 11: Heat-map for "Never have I seen such a ghastly look on any man's face".

The relative pronoun *what* in Fig. 10 already slows down L2 readers, but zero-relative pronouns are processed with even more difficulties.

CHECK	O/E - 1	across 5 OI	across 5 RT
It	-0.224	0.036	1113.077
was	0.700	0.860	1125.923
one	0.383	1.342	1316.923
of	-0.040	0.791	926.846
the	0.359	1.177	852.538
longest	0.067	1.469	953.846
and	1.000	1.770	889.846
blackest	0.144	1.530	985.923
I	0.637	2.207	967.231
have	0.469	2.317	1231.538
ever	-0.256	1.993	1140.000
seen	0.520	1.513	1251.923

Fig. 12: Heat-map for "It was one of the longest and blackest I have ever seen".

Fig. 12 shows the effect of such a zero-relative clause. The absence of the relative pronoun (*that I have ever seen*) seems to trigger a considerably longer processing

time in non-native readers. The fact that personal pronouns in the nominative case are often a good indicator for being subjects of a subordinate relative clause may be better known to native readers than to language learners, who have less routine.

Rare vocabulary items are a difficulty that L2 readers often face – the probability that it is unknown or in the case of *pince-nez* in Fig. 13 may be retrieved via the other foreign language French creates a delay. Observe that this delay is much more local (across 5 OE drops to and even below 1 three words later) than the one seen in Fig. 12, where a large region of surrounding words is affected (across 5 OE stays above 1.5 until the end of the sentence).

CHECK	O/E - 1	across 5 OI	across 5 RT
He	0.392	1.761	935.154
wore	0.352	1.476	1300.308
gold-rimmed	0.283	1.291	1626.923
pince-nez	0.282	1.829	2523.308
and	0.057	1.367	2459.385
had	0.321	1.296	2617.615
a	-0.087	0.857	2237.385
curious	0.058	0.631	1986.846
impassivity	0.266	0.615	1512.385
of	-0.582	-0.024	1335.462
feature	-0.059	-0.404	1506.846

Fig. 13: Heat-map for "He wore gold-rimmed pince-nez and had a curious impassivity of feature".

Nominalisations, particularly if they occur in a very formal register, can challenge L2 readers. The old-fashioned formulation *in the main* can be seen in Fig. 14. The frequency of *in the main* reduces in the corpus of historical American English (COHA, Davies 2010) from 0.1 per 10000 words around the year 1900 to only 0.04 around the year 2000. L1 readers have typically had more exposition to rarer registers, literary genres, and retreating constructions.

CHECK	O/E - 1	across 5 OI	across 5 RT
Mrs	0.208	0.883	1468.538
Inglethorp	0.306	1.031	1439.615
kind	-0.190	0.528	1333.154
as	0.479	1.059	1380.385
she	0.488	1.291	1543.923
might	0.084	1.167	1456.308
be	0.772	1.633	1145.077
in	0.585	2.409	1155.077
the	0.493	2.422	1392.000
main	0.162	2.096	1179.000

Fig. 14: Heat-map for "Mrs Inglethorp, kind as she might be in the main".

Fig. 15 shows a heavy nominalisation, *neatness*, in combination with the rare and old-fashioned word *attire*. The zone ending at *attire* takes L2 readers more than twice as long to read.

CHECK	O/E - 1	across 5 OI	across 5 RT
The	0.470	1.489	1095.154
neatness	0.331	1.588	1296.769
of	0.473	1.831	1161.077
his	0.440	1.962	1346.769
attire	0.384	2.098	1433.846
was	0.061	1.689	1304.385
almost	0.020	1.378	1112.769
incredible	0.074	0.979	1299.385

Fig. 15: Heat-map for "The neatness of his attire was almost incredible".

Long attachments, i.e., phrases that are quite far away from their governor, are also more difficult for L2 readers. Fig. 16 gives the example of the subordinate clause *to come* which only starts after two inserted prepositional phrases (*to him* and *over her shoulder*). The structure in Fig. 17 also contains a fronted element (*what*). Fronting has the effect that the distance between the object *what* and its governing verb *been* are quite long. Further, it is a complex phrasal verb, Phrasal verbs are a feature of spoken language, a register with which L2 readers may be less familiar than L1 readers. As a result, processing speed decreases considerably.

CHECK	O/E - 1	across 5 OI	across 5 RT
Cynthia	0.112	0.695	1355.769
called	0.021	0.303	1051.154
to	-0.054	0.480	1029.385
him	0.128	0.326	827.077
over	0.503	0.709	1273.000
her	0.249	0.846	1242.077
shoulder	0.420	1.246	1388.538
to	0.605	1.905	1612.077
come	0.628	2.406	1810.000
and	0.128	2.031	1334.154
join	0.267	2.049	1249.000
us	0.004	1.633	980.615

Fig. 16: Heat-map for "Cynthia called to him over her shoulder to come and join us".

CHECK	O/E - 1	across 5 Ol	across 5 RT
Have	-0.128	-1.650	359.538
some	0.547	-0.738	681.692
coffee	0.098	-0.878	806.846
and	0.270	0.391	965.077
tell	0.190	0.976	1203.000
us	0.799	1.904	1159.769
what	0.319	1.675	1077.000
you	0.634	2.212	1048.231
have	-0.109	1.833	986.308
been	0.252	1.895	941.923
up	0.910	2.006	978.308
to	0.108	1.795	896.923

Fig. 17: Heat-map for "Have some coffee and tell us what you have been up to".

Infrequent constructions are generally difficult. This is illustrated in Fig. 18, which gives the example of a participial, *having been occupied*.

CHECK	O/E - 1	across 5 Ol	across 5 RT
and	0.678	1.043	1094.923
that	0.309	1.000	1027.923
there	0.547	1.497	1215.538
was	0.201	1.623	1031.154
no	0.321	2.057	1103.385
sign	0.272	1.651	1090.385
of	0.345	1.687	1069.769
the	0.820	1.960	1153.000
room	0.252	2.011	1398.769
having	0.321	2.011	1573.692
been	-0.079	1.660	1445.692
occupied	0.107	1.421	1501.615

Fig. 18: Heat-map for " . . . and that there was no sign of the room having been occupied".

Unusual and archaic word meanings are also difficult. The word *gay* with the meaning *joyful* seems to be less familiar to L2 readers, as shown in Fig. 19.

CHECK	O/E - 1	across 5 Ol	across 5 RT
But	0.504	0.781	961.846
they	0.569	1.226	1299.231
were	0.202	1.437	1366.923
both	0.752	2.267	2415.769
gay	0.257	2.284	2414.846
enough	0.253	2.033	2861.615
this	0.120	1.584	2735.923
afternoon	0.288	1.670	3158.154

Fig. 19: Heat-map for "But they were both gay enough this afternoon".

Complex preposition and phrasal verb constructions, as we have already seen in Fig. 16, can lead to slower processing by L2 readers. In Fig. 20, we can see a sequence of four PPs that all attach to the main verb *went*.

CHECK	O/E - 1	across 5 OI	across 5 RT
and	0.334	0.249	982.538
went	0.330	0.587	1095.769
rapidly	0.588	1.114	1612.692
past	0.192	1.149	1605.846
me	0.259	1.703	1661.615
down	0.350	1.720	1875.308
the	0.780	2.169	1791.385
stairs	0.214	1.795	1434.846
across	0.318	1.921	1500.077
the	0.192	1.855	1290.769
hall	0.393	1.897	1187.077
to	0.151	1.269	1006.385
the	0.218	1.273	858.769
boudoir	-0.189	0.765	761.692

Fig. 20: Heat-map for ". . . and went rapidly past me down the stairs across the hall to the boudoir".

Next, as Siyanova–Chanturia/Conklin/Schmitt (2011) have shown, idioms are often more difficult to process for L2 readers than for L1 readers, as Fig. 21 illustrates on the basis of *wit's end for money*. The meaning of this idiom is to be puzzled, and not knowing what to do.

CHECK	O/E - 1	across 5 OI	across 5 RT
I	1.000	0.443	620.462
don	0.183	0.633	474.038
t	0.183	1.063	486.615
mind	0.579	1.731	737.000
telling	0.180	2.126	766.769
you	-0.037	1.089	834.385
that	0.422	1.328	927.962
I	0.351	1.496	908.923
m	0.351	1.267	710.462
at	0.406	1.493	756.923
my	0.165	1.695	768.923
wit	0.519	1.791	818.962
s	0.519	1.959	981.615
end	0.547	2.155	1319.769
for	0.506	2.256	1360.615
money	0.101	2.192	1467.462

Fig. 21: Heat-map for "I don't mind telling you that I'm at my wit's end for money".

Finally, we turn to an example from morphology. The irregular verb *fling* causes a local slowdown for many L2 readers, see Fig. 22.

CHECK	O/E - 1	across 5 Ol	across 5 RT
John	0.155	1.111	1560.538
flung	0.428	1.351	1769.692
the	0.335	1.576	1741.000
match	-0.058	0.912	1436.769
into	-0.027	0.833	1277.923
an	0.046	0.725	1175.923
adjacent	0.093	0.390	1119.154
flower	-0.151	-0.096	957.000
bed	-0.311	-0.349	881.000

Fig. 22: Heat-map for "John flung the match into an adjacent flower bed".

6 Discussion and Conclusion

We have used language models such as surprisal and regression as a cognitive model in order to predict RT of native speakers (L1) and language learners (L2) with a linear regression method using eye tracking data, especially the GECO corpus. Our goal is both application-driven, aiming to accurately predict reading behaviour of L1 and L2 readers, and also cognitive, aiming to assess the most important factors, and the differences between L1 and L2 readers. Let us revisit our research questions from the introduction again.

In addition, we have seen that individual variation between the readers is very strong. Fast readers exhibit a better model fit, potentially because they can concentrate better on the task. L1 readers are faster than L2 readers, and L1 readers exhibit a better model fit, they are more efficient readers both in terms of speed and model fit. RT predictions are off by 34–40% in our linear regression. Our prediction errors are considerably below individual variation, which means that our models predict a plausible reader.

We have also seen some evidence on which constructions are harder for L2 in the qualitative results section. Let us revisit our research questions from the Introduction.

1) Which features correlate to reading times?
We have seen that all four features that we selected (word length, presence of punctuation, distance to previous occurrence of same word, surprisal) are highly correlated to RT and are significant predictors in a regression model. There are strong correlations between reading times and surprisal, but there are also other

factors. In particular, for predicting reading times, we have seen the following order of features:

Word length > punctuation ≥ distance to previous occurrence of same word > surprisal.

Word length is a trivial predictor, longer words simply take longer to read. Presence of punctuation, mostly commas and full stops, lead to significantly slower RT, because the meaning of the clause is processed by the reader. The discourse-related feature of the distance to the previous last occurrence of the same word shows how much knowledge of the semantic background of the individual discourse, here a novel, helps readers to integrate new information, and how much introduced entities are salient on the readers' mind, expecting their re-appearance (Church 2000). To be able to assess the impact of discourse was a motivation for collecting the GECO corpus, with the aim "to evaluate the generalizability of . . . language theories and models to the reading of long texts and narratives" (Cop et al. 2017: 602). We could profit from this potential in our study.

Surprisal, although a highly significant feature, turns out be less important than the discourse feature of last occurrence of the same word. While the strong influence of surprisal is well known (Demberg/Keller 2008; Smith/Levy 2013) we could place it more precisely in the hierarchy of significant factors. The ranking that we obtained by linear regression was also confirmed by stepwise regression (section 4.2.4) and by feature ablation experiments (section 4.2.2).

2) Are the features and their weights similar for L1 and L2 readers?
The order of feature weights is similar, but there are two notable differences: first, surprisal is less significant for L2 readers than for L1 readers. This result is in line with Underwood/Schmitt/Galpin (2004), Siyanova–Chanturia/Conklin/Schmitt (2011), and Schilk (2017), in which L2 readers found idioms and formulaic word sequences more difficult to process, even if considering retrieval time for individual words. In other words, the lower level of routinization of L2 readers is apparent, as already anticipated by Pawley/Syder (1983). Second, the presence of punctuation symbols (mostly these are commas and full stops) is a less important feature for L2 readers than for L1 readers. In terms of F-value, punctuation is three times stronger than the distance to the previous occurrence of the same word, while for L2 readers, these two features are similarly important. It seems that native speakers manage better to read clauses as a single unit. This observation also supports the view that idiomatic units are processed faster and as single units by L1 readers, and that they exhibit a less linear reading behaviour, pausing at semantic boundaries rather than at difficult words or constructions, as we have qualitatively assessed in section 5.

3) Is the individual variation between the readers bigger or smaller than the difference between L1 and L2?
The difference between L1 and L2 RT has a mean of 67 ms. As the standard deviation of RTs is 43 ms for L1 and 34 ms for L2, the between-group differences are only slightly bigger. In other words, individual variation is very strong in L1 and in L2. The very strong individual variation means that individual reading times are not very strongly correlated: the reading time of a different reader is as good a predictor as surprisal. At the same time, individual variation is too unsystematic to serve as a useful random effect in a mixed model. This is why we decided to pool the readers as and predict RT means across the individuals.

Pooling participants is less common than using a mixed-effects model, in which the individual is a random effect. Experiments with mixed models on GECO (Schneider accepted) revealed, however, that individual variation is not systematic. The standard deviation of the random effect of the individuals is more than seven times smaller than the residual. We thus use pooling participants as a noise reduction method. While the method of predicting average reading time can be seen as a shortcoming, it also offers a number of attractive characteristics. First, it allows us to keep a simpler, parsimonious model. Second, for the task of predicting typical reading times, irrespective of individual behaviour, for instance as a proxy to reading difficulty, it is an appropriate and simple smoothing technique. Third, it leads to better performance in downstream applications aiming to model typical readers (Hollenstein 2020; Klerke/Plank 2019). Hollenstein (2020) states that for the aim of predicting typical readers, averaging is a good option: "The eye movement measurements were averaged over all native-speaking readers of each dataset to obtain more robust estimates." (41). Fourth, averaging greatly reduces the number of skipped words, for which the data set gives reading times of 0 ms, about 39% of all words are skipped, be that due to parafoveal reading (Rayner 1998) or a low sampling rate of the eye tracker (Andersson/Nyström/Holmqvist 2010). When using readers' means for each word, less than 1% of all words have RT of 0 ms in the GECO corpus. The fact that there are very few words that are skipped by all readers is a further indication that individual variation may be viewed as noise (unsystematic variation) rather than a signal (systematic variation).

4) Can L2 reading times be predicted similarly well as L1 times?
Both model fit and prediction accuracy of L2 readers is much lower than of L1 readers. R^2 for the prediction of L1 RT is 0.475, but only 0.332 for L2. L2 readers show more variability, less systematicity, and are harder to predict. This is also related to the observation that slower readers generally have lower correlation to surprisal, indicating that they are less efficient not only in the task of reading

but probably also in knowing word sequences, idioms, and how a sentence is likely to continue.

5) Do increased reading times of L2 readers reveal to us which constructions are particularly taxing for L2 readers?
In the qualitative analysis in Section 5 above, we presented a selection of phenomena that stood out in the heat-map visualisation, and gave our interpretation. Salient phenomena that take L2 readers longer to process include fronting (non-canonical word order), zero constituents, rare words and constructions, nominalisations, long attachments, unusual word meanings, complex preposition and phrasal verbs, idioms, and irregular morphology. WE could detect these differences between L1 and L2 readers in a data-driven fashion, without selecting candidate phenomena beforehand.

Our study has several limitations. First, the list of features that we have selected, following Demberg/Keller (2008), Smith/Levy (2013), and Schneider (in press) is unlikely to be complete. We spent considerable time on testing further features, but some strong predictors may have escaped us, and we have also excluded two significant features, POS tag and tagger confidence, which we would like to include in future studies. Second, the fact that the L2 data contains native speakers of Dutch and no other language may add a bias. Particularly as Dutch is typologically related to English, our observations cannot be generalized to very different L1 languages. It would be interesting to include Readers with native languages from non-Indo-European backgrounds, for instance Finnish or Basque. Also, languages with considerably freer word order and stronger inflectional systems (e.g., Russian or German) or head-final languages like Japanese would be a desideratum.

We envisage many applications of our research, ranging from cognition to stylistics, automatic style checking and essay grading, understanding learner language, and language simplification.

Future research should include the significant feature of POS tag (section 4.2.1), more syntactic features, and further language models like BERT or neural networks. In cognitive linguistics, we would like to further distinguish pragmatic effects of world knowledge, for instance by including word embedding, discourse knowledge (our feature of the last occurrence of the same word, but also adding anaphora resolution), language sequence and idioms (surprisal) and syntactic features. For a language learning application, one can focus on phenomena and words that L2 readers find particularly hard, both generally, or from specific L1 backgrounds. Also on the individual level, eye tracking or self-paced reading reveals weaknesses and important study areas to which a given student should give particular focus.

References

Andersson, Richard/Nyström, Marcus/Holmqvist, Kenneth (2010): "Sampling frequency and eye–tracking measures: How speed affects durations, latencies, and more." In: *Journal of Eye Movement Research* 3, 1–12.

Altenberg, Bengt (1998): "On the phraseology of spoken English: The evidence of recurrent word combinations." In: A. P. Cowie (Ed.): *Phraseology. Theory, Analysis, and Applications*. Oxford: Oxford University Press.

Aston, Guy/Burnard, Lou (1998): *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Biber, Douglas/Conrad, Susan (2009): *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Bresnan, Joan/Cueni, Anna/Nikitina, Tatiana/Baayen, Harald (2007): "Predicting the dative alternation." In: Gerlof Bouma/Joost Zwarts/Irene Krämer (Eds.): *Cognitive Foundations of Interpretation*. Amsterdam: Royal Netherlands Academy of Science, 69–94.

Bresnan, Joan/Nikitina, Tatiana (2009): "The gradience of the dative alternation." In: Linda Uyechi/Lian Hee Wee (Eds.): *Reality Exploration and Discovery. Pattern Interaction in Language and Life*. Stanford: CSLI Publications, 161–184.

Bybee, Joan (2007): *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.

Conklin, Kathy/Schmitt, Norbert (2012): "The processing of formulaic language." In: *Annual Review of Applied Linguistics* 32, 45–61.

Conklin, Kathy/Pellicer–Sánchez, Ana/Carrol, Gareth (2018): *Eye–Tracking. A Guide for Applied Linguistics Research*. Cambridge: Cambridge University Press.

Church, Kenneth (2000): "Empirical estimates of adaptation: The chance of two noriegas is closer to p/2 than p^2." In: *Proceedings of the 17th conference on Computational linguistics*, 180–186.

Cop, Uschi/Dirix, Nicolas/Drieghe, Denis/Duyck, Wouter (2017): "Presenting GECO: An eye tracking corpus of monolingual and bilingual sentence reading." In: *Behavior Research Methods* 49, 602–615.

Davies, Mark (2010–): *The Corpus of Historical American English (COHA)*. Online at: https://www.english–corpora.org/coha/ <14.04.2022>.

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

Divjak, Dagmar/Levshina, Natalia/Klavan, Jane (2016): "Cognitive linguistics: Looking back, looking forward." In: *Cognitive Linguistics* 27, 447–463.

Ellis, Nick C. (2013): "Construction grammar and second language acquisition." In: Thomas Hoffmann/Graeme Trousdale (Eds.): *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press, 365–378.

Ellis, Nick/Frey, Eric/Jalkanen, Isaac (2009): "The psycholinguistic reality of collocation and semantic prosody (1): Lexical access." In: Rainer Schulze/Ute Römer (Eds.): *Exploring the Lexis–Grammar Interface*. Studies in Corpus Linguistics. Amsterdam: John Benjamins, 89–114.

Frank, Stefan L./Monsalve, Irene F./Thompson, Robin L./Vigliocco, Gabriella (2013): "Reading–time data for evaluating broad–coverage models of English sentence processing." In: *Behavior Research Methods* 45, 1182–1190.

Frank, Stefan L. (in press): "Towards computational models of multilingual sentence processing." *Language Learning* (special issue *What is special about multilingualism?*).
Glynn, Dylan/Fischer, Kerstin (2010): *Quantitative Methods in Cognitive Semantics. Corpus–Driven Approaches*. Berlin/New York: Mouton de Gruyter.
Grön, Gerbrand J. (1996): "Cognitive slowing in patients with acquired brain damage: An experimental approach." In: *Journal of Clinical Experimental Neuropsychology* 18, 406–415.
Goldberg, Adele E. (2006): *Constructions at Work. The Nature of Generalization in Language*. Oxford: Oxford University Press.
Graën, Johannes/Alfter, David/Schneider, Gerold (2020): "Using multilingual resources to evaluate CEFRLex for learner applications." *Proceedings of the 12th Language Resources and Evaluation Conference*, 346–355. Online at: https://www.aclweb.org/anthology/2020.lrec-1.43.pdf <14.11.2022>.
Gries, Stefan (2015): "The most under–used statistical method in corpus linguistics: Multi–level (and mixed–effects) models." In: *Corpora* 10, 95–125.
Gries, Stefan Th./Wulff, Stefanie (2005): "Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora." In: *Annual Review of Cognitive Linguistics* 3, 182–200.
Hilpert, Martin (2019): "Constructional approaches." In: Bas Aarts/Jill Bowie/Gergana Popova (Eds.): *The Oxford Handbook of English Grammar*. Oxford: Oxford University Press, 106–123.
Hollenstein, Nora (2020): *Leveraging cognitive processing signals for natural language understanding*. Zurich: ETH Zurich dissertation.
Janda, Laura A. (2013): *Cognitive Linguistics. The Quantitative Turn*. Berlin: Mouton de Gruyter.
Kennedy, Alan/Hill, Robin/Pynte, Joël (2003): *The Dundee Corpus*. Paper presented at the 12th European Conference on Eye Movement, Dundee, Scotland.
Kennedy, Alan/Pynte, Joël/Murray, Waine/Paul, Shiley–Anne (2013): "Frequency and predictability effects in the Dundee Corpus: An eye movement analysis." In: *Quarterly Journal of Experimental Psychology* 66, 601–618.
Klerke, Sigrid/Plank, Barbara (2019): "At a glance: The impact of gaze aggregation views on syntactic tagging." In: *Proceedings of the Beyond Vision and LANguage. inTEgrating Real–world kNowledge (LANTERN)*, 51–61.
Larsen-Freeman, Diane (1997): "Chaos/complexity science and second language acquisition." In: *Applied Linguistics* 18, 141–65.
Larsen–Freeman, Diane/Cameron, Lynne (2008): *Complex Systems and Applied Linguistics*. Oxford: Oxford University Press.
Levy, Roger/Jaeger, T. Florian (2007): "Speakers optimize information density through syntactic reduction." In: *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*.
Luke, Steven G./Christianson, Kiel (2018): "The Provo Corpus: A large eye–tracking corpus with predictability norms." In: *Behaviour Research Methods* 50, 826–833.
MacWhinney, Brian J. (2001): "Psycholinguistics: Overview." In: Neil J. Smelser/Paul B. Baltes (Eds.): *International Encyclopedia of the Social & Behavioral Sciences*. Pergamon, 12343–12349.
MacWhinney, Brian/Bates, Elisabeth (Eds.) (1989): *The Crosslinguistic Study of Sentence Processing*. New York: Cambridge University Press.

Molnar, Christoph (2020): *Interpretable. Machine learning: A guide for making black box models explainable*. Unpublished manuscript. Online at: https://christophm.github.io/interpretable–ml–book/ <01.10.2021>.

Nesselhauf, Nadja (2005): *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

Newman, John/Rice, Sally (2010): *Experimental and Empirical Approaches in the Study of Conceptual Structure, Discourse, and Language*. Stanford: CSLI Publications.

Norman, Rocío S./Shah, Manish N./Turkstra, Lyn, S. (2019): "Reaction time and cognitive–linguistic performance in adults with mild traumatic brain injury." In: *Brain Injury* 33, 1173–1183,

Pawley, Andrew/Syder, Frances Hodgetts (1983): "Two puzzles for linguistic theory: Native–like selection and native–like fluency." In: *Language and Communication* 191–226.

Racine, John P. (2014): "Reaction time methodologies and lexical access in applied linguistics." In: *Vocabulary Learning and Instruction* 3, 66–70.

Rayner, Keith (1998): "Eye movements in reading and information processing: 20 years of research." In: *Psychological Bulletin* 124, 372–422.

Rayner, Keith/Duffy, Susan A. (1986): "Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity." In: *Memory and Cognition* 14, 191-201.

Rayner, Keith/Raney, Gary E./Pollatsek, Alexander (1995): "Eye movements and discourse processing." In: Robert F. Lorch Jr./Edward. J. O'Brien/Robert F. Lorch (Eds.): *Sources of Coherence in Reading*. Hillsdale, NJ: Lawrence Erlbaum Associates, 9–36.

Rodríguez, Germán (2020): *Introducing R*. Unpublished manuscript, Princeton University. Online at: https://data.princeton.edu/R <01.10.2021>.

Sahlgren, Magnus (2006): *The Word–Space Model: Using distributional Analysis to represent syntagmatic and paradigmatic relations between words in high–dimensional vector spaces*. Stockholm: Stockholm University dissertation.

Saville–Troike, Muriel/Barto, Karen (2016): *Introducing Second Language Acquisition*. 3rd edn. Cambridge: Cambridge University Press.

Segalowitz, Norman S./Segalowitz, Sidney J. (1993): "Skilled performance, practice, and the differentiation of speed–up from automatization effects: Evidence from second language word recognition." In: *Applied Psycholinguistics* 14, 369–385.

Schilk, Marco (2017): *Language processing in advanced learners of English: A multi–method approach to collocation based on corpus linguistic and experimental data*. Hildesheim: Universität Hildesheim Habilitation Thesis.

Schneider, Gerold (in press): "Correlations between reading times, collocation and surprisal." In: Manfred Krug/Ole Schützler/Fabian Vetter/Valentin Werner (Eds.): *Perspectives on Contemporary English*. Peter Lang: Series Bamberg Studies in English Linguistics.

Schneider, Gerold/Gaëtanelle, Gilquin (2016): "Detecting innovations in a parsed corpus of learner English." In: *International Journal of Learner Corpus Research* 2, 177–204.

Schneider, Gerold/Grigonyte, Gintare (2018): "From lexical bundles to surprisal and language models: Measuring the idiom principle on native and learner language." In: *Patterns in text: Corpus–driven methods and applications*. Studies in Corpus Linguistics Series. John Benjamins.

Schneider, Gerold/Lauber, Max (2019): *Statistics for Linguists*. Zürich: Pressbooks.

Shain, Cory (2019): "A large–scale study of the effects of word frequency and predictability in naturalistic reading." In: *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol 1. 4086–4094.

Shannon, Claude E. (1951): "Prediction and entropy of printed English." In: *The Bell System Technical Journal* 30, 50–64.

Sinclair, John (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Smith, Nathaniel/Levy, Roger (2013): "The effect of word predictability on reading time is logarithmic." In: *Cognition* 128, 302–319.

Siyanova–Chanturia, Anna/Conklin, Kathy/Schmitt, Norbert (2011): "Adding more fuel to the fire: an eye–tracking study of idiom processing by native and non–native speakers." In: *Second Language Research* 27, 251–272.

Speelman, Dirk/Heylen, Kris/Geeraerts, Dirk (Eds.) (2018): *Mixed–Effects Regression Models in Linguistics*. New York: Springer.

Underwood, Geoffrey/Schmitt, Norbert/Galpin, Adam (2004): "The eyes have it: An eye movement study into the processing of formulaic sequences." In: Norbert Schmitt (Ed.): *Formulaic Sequences. Acquisition, Processing and Use*. Amsterdam: John Benjamins, 153–172.

Winter, Bodo (2013): *Linear Models and Linear Mixed Effects Models in R with linguistic applications*. Online at: https://arxiv.org/pdf/1308.5499.pdf <01.10.2021>.

Wulff, Stefanie (2008): *Rethinking Idiomaticity*. London: Continuum.