Robustness and processing difficulty models. A pilot study for eye-tracking data on the French Treebank

Stéphane Rauzy, Philippe Blache

Robustness and processing difficulty models. A pilot study for eye-tracking data on the *French Treebank*

Stéphane RAUZY Philippe BLACHE
Aix-Marseille Université & CNRS
Laboratoire Parole & Langage
Aix-en-Provence, France
stephane.rauzy@lpl-aix.fr, philippe.blache@lpl-aix.fr

ABSTRACT
We present in this paper a robust method for predicting reading times. Robustness first comes from the conception of the difficulty model, which is based on a *morpho-syntactic surprisal* index. This metric is not only a good predictor, as shown in the paper, but also intrinsically robust (because relying on POS-tagging instead of parsing). Second, robustness also concerns data analysis: we propose to enlarge the scope of reading processing units by using syntactic chunks instead of words. As a result, words with null reading time do not need any special treatment or filtering. It appears that working at chunks scale smooths out the variability inherent to the different reader's strategy. The pilot study presented in this paper applies this technique to a new resource we have built, enriching a French treebank with eye-tracking data and difficulty prediction measures.

KEYWORDS: Linguistic complexity, difficulty models, morpho-syntactic surprisal, reading time prediction, chunks.

1 Introduction

Eye-tracking data are now often used in the study of language complexity (e.g. difficulty metrics evaluation) as well as finer syntactic studies (e.g. relative complexity of alternative constructions). However, only few resources exist, for a small number of languages. We describe in this paper a pilot study aiming at developing a high-level resource enriching a treebank with physiological data and complexity measures. This work have been done for French, with several objectives : (1) building a new large resource for French, freely available, associating syntactic information, eye-tracking data and difficulty prediction at different levels (tokens, chunks and phrases) (2) validating a difficulty model for French in the line of what has been done for other languages (Demberg and Keller, 2008), (Boston et al., 2008) relying on a robust surprisal index described in (Blache and Rauzy, 2011).

This pilot study, on top of building a new resource, had important side-effects. First, this work led us to examine carefully the question of *data analysis*. In particular, we found that working with larger units (syntactic chunks) instead of tokens makes it possible to take into consideration the entire set of data. In other words, it is not anymore necessary to eliminate data that are usually considered for different reasons as problematic (tokens ending lines, before punctuations, etc.). This result is important for several reasons. First, it avoids the use of truncated data (which is problematic in a statistical point of view). Second, it supports the hypothesis that chunks are not only functional, but can also be defined in linguistic terms by means of syntactic relation strength. Another interesting result is the influence of the syntactic parameter on the global model: we show that (morpho)syntax has modest impact in comparison with word frequency and word length. Finally, at the technical level, we have developed an entire experimental setup, facilitating data acquisition when using *Tobii* devices. Our environment proposes tools for the preparation of the experimental material (slide generation) as well as data post-processing (e.g. lines model detection).

2 Background

The study of language complexity first relies on theoretical difficulty models. Several proposals can be found in the literature, exploring the influence of different parameters on the parsing mechanism (Gibson, 1998, 2000), (Hawkins, 2001), (Vasishth, 2003). One important problem is the possibility to quantify the difficulty level: some metrics have been proposed such as *Dependency Locality Theory* (Gibson, 1998) which uses the number of new discourse referents in an integration region. Evaluating such models relies on the comparison of similar constructions, one being known to be more difficult than another (for example, object vs. subject relative clauses). Prototypical examples of such alternations are built and the model applied incrementally, estimating at each word the integration costs. Such models rely on high-level linguistic information, capable of bringing together syntactic and lexical semantic information, as well as integrating frequency information. In such cases, difficulty estimation is done manually, the validation applied only to few examples.

Recently, the development of probabilistic NLP techniques opened a new way in difficulty estimation. The idea consists in using the probability of the integration of a word into a partial parse as a predictor for human difficulty. The *Surprisal* index (Hale, 2001) implements this proposal: the mechanism consists in evaluating at each word the difference between probability of the set of trees before the word and that integrating the word. Several works such as (Demberg and Keller, 2008) have shown that *Surprisal* can be a predictor for reading time and, as a consequence, for language processing difficulty. The interest in these experiments is that,

thanks to automatic difficulty evaluation, it becomes possible to work on larger amounts of data, offering the possibility to study language in more natural contexts.

We present in the remaining of this section an overview of different works addressing this question and propose an analysis of their characteristics, in particular with respect to the kind of data they use.

2.1 Experimental evaluations of complexity models

(Demberg and Keller, 2008) proposes an evaluation of two syntactic complexity theories (*DLT* and *Surprisal*) for the prediction of readers difficulty. Linear mixed effects models are experimented, taking into account non-syntactic predictors besides complexity measures. Such predictors are low-level variables known to have an impact on reading times[1]: word frequency, word length, position in the sentence (final words in the sentence are read faster). Oculomotor variables also have to be considered: fixation of a previous word, number of characters between two fixations, position of the fixation in the word. Higher level contextual variables are also proposed: forward transitional probability (probability of a word knowing the previous one) and backward transitional probability (probability of a word knowing the next one). As for the surprisal parameter, two different version have been used: one calculating surprisal taking into consideration the word forms, the other the POS tags. The experimental data rely on the English part of the Dundee corpus (Kennedy et al., 2003). This corpus comprises 51,502 tokens, from 20 newspaper articles (from *The Independent*). Eye-tracking data have been acquired for 10 subjects. Different eye-tracking measures are considered: *first fixation duration (FFD)* in a region, *first pass duration (FPD)* (total of all the fixations in a region when reading it for the first time) and *total reading time (TRD)* of a region (all the fixations, including those when going back into a region that has already been read).

In the experiment, (Demberg and Keller, 2008) eliminates from the original corpus several data: first and last tokens of each line, token followed by a punctuation, region of 4 words with no fixations and words with zero value for FFD and FPD . Finally, this experiment retains a total of 200,684 data points, which means 20,068 tokens read by 10 subjects.

The results of this study show that unlexicalized surprisal can predict reading times, whereas the lexicalized formulation does not. However, (Monsalve et al., 2012) pointed out recently that when using independent sentences, both lexicalized and unlexicalized surprisal measures are significant predictors of reading time (measures done with corpus of around 2,500 words and 54 participants).

These different studies focus on lexical and syntactic effects. In a complementary direction, (Pynte et al., 2009) analyzed the influence of superficial *lexical semantics* on fixation duration. (Mitchell et al., 2010) integrates this parameter into *Surprisal*. This work shows the effect of semantic costs in addition to syntactic surprisal for reading time prediction. It also addresses in a specific way the question of modeling: experimental studies usually use linear mixed effect models, including random effects (e.g. participants characteristics) and fixed ones (e.g. word frequency). In these approaches, many different parameters are brought together. As authors pointed out, the use of a unique measure for predicting complexity is preferable than a set of factors, not only for simplicity, but also because it is difficult to analyze the effective contribution of a factor: one can evaluate whether adding it into a model improves it fits, but cannot explain the reasons.

[1] See (Demberg and Keller, 2008) p.196 for a precise description.

2.2 Parameters and data

The different experiments have shown that *Surprisal* can play a significant role in a complexity model. All such models bring together different parameters at different levels: oculomotor (positions of the fixations), lexical (properties of the lexical items) and syntactic (contextual characteristics). Moreover, surprisal presents the advantage to be calculated for lexical items (taking into account the specific properties of each token, including co-occurrence) as well as POS, the last case being apparently more robust.

The complexity models in these different studies are linear mixed-effects and make use of many predictors. The following table recapitulates the main parameters used in the different studies[2]:

	Demberg08	Mitchell10	McDonald03	Monsalve12	Boston08	Roark09
Word length	+	+	+	+	+	+
Word freq.	+	+	+	+	+	+
Sentence position	+			+		
Word position				+		
Landing position	+	+	+			
Launch distance	+	+				
Previous word RT	+	+		+		
Lexicalized surp.	+			+	+	
Unlexicalized surp.	+			+		+
Bigram prob.		+	+		+	
Forward trans.	+		+	+		
Backward trans.	+		+	+		
Integration costs	+					
Lexical surp. entropy					+	
Synt supr. entropy					+	
Derivation steps					+	
Semantic		+				
Predictability			+		+	
Retrieval					+	

Arbitrarily, we distinguish in this table between low and high level predictors, the first usually being the baseline. As expected, word length and word frequency are used in all considered models, other predictors being less systematic. One can observe that the combinatory is very important and many different models have been experimented.

By another way, these experiments have shown the importance of input data. Until recently, studies on linguistic complexity was done on controlled material (artificially built sentences, out of context, small corpora). *Surprisal* relying on well-known NLP techniques, it offers the advantage to be applied to unrestricted corpora. (Demberg and Keller, 2008) evaluates this measure against a large corpus of newspaper articles, which constitutes an important step towards the treatment of *natural data* (even though the idea of contextualized material has been challenged by (Monsalve et al., 2012)). However, the main problem with the size of input data is that only few corpora with eye-tracking data are available. The Dundee corpus is, to the best of our knowledge, the only one with a reasonable size in a NLP perspective. Other existing corpora are much smaller, such as the *Embra* (McDonald and Shillcock, 2003) which comprises around 2,600 words. Another problem when dealing with large amount of data is the sensibility of the measures to parsers efficiency. No precise indication is given in these works, in spite of the fact that this constitutes a big issue (parsers F-scores being usually close to 85%).

A last feature shared by these different experiments lies in data cleaning. For different reasons, large part of the input material is excluded: position in the line, fixation duration, even in some

[2]For sake of place, these predictors are not described here. Their definition can be found in the corresponding papers.

cases morpho-syntactic category. Even though such pre-processing is usual in psycholinguistics, it constitutes a problem, in particular in terms of data analysis, as it will be explained later.

3 Experiment

As shown in the previous section, corpus used in the different experiments are very different in size and nature. (Demberg and Keller, 2008) explicitly focuses on naturalistic data. On the opposite, (Boston et al., 2008) relies on a very small corpus, but with large amount of subjects. The following table presents the main features of the different corpora. It mentions the number of token presented to the readers, the number of subjects participating to the experiment, the number of data points (roughly speaking fixation points) taken into account in the evaluation (after eliminating problematic data), the average number of tokens read by the subjects and taken into account after data filtering (data points are more or less the number of participants times the number of remaining tokens) and the experimental method.

	Tokens	Participants	Data points	Remaining tokens	Method
Demberg08	51,502	10	200,684	20,000	Eye-tracking
Mitchell10	5,370	10	53,704	5,300	Eye-tracking
McDonald03	2,262	23	31,242	1,350	Eye-tracking
Monsalve12	?	54	132,298	2,449	Self-paced reading
Boston08	1,138	222	167,499	754	Eye-tracking
Roark09	883	23	20,309	883	Self-paced reading

For similar study on French, there exists only one resource (the French part of the Dundee corpus (Kennedy et al., 2003)), but which is not publically available. This situation leads us to the project to build a new large resource for French, associating syntactic information, eye-tracking data and difficulty prediction. The pilot study presented hereafter has been realized in order to check the viability of the overall project.

3.1 Experimental design

One of our goal is to validate the experimental design. Our pilot study consisted in acquiring eye-movement data for 13 subjects reading an extract of the French Treebank (herefater *FTB*, (Abeillé et al., 2003)). The *FTB* is a set of articles from the newspaper *Le Monde*. Most of these articles are in the economical field, which does not fit well with the idea of natural reading. However, we selected from this corpus several extracts that seemed to us less technical in terms of semantic contents.

The eye-tracking device is a *Tobii 60 Hz* [3]. The selected subcorpus used in this experiment is made of 6 articles of variable length (from 3 to 6 minutes of reading time), each of them presented to the reader as a succession of slides. Participants have to press a key to access to the next slide. Once the key pressed, an empty frame with a target cursor indicating the position of the first line beginning the next slide is presented during three seconds, followed by the text slide. A calibration of the *Tobii* machine is proposed before reading each article and a three minutes pause between articles has been observed, filled by an informal discussion with the experimenter about the content of the article. The overall session last 45 minutes in average for each participant.

Each slide contains from 4 to 7 lines. Sentences were constrained to appear on a single slide, and the text is not right justified, tokens too long to enter the current line are printed on the

[3] In parallel, we will compare our data with their counterparts obtained using an Eye-link II system (these data are on the process to be acquired at LLF by B. Hemforth).

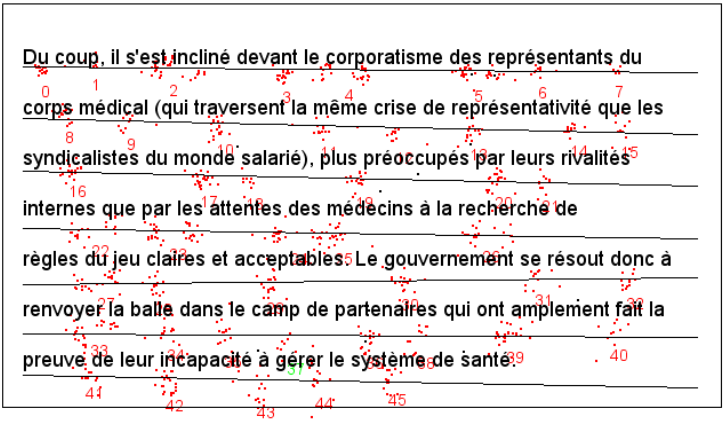

Figure 1: *An example of the slides presented. Red dots give the gaze positions recorded by the Tobii system at a* 17 *milliseconds rate. Horizontal lines represent the lines model fitted to the slide. The lines model allows to associate the gaze fixations (the clusters of points appearing on the figure) with the words of the text.*

next line. The text is printed on 800 × 600 pixels slides using an Arial font of size 18 with line spacing of size 26 pixels (an example is presented figure 1). The participant is positioned at a 60 cm distance of the screen, which implies a 30 pixels precision on *Tobii* measurements or equivalently a two characters horizontal precision and half line spacing in vertical precision.

The design of the experiment has been done thanks to a software we have developed (the generic designing software coming with *Tobii* being not suited for a full-text reading experiment). Our system automatically generates the slides and associates to each word its size in pixels as well as its precise spatial location. This renders straightforward the specification of each word (or set of words) as "*area-of-interest*" for the eye-tracking system. The overall corpus is made of 80 slides, 198 sentences split on 549 lines, which contains 6,572 tokens (5,696 words and 876 punctuation marks), which comes to 75,077 data points (a reasonable size in comparison with existing resources, see previous section).

3.2 Data post-processing

Our software also takes in charge data post-processing. In particular, one of the main problem consists in associating a sequence of eye movements with a line: the fact that backward movements (i.e. regressions) as well as line jumps are frequent renders difficult the association of a fixation aera with a word. We developed a specific algorithm to fit a line model to gaze measurements (see figure 1). The lines model allows to establish a geometrical relation between the set of fixations and the tokens of the slide. A parameter measures the quality of the fit. It is used to discard the slides for which the matching between fixations and tokens is problematic. For the present pilot experiment, the ratio of discarded slides reaches 12%. However, all the slides presented possess valid measurements for at least 9 participants over the group of 13 subjects.

Fixations are formed from individual gaze measurements by use of standard clustering tech-

niques[4]. The minimal duration time has been fixed to 85 ms and a maximal clustering length of 30 pixels has been adopted (or a two characters length, which is the precision of the *Tobii* device). First and last fixations of the slide are trimmed if problematic (e.g. at the end of the reading, it is not rare that the reader's gaze wanders over the slide before pressing the next slide key). We therefore obtain the list of fixations and their associated parameters (position, starting time, ending time, ...). Thanks to the lines model (which gives the line the fixation belongs to) and the horizontal coordinate of the fixation, the closest token is associated with the fixation. Herein, we choose to associate fixation only to words, so by construction punctuation marks have zero reading time.

From the fixations list, we collect for each token of the slide and for each participant the oculomotor quantities of interest such as the first pass duration time, total reading time, number of fixations, and so on. This information is enriched for each token by metric and positioning information (length in pixels, number of characters, line index, position index in the line, ...) and later on in the analysis with linguistic information (morphosyntactic category, lexical frequency, ...). For the overall $6,572$ tokens of the corpus, we finally obtain $75,077$ oculomotor measurements for the set of 13 participants ($10,359$ over $85,436$ have been discarded due to lines model problem). Among them, $34,598$ have a null total duration time ($11,388$ correspond to punctuation marks, the $23,210$ remaining correspond to skipped words, i.e. words with no associated fixation). The ratio of skipped words (over the total number of words) is around 36% for our corpus of french newspaper articles.

The comparison of our pilot experiment with similar works (e.g. the french part of the Dundee corpus (Kennedy et al., 2003)) does not reveal significant difference concerning the global reading parameters such the mean fixation duration, saccade ratio, regression ratio, ... It means that the experimental setup chosen (e.g. large font size, spacious layout, ...), even if far from ecological reading condition, does not pertube the participant reading task. Similarly, the low sampling rate (one measurement each 17 milliseconds) and the relatively poor spatial precision of the Tobii device does not affect the average values of the global reading parameters. An accurate comparison of the Tobii and Eye-link II results will be conducted as soon as the Eye-link II data will be available for our reading material.

4 Analysis

The analysis relies on the paradigm that the reading times are a tracer of the linguistic complexity. In the present pilot study, our main objective restricts to study what can we learn about linguistic difficulty from reading time measurements. In particular, to model the reading strategy (e.g. when and where fixations occur) is out of the scope of the analysis. Therefore, the model we propose does not contain low-level variables describing reading strategy except the word length which accounts for the time spent to decode the characters of the words.

Motivations leading us to choose this strategy are twofold. First, we desire to draw robust conclusions concerning the linguistic difficulty, independent of a peculiar choice for the model describing the reading strategy. Second, as far as possible, we will try to limit the number of variables entering the statistical model. Indeed, the difficulty to interpret the resulting fitted values of a linear model (mixed or not) increases with the number of dimensions (i.e. the number of variables), especially when all these variables are strongly statistically dependent.

[4]A complete presentation of the algorithms implemented herein as well as a comparison with the state-of-the-art (see (Holmqvist et al., 2011)) will be proposed in a forthcoming paper.

In that case, the parameters space becomes highly instable, and the addition (or removal) of one variable in the model may dramatically change the resulting fitted coefficients. This effect has to be avoided since the final interpretation eventually relies on the values of these fitted coefficients.

In the following subsection, we introduce the basic ingredients of the model. The multivariate regression analysis is performed subsection 4.2 where the main results are discussed.

4.1 The variables of the model

4.1.1 Reading time

In the present study, we will focus on the total reading time measurement, defined as the sum of duration lengths for all the fixations on the area spanned by the token, including backward fixations.

In order to compare the token reading times measured for the different participants, we will first proceed to a normalization. Each participant P possess its own reading velocity $V(P)$ which can be estimated on the corpus. For each participant, the sum over the slides not discarded of the tokens total reading time $D(P)$ and tokens length $L(P)$ (for example the length in pixels) are computed. The mean reading velocity of the participant is then given by $V(P) = L(P)/D(P)$. By introducing the average reading velocity over the participants $\overline{V}$, we can form the normalized total reading time for token t and participant P :

$$D(t,P) = \frac{V(P)}{\overline{V}} \times \texttt{total reading time}(t,P) \tag{1}$$

Note that this transformation affects also the minimal threshold of 85 milliseconds (i.e. the minimal duration for a fixation).

Since participants were asked to read the same texts, it could be interesting to introduce the notion of *average reader*. The token reading time of the average reader $\overline{D}(t)$ is defined as the average of the normalized reading times over the participants (when this measurement is available) :

$$\overline{D}(t) = \sum_{\mathrm{P}} D(t,P) \Big/ \sum_{\mathrm{P}} 1 \tag{2}$$

It has been observed (Lorch and Myers, 1990) that averaging over participants is source of information loss for the low-level variables describing reading strategy (e.g. landing position, launch distance, ...). However, we are herein not concerned by this potential problem since low-level variables are not included in our model.

4.1.2 Word length

Reading times are known to depend on the word lengths (see (Rayner, 1998) for a review of the literature). For a token t, we choose to include this metric information by considering the number of characters of the token :

$$L(t) = \texttt{number of characters}(t) \tag{3}$$

The $L(t)$ variable accounts for the time spent to decode the characters of the token. Other metric information (landing position, previous word fixated, ...) is herein not considered.

4.1.3 Lexical information

The frequency of the word is another variable of our model. Frequent words are red faster, which can be interpreted either as a lexical access facility or as a predictability effect. The variable used herein is minus the logarithm of the lexical probability of the token form :

$$F(t) = -\log P(\texttt{form(t)}) \tag{4}$$

This quantity is computed from the frequencies obtained in the LPL French lexicon augmented by the words of the *French Treebank*. Tokens not in the lexicon (punctuation marks, numbers, ...) have received a special treatment.

4.1.4 Morphosyntactic surprisal

The classical surprisal model being very sensitive to the parser performance, we use a new measure relying on morphosyntactic analysis (Blache and Rauzy, 2011). The idea consists in making the same kind of differential measure as for surprisal (Hale, 2001), but using POS-tagging instead of parsing.

POS-tagging builds during the process a set of solutions for the sequence of tokens. Each solution corresponds to an alternative when associating the set of morphosyntactic categories (tags) to the lexical form of the token (POS). Let's call $Sol_i(t)$ the i^{th} solution at position t,

$$Sol_i(t) = c_{1,i}...c_{t,i} \tag{5}$$

where $c_{t,i}$ is the morphosyntactic category associated to the token at position t for solution $Sol_i(t)$. The probability of the solution $Sol_i(t)$ is obtained recursively by Bayes formulae :

$$P(Sol_i(t)) = P(c_{t,i}|Sol_i(t-1)) \times P(Sol_i(t-1)) \tag{6}$$

where $P(Sol_i(t-1))$ is the probability of the solution i at position $t-1$ and $P(c_{t,i}|Sol_i(t-1))$ is the conditional probability of category $c_{t,i}$ given the left context $Sol_i(t-1) = c_{1,i}...c_{t-1,i}$. The relative contribution of each solution can be obtained thanks to the introduction of the density function $\rho_i(t)$:

$$\rho_i(t) = \frac{P(Sol_i(t))}{A(t)}, \texttt{ with } A(t) = \sum_i P(Sol_i(t)) \tag{7}$$

Following (Hale, 2001), the morphosyntactic surprisal at position t for each solution $Sol_i(t)$ is :

$$S_i(t) = -\log \frac{P(Sol_i(t))}{P(Sol_i(t-1))} = -\log P(c_{t,i}|c_{1,i}...c_{t-1,i}) \tag{8}$$

and the overall surprisal is :

$$S(t) = \sum_i \rho_i(t) S_i(t) \tag{9}$$

The morphosyntactic surprisal is an *unlexicalized* surprisal (see (Demberg and Keller, 2008)) in the sense that it does not capture the lexical probability of the form (that information is however included in the model section 4.1.3). The morphosyntactic surprisal accounts for two distinct types of difficulty: one related to the predictability of the proposed tag in context (high predictibility leads to low surprisal), the other coming from the effective number of solutions

maintained in parallel due lexical form ambiguity (the higher is this effective number, the higher is the surprisal).

Without entering into details (a complete presentation can be found in (Blache and Rauzy, 2011)), the contextual probabilities entering equations 6 and 8 are learned on the *GraceLPL* French corpus augmented by the *French Treebank*. Adding the corpus under treatment allows to avoid infinite value for surprisal (e.g. the cases present in the corpus to tag but no met in the training corpus).

4.2 Model and results

The aim is herein to quantify the relative effects of the variables mentioned above on reading time measurements. At first approximation, a simple linear model is assumed :

$$D = \alpha_L L + \alpha_F F + \alpha_S S + D_0 + \epsilon \tag{10}$$

where the slopes α_L, α_F and α_S measure the strength of the effect of the explanatory variables L, F and S respectively, D_0 is the intercept of the model and the residuals ϵ account for what remains unexplained by the model.

4.2.1 Analysis at the token scale

We applied a multivariate linear regression to the 75,077 individual normalized reading time measurements. For convenience, the explanatory variables have been previously scaled (zero mean and unit variance), in such way that the slope gives directly the strength of the effect on the duration time. All the slopes are found positive (which was expected) and highly significant. However, a closer analysis reveals that the residuals of the model are strongly dependent on the predicted values (see figure 2).

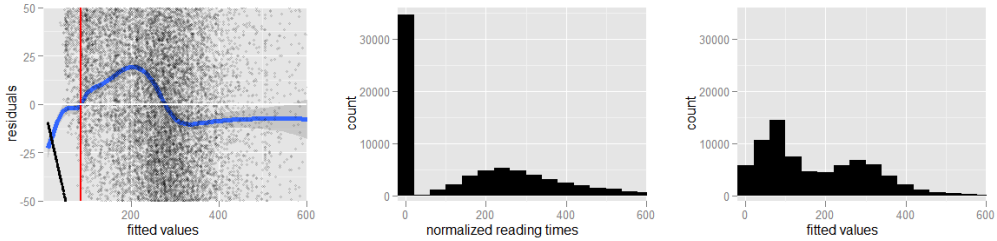

Figure 2: *For individual reading time measurements, the residuals of the linear model fit are plotted versus the fitted values. For a valid fit, the moving average of residuals (blue curve) is expected to match the x-axis within its error bars. The minimal fixation duration is represented by the vertical red line. The histogram of the normalized reading times and of the fitted values are also shown.*

The inspection of the normalized reading times and predicted values histograms of figure 2 explains why the linear model fails to fit reading time measurements. About 46% of the tokens have null reading time (67% of them are skipped words, the remaining 33% consists in punctuation marks which have null reading time by construction). The explanatory variables entering the right term of equation 10 does not present such discrete bimodal distribution.

There is therefore little hope that a linear combination of these variables can successfully describe the data.

In order to minimize the problem of null reading times, two modifications are brought to the model. First, a binary parameter N_{pm} which specify whether the token is a punctuation mark or not is added to the linear model, i.e.

$$D = \alpha_L L + \alpha_F F + \alpha_S S + \alpha_{pm} N_{pm} + D_0 + \epsilon \tag{11}$$

The second modification concerns the reading times to fit. Because of the average over the participants, the average reading times introduced section 4.1.1 is less susceptible to present a bimodal distribution. The multivariate regression is thus applied on the $6,572$ average reading times of the corpus including the binary parameter to deal with punctuation marks. The results are presented figure 3. The modified linear model is unable to describe the average reading times distribution. As expected, the distribution of the average reading times does not present the bimodal trend of the individual reading times histogram. However, the same dependency is found between the predicted values and residuals of the fit: short predicted reading times are predicted not enough short and long ones not enough long. This observation suggests that skipped words are not just skipped because they are frequent and short (in that case, the model will have explained the effect) and that this skipping word strategy is shared by the group of participants. The linear model misses an ingredient to account for this effect.

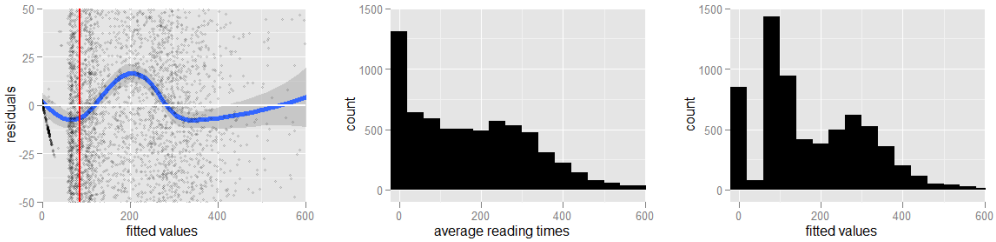

Figure 3: *Same plots as figure 2 for the average reading times over the participants.*

The problem is mainly due the presence of null reading time measurements in the data. One solution could be to remove them from the analysis. However statistics on truncated data, even if feasible in theory (see for example (Breen, 1996)), are often a tricky business in practice. Because a part of the genuine distribution has been removed, standard statistical recipes do not apply securely and the estimators of the model parameters are found biased in general. While some techniques exist to correct on these bias, they may require a full knowledge of the explanatory variables distributions and their dependencies, which is difficult to achieve in practice. We will not pursue this way.

A second solution could be to make use of a reading model which account for the skipped word phenomena. However again, our original aim was to make use of reading time measurements to learn about the syntactic complexity. As far as it is possible, we would like that our conclusions remain independent of a particular choice concerning the reading model used. We propose next subsection an alternative solution.

4.2.2 Analysis at larger scale

Our alternative solution is based on the following remark. All the variables entering the linear model are *extensive variables*[5], which means that they are globally additive under scale change. For example, the total duration time for a group of N tokens is the sum of the individual total reading time of the N tokens. Similarly, the property holds for the tokens length, the tokens frequency and as mentioned by (Smith and Levy, 2008), for the surprisal measure. Therefore, nothing prevents us to change the scale of the analysis, by considering group of adjacent tokens rather than working at the token scale.

We experimented this approach by forming groups of consecutive tokens (with the additional constraint that the tokens belong to the same line). The mutivariate regressions were performed on the summed quantities (summed average reading times, summed lengths, ...). The dependency between the predicted values of the fit and the residuals decreases as the size of the group increases. The fit becomes acceptable above the scale of 5 tokens. At this scale, it seems that the erroneous predicted reading times compensate each others (i.e. short versus long reading times) and provide us with a valid prediction for the reading time of the group as a whole.

This observation leads us to search for a natural scale grounded on linguistic information. Figure 4 displays for each morphosyntactic categories the boxplot of the number of participants having fixated the tokens. We remark that two populations emerges: the content words (adjectives, adverbs, nouns and verbs) with a high fixated count and the function words (determiners, auxiliaries, prepositions, ...) with a low fixated count.

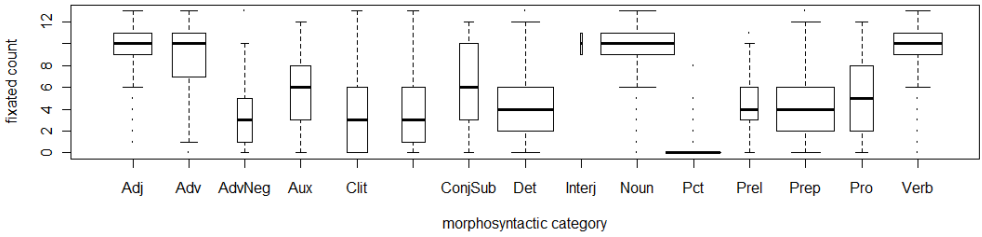

Figure 4: *Boxplot of the number of participants having fixated the tokens in function of the morphosyntactic category of the tokens.*

In the field of syntax, there exists a unit which groups the function words with their associated content word: the *chunk* (Abney, 1991). It remains to check whether the chunk scale is a good candidate for our analysis. Because chunks have variable sizes, we added to the linear model the variable N which represents the number of tokens in the chunk. The equation becomes :

$$D = \alpha_L\, L + \alpha_F\, F + \alpha_S\, S + \alpha_{pm}\, N_{pm} + \alpha_N\, N + D_0 + \epsilon \tag{12}$$

Our corpus contains $2,842$ chunks, the average number of tokens by chunk is 2.31. The results of the multivariate regression fit are shown figure 5. A slight dependency of the residuals is still

[5]The notion of *extensive* versus *intensive* variables comes from Thermodynamics and Statistical Physics.

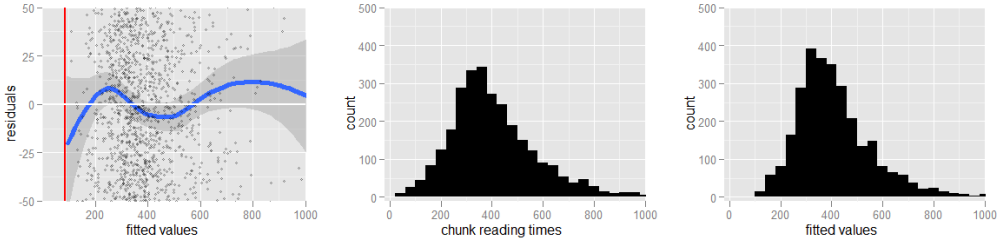

Figure 5: *Same plots as figure 2 for average reading times at the chunk scale. The grey envelope represents 1-σ error bars on the moving average.*

Variable	Estimate	Std. Error	Pr(>|t|)
(Intercept)	422.775	2.307	<2e-16 ***
L_{scaled}	89.388	2.798	<2e-16 ***
F_{scaled}	91.527	4.429	<2e-16 ***
S_{scaled}	22.345	2.696	<2e-16 ***
N_{scaled}	-35.382	4.618	2.51e-14 ***
$N_{punctuation}$	-37.156	4.248	<2e-16 ***

Table 1: The slopes, standard errors and statistical significance for the variables entering the linear fit.

present (the maximal amplitude is about 7 milliseconds on residuals), but its effect has been considerably lessening if compared with the analysis at the token scale (see figure 3).

Table 1 summarizes the amplitudes of the effect for each variable of the linear model. The residuals standard error is of 101 ms and the multiple R-squared of 0.687. In average, a chunk is red in 422 ms. The influence of the chunk length and the chunk frequency are of the same order (around 90 ms, or 20% of the average reading time). The contribution of morphosyntactic surprisal is slighter, 22 ms or 5% of the signal. A negative effect is found for the number of tokens. At equal values for length, frequency and morphosyntactic surprisal, chunks containing more tokens are red slower. Note that the amplitudes of all these effects are considerably larger than the 7 ms maximal dependency bias remaining in the fit. We can thus conclude securely that these effects are real.

5 Results and perspectives

The first goal of this work was to develop and evaluate a difficulty model based on *morpho-syntactic surprisal*. The results obtained with eye-tracking data show that our model is a good reading time predictor. This result is interesting for several reasons. First, it replicates for French similar results obtained for other languages. Second, it shows that morpho-syntactic surprisal is a good predicting variable. Because this difficulty measure is very robust and independent from any syntactic formalism, it is possible to use for any linguistic material, including spoken language: this opens the way to future experiments on predicting difficulty in natural interaction.

Evaluating this model led us to other interesting theoretical, methodological and technical results. In particular, we have shown that it is possible to keep all original data, including null

reading time tokens. Variables of the linear model being additive under scale change, it becomes possible to take into consideration set of tokens as fixation area. Interestingly, considering syntactic chunks as fixation area provides very good result (reducing in a considerable extent the dependency of the residuals). This observation allows to avoid the important data reduction usually applied by other works. Moreover, it gives an experimental support to the idea that reading is done at he level of chunks instead of words.

More generally, these results has to be situated in the perspective of the development of a generic difficulty model that would integrate (1) parameters from different linguistic domains and (2) high level effects such as *cumulativity* (Keller, 2005) or *compensation* (Blache, 2011), increasing or decreasing difficulty. Our objective with such a generic model is to answer at three questions: where, how and why difficulties occur. This long-term goal is based on the idea that the basic elements of the integration process are variable in granularity: this process can indeed relies on words, but also on larger units such as phrases, prosodic units or discursive segments.

Last, but not least, this study led to the construction of a high-level linguistic resource: a treebank enriched with eye-tracking data plus difficulty measures. Such resource will be of great interest in the perspective of the new field of experimental syntax.

Acknowledgments

Vera Demberg, Stéphanie Ducrot, Sophie Dufour and John Hale are cheerfully thanked for fruitful discussions.

References

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for french. In Abeillé, A., editor, *Treebanks*, Kluwer, Dordrecht.

Abney, S. (1991). Parsing by chunks. In *Principle-Based Parsing. Kluwer Academic Publishers*, pages 257–278.

Blache, P. (2011). Evaluating language complexity in context: New parameters for a constraint-based model. In *CSLP-11, Workshop on Constraint Solving and Language Processing*.

Blache, P. and Rauzy, S. (2011). Predicting linguistic difficulty by means of a morpho-syntactic probabilistic model. In *Proceedings of PACLIC 2011, december 2011*, Singapour.

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

Breen, R. (1996). *Regression models : Censored, Sample selected or Truncated Data*. Sage Publications Ltd.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. In *Cognition*, volume 109, Issue 2, pages 193–210.

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

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In *Image*. A. Marantz, Y. Miyashita, W. O'Neil (Edts).

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

Hawkins, J. (2001). Why are categories adjacent. *Journal of Linguistics*, 37.

Holmqvist, K., Nystrom, M., Anderson, R., Dewhurst, R., Jaradzka, H., and van de Weijer, J. (2011). *Eye Tracking: A comprehensive guide to methods and measures*. Oxford Press.

Keller, F. (2005). Linear Optimality Theory as a Model of Gradience in Grammar. In *Gradience in Grammar: Generative Perspectives*. Oxford University Press.

Kennedy, A., Hill, R., and Pynte, J. (2003). The dundee corpus. In *Proceedings of the 12th European Conference on Eye Movements*.

Lorch, R. F. and Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(1):149–157.

McDonald, S. A. and Shillcock, R. C. (2003). Low-level predictive inference in read- ing: The influence of transitional probabilities on eye movements. *Vision Research*, 43.

Mitchell, J., Lapata, M., Demberg, V., and Keller, F. (2010). Syntactic and semantic factors in processing difficulty: An integrated measure. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 196–206.

Monsalve, I. F., Frank, S. L., and Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceeding of EACL*.

Pynte, J., New, B., and Kennedy, A. (2009). On-line contextual influences during reading normal text: The role of nouns, verbs and adjectives. *Vision Research*, 49(5):544–552.

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

Smith, N. J. and Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 595–600.

Vasishth, S. (2003). Quantifying processing difficulty in human sentence parsing: The role of decay, activation, and similarity-based interference. In *Proceedings of the European Cognitive Science Conference 2003*.