Human and computer estimations of Predictability of words on written language

Bruno Bianchi[1], Facundo Carrillo[2], Diego Fernndez Slezak[2], Juan E. Kamienkowski[1,2], and Diego E. Shalom[1]

[1]Laboratorio de Neurociencia Integrativa, Universidad de Buenos Aires, Argentina
[2]Laboratorio de Inteligencia Artificial Aplicada, Universidad de Buenos Aires, Argentina

Abstract

When we read printed text, we continuously predict the follow words in order to integrate information and direct future eye movements to forthcoming words. Thus the Predictability has become one the most important variables when explaining human behavior and information processing during reading. In this study we present results of word predictability in long Spanish texts, estimated from human responses in a massive web-based task. We used Latent Semantic Analysis (LSA) as a way to estimate human-based predictability values computationally. We validated the human estimation of predictability with local and global properties of the text, and we showed that LSA-distance on adequate timescale captures some semantic aspects of the prediction.

1 Introduction

One of the most challenging paradoxes in human cognition is that, although we sample the environment in a very discrete manner, both in space and time, we perceive both as continuous and smooth. This is done by continuously integrating the past and predicting the upcoming stimuli. A special case of this capacity is used in reading, where humans move their eyes word-by-word. Future eye movements through the text are planned, based on expectancy of the upcoming words. Thus, the Predictability has become one the most important variables when explaining eye movements and information processing in reading [10, 16, 6, 7, 15]. The classical approach to measure Predictability in neurolinguistics studies is asking participants to complete a sentence with the word they believe that follows. This simple task is called *Cloze-Task*[18]. A caveat with this measure is that it's very expensive in time and effort. During the last decade, several computational alternatives have been proposed and evaluated to understand and replace the Cloze-Task's Predictability measure. For example: The pure brute force while estimating *Transition Probabilities* [5, 9, 12], pure semantic relatedness measures as *Latent Semantic Analysis* [11, 14, 13], or pure grammar based measures as *Surprisal* [1, 2, 4]. But, none of them has been completely successful. The main reason of this is probably because humans make use of a combination of various features to predict the following word. A good model that approximates the Predictability could give us not only a cheaper way to estimate and use the Predictability in neurolinguistics studies, but also an insight on how the brain uses those semantic and syntactic queues in the process. Finally, brain inspired prediction of the forthcoming word could be used to improve aid typing applications, as for example currently used in cell phones.

In the present work we first aimed to measure the Predictability in an Spanish corpus of short stories (for which we also had the measures of eye movements in a separate study). And then, we aimed to find a better estimation of the Predictability, incorporating new measures from larger corpus, and combining the existing measures.

2 Massive measures in humans

As a starting point we measure the Cloze-task Predictability in human behavior as it is usually done in psycholinguistics studies, in order to get a fair comparison with previous bibliography. Since we use corpus in which we also have measures of eye movements while participants read those texts, we can compare not only with humans aware reports of Predictability but also with human unaware eye movements during reading.

2.1 Methods

2.1.1 Cloze task measures of Predictability

The traditional measure of Predictability in humans is performed using a modified Cloze task. Participants have to read an incomplete sentence and report the word. For example, participants read "I want to climb a .....", and they report a single word for that blank space. Then the Predictability of a word given that context is defined dividing the number of reports of that word over the total number of reports. After the report, the sentence is completed until the following missing word. A minimum number of responses (of the order of N=10) is needed for each individual word. For increasingly long texts this task becomes quickly impractical.

The Predictability of the original word is finally calculated as the number of responses that matched that word over all responses, thus Predictability ranges between 0 and 1. But because it is usually accumulated around values near zero, the Predictability is typically presented as the *logit* transformation: $logit(p) = log\left(\frac{p}{1\ p}\right)$.

2.1.2 Implementation in the web

The eight stories included in our corpus are about 3300 words long (range: 1975-4640). First we select a subset of target words (range: 836-2559) per text that were selected based on a separate eye movement study (we excluded first and last word of each sentence, first and last word of each line, and words that were repeated more than 10 times in each the text). We end up with 12289 target words to evaluate. Hence we need about 120000 reports to be able to estimate the predictability of all our target words. This task is largely impractical to perform in laboratory conditions. We implemented it as an on-line game platform (Figure 1, Upper-left). Each participant completed between 50 and 125 words in each text, which corresponds to one word every about 37 presented words. The distribution of inter-response time has a reasonable log-normal distribution (Figure 1, Lower-left), with a mode at 11 seconds, which correspond to a reading speed of the order of 200 words per minute.

We used a modified Django template, which used a Posgre database for data storage. A Twitter Bootstrap framework was used to improve the graphical design, to confer a modern view to the site. We spread the link through various mailing lists (Google groups, private contacts, etc.) and different social networks (Facebook, Twitter, etc). A prize of 200 Argentinean pesos was raffled every two weeks among the participants, and the winners were announced by e-mail and in the social networks. The responses are mostly driven by these announcements (Figure 1, Lower-right, vertical dashed lines correspond to the dates of the draws). Up to date, 965 subjects have logged in, a total of 107000 words were completed. We have 1184 completed texts (we consider that a text is "complete" if more than 50% of the target words were answered) (Figure

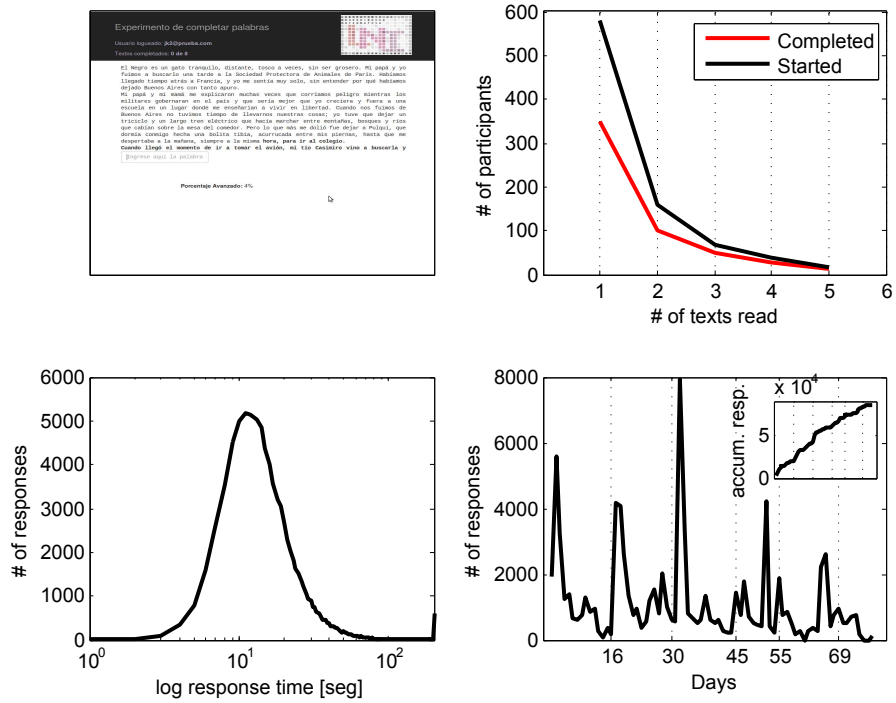

Figure 1: General description of the web experiment. Upper-left: A screenshot of the web interface. Upper-right: Number of text completed (and started) by each participant. Lower-left: Histogram of the inter-response time. Lower-right: Number of responses vs time.

1, Upper-right). Currently, 5 out of the 8 stories have enough data for us to begin the statistical analysis.

2.2 Results and discussion

As it usually happens with every single feature that characterize a word, it covariates with many other features. Two of the most common features are the word Frequency and the word length, which present a strong correlation: most frequent words tend to be also short words. Predictability is also expected to have strong correlations with other features of the words. We estimate the Frequency of the word in the lexicon from the LEXESP corpus [17]. As expected, the Predictability increases with the Frequency, indicating that participants tend to report more Frequent words (Figure 2, Upper-left). On the other hand, shorter words are more predictable than longer words, which is also expected from previous works (Figure 2, Upper-right). Interestingly, although shifted, Content and Function words have nearly the same behavior.

Previous works in the fields of reading and linguistics focused their attention to Predictability of words in sentences. In this study we used larger texts (of around 3000 words), thus we can evaluate other factors that affects the Predictability. Two of those factors that potentially affect the Predictability are the Repetition and the relative position of the word in the text [8]. Results show that the Repetition strongly influenced the Predictability (Figure 2, Lower-left). It increases with the repetition number, with a larger slope for the first five words and the it seems to reach a plateau. This effect is weaker for the Function words, which presented a smaller slope. These effects are not observed in for the relative position of the word in the text (Figure 2, Lower-right), indicating that the effect of repetition (that *a priori* is tightly correlated with the position in the

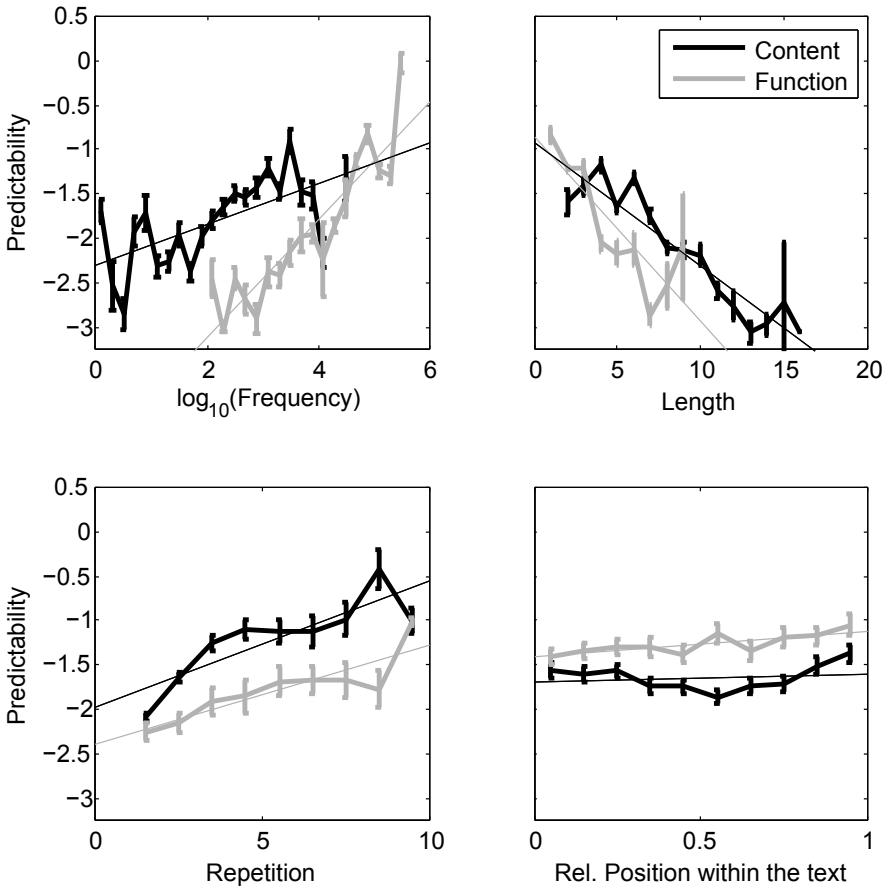

Figure 2: Predictability as function of local and global word properties, for content and function words.

text) holds by itself.

The grammatical category of a word also plays a role (Table 1). Whether the missing word is a Noun or a Preposition, the clues used to predict it are most certainly different in nature. Conjunctions and Prepositions are probably predicted using strong syntactical clues, which leads to high values of Predictability. On the other hand, the clues exploited to predict content words (Nouns, Verbs, Adjectives) are mostly semantic in nature. However, the low value of Predictability of verbs is probably due to the richness of verb conjugations in Spanish. To reduce this effect, lemma predictability was also calculated. While adjectives and nouns are almost insensitive to lemmatization, predictability of verbs increases substantially when lemmatized.

Overall, Predictability measured in this corpus of long texts in Spanish behaves in a reasonable manner, in line with what it is expected from previous works in sentences in English, French and German [6, 7, 15]. This gives us a very good starting point to evaluate different computational algorithms to estimate Predictability in an automatized form. Some efforts in this line are presented in the next section. Briefly, Predictability estimation by humans is clearly a combination of many factors. Semantic, morphological, grammatical and syntactic relations between the future word and its past could be involved, thus a good final algorithm must make some use of those relations.

Table 1: Predictability of grammatical classes

Word Type	Predictability mean (s.e.m.)	Lemma predictability mean (s.e.m.)	Number of responses per word	Number of words
Adjective	11.3% (0.9%)	11.6% (0.9%)	12.0	563
Noun	30.6% (0.8%)	30.7% (0.8%)	11.6	1875
Verb	15.5% (0.6%)	18.0% (0.6%)	11.4	1621
Article	22.2% (1.0%)	28.0% (1.0%)	11.2	705
Determinant	12.2% (1.3%)	17.4% (1.6%)	11.4	231
Pronoun	20.0% (1.2%)	23.7% (1.2%)	11.3	462
Conjunction	32.7% (1.5%)	32.7% (1.5%)	11.1	501
Preposition	37.4% (1.2%)	37.4% (1.2%)	11.1	861
TOTAL	**24.2% (0.4%)**	**25.9% (0.4%)**	**11.4**	**6819**

3 Computational estimates

As a first step into a deep description of the Predictability, we used Latent Semantic Analysis [11, 14, 13] to predict the participants' performance in the online experiment, as function of history length across the text.

3.1 Methods

We used Latent Semantic Analysis (LSA) to measure the semantic relationship between words that participants completed in the online experiment and the preceding words in the text. LSA is a natural language processing technique that proposes that words with close meaning will occur at similar frequency in texts. In short, LSA decomposes a word-by-document occurrence matrix X - with each row corresponding to a unique word in the corpus (n) and each column corresponding to a document (m) - by using Singular Value Decomposition (SVD). Then, the decomposition (U,S,V) is reduced to k dimensions, preserving as much as possible the similarity structure between rows, i.e., preserving the rank of the matrix X. Landauer and Dumais studied the importance of this parameter, concluding that the optimal value of k is around 300 components [11]. Then, LSA-distance between two words is calculated by taking the cosine of the angle between the two vectors corresponding to the words. Values close to 1 represent very related words while values close to 0 represent unrelated words. LSA also depends on the training corpus from where the relation of documents and words is learned. In the present work a set of articles from the Argentinean newspaper "Pgina 12" was used as training corpus [3] (Corpus size 326,466 documents) [1], and k=300 dimensions to trim the decomposition matrix.

In order to predict the participant's responses we calculated the LSA-distance between each responded word and the preceding words. These preceding words were taken from a window of variable size (M). Then, we sum all those values into one LSA value per responded word and window size.

We calculated the Pearson's correlation coefficient between the LSA values and the Predictability for each window size. The significance was estimated with a criterion of p-value $<$ 0.05. The p-value was computed by transforming the correlation to create a t-statistic having N-2 degrees of freedom, where N is the number of rows of X. This is valid for large samples. To evaluate the validity of this assumption we compared the confidence intervals of 95% with this procedure and with a bootstraping procedure (with 5000 replications). The bounds differed in less than 5%, and didn't change the significance.

[1] http://www.pagina12.com.ar/usuarios/anteriores.php

3.2 Results and discussion

As a first effort in assessing the Predictability computationally, we investigated how semantic relations with the history predicts the participant's responses in the online experiment, as function of the window size. We first analyzed the complete dataset of responded words (Fig 3, left, grey), and found significant correlations for all window sizes. These correlations are slightly stronger when we exclude function words (Fig 3, left, black), indicating that LSA captures some aspects of the semantic relations.

On the other hand, when we include in the analysis only the words that were actually part of the original texts (Fig 3, right, grey), no significant correlation is found for any window size. But when we exclude function words, significance is reached for window size 10 and larger (Fig 3, right, black). Interestingly, the LSA estimation of participant's responses increases with history reaching a plateau around the 100 words. This is thought to be the size of a thematic unit, the range in which words correspond to a single semantic neighborhood.

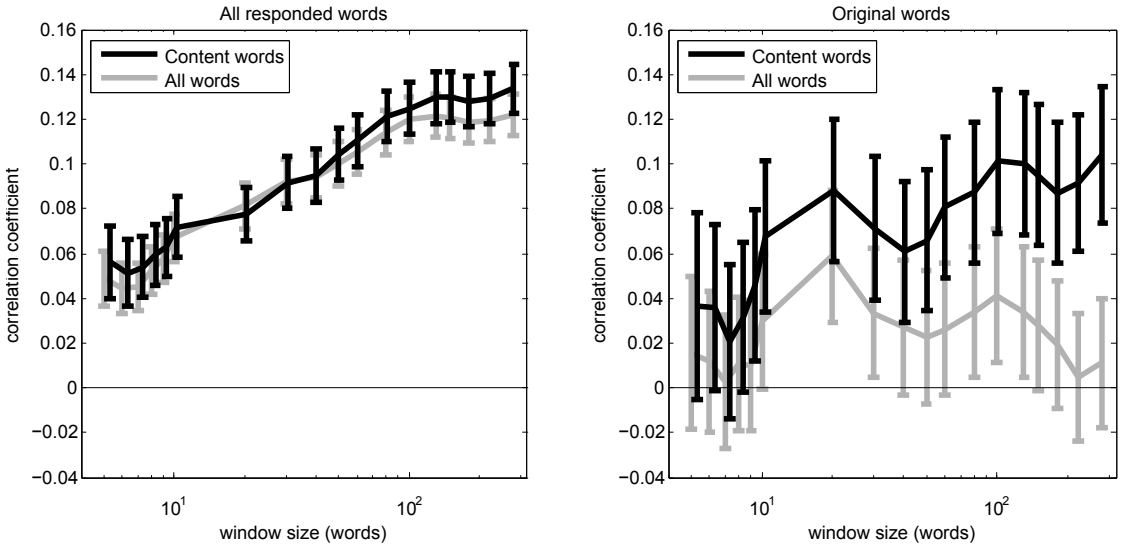

Figure 3: Correlation between LSA values and Predictability as function of window size. Left, all responded words (present or not in the original text) (grey lines), and only content words (black lines). Right, all original words (grey lines), and only content words (black lines). Error bars represent the standard error of the correlation coefficient.

The LSA estimation of participant's responses is better than chance even for very small value of history like 10 words. This indicates that for better predictions of the following word, both near and further past must be taken into account. Our dataset differs from the previous datasets used in testing Predictability estimations in that we used longer texts. This allow us to quantitatively evaluate the influence of larger context or the topic on the estimation.

4 Conclusions and Future Directions

In summary, we presented a dataset where participants predict the following word based on the previous read text. This Predictability was validated studying its known relations with some classical features of isolated word (frequency, length and grammatical category). More interestingly, we showed that it also depends of global variables such as repetition number and

position within the text. These global variables extend current knowledge of isolated sentences further into the range of paragraphs and stories.

Finally, we evaluated the incorporation of LSA-distance to computationally estimate cloze predictability in Spanish text corpus. The novelty is to evaluate the predictability of pieces of text longer than isolated sentences, which allowed us to study not only local context but also global semantic topics in which sentences are embedded. Actually, we observed that LSA-distance is more relevant for window sizes larger than 10 words, while previous works evaluated shorter contexts, using only previous word or current sentence [14].

Further efforts should be directed to incorporating not only semantic aspects but also grammar and syntactic features as predictors. Also, we expect to improve the LSA estimation by expanding the training corpus, in order to have LSA more representative of the whole Spanish language.

References

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

[2] Boston, M.F., Hale, J.T., Vasishth, S., Kliegl, R.: Parallelism and syntactic processes in reading difficulty. Language and Cognitive Processes 26(3), 301–349 (2011)

[3] Carrillo, F., Cecchi, G., Sigman, M., Fernandez Slezak, D.: Evaluation of lsa performance in spanish using multiple corpus of text. In: Proceedings of 14th Argentine Symposium on Artificial Intelligence. JAIIO, Córdoba, Argentina (2013)

[4] Demberg, V., Keller, F.: Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition 109(2), 193–210 (2008)

[5] Frisson, S., Rayner, K., Pickering, M.J.: Effects of contextual predictability and transitional probability on eye movements during reading. Journal of Experimental Psychology: Learning, Memory, and Cognition 31(5), 862 (2005)

[6] Inhoff, A.W., Rayner, K.: Parafoveal word processing during eye fixations in reading: effects of word frequency. Percept Psychophys 40(6), 431–9 (1986)

[7] Just, M.A., Carpenter, P.A.: A theory of reading: From eye fixations to comprehension. Psychological review 87, 329–354 (1980)

[8] Kamienkowski, J.E., Carbajal, M.J., Sigman, M., Shalom, D.E.: Repetition effect within a short stories: An eye movements and linear-mixed models study on a spanish corpus [abstract]. 17th European conference on eye movements. Book of abstracts 17, 432 (2013)

[9] Keller, F., Lapata, M.: Using the web to obtain frequencies for unseen bigrams. Computational linguistics 29(3), 459–484 (2003)

[10] Kliegl, R., Nuthmann, A., Engbert, R.: Tracking the mind during reading: the influence of past, present, and future words on fixation durations. J Exp Psychol Gen 135(1), 12–35 (2006)

[11] Landauer, T.K., Dumais, S.T.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review 104(2), 211 (1997)

[12] McDonald, S.A., Shillcock, R.C.: Eye movements reveal the on-line computation of lexical probabilities during reading. Psychological Science 14(6), 648–652 (2003)

[13] Ong, J.K.Y., Kliegl, R.: Conditional co-occurrence probability acts like frequency in predicting fixation durations. Journal of Eye Movement Research 2(1), 3 (2008)

[14] Pynte, J., New, B., Kennedy, A.: On-line contextual influences during reading normal text: a multiple-regression analysis. Vision Res 48(21), 2172–83 (2008)

[15] Rayner, K., Duffy, S.A.: Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. Memory & Cognition 14(3), 191–201 (1986)

[16] Rayner, K., Warren, T., Juhasz, B.J., Liversedge, S.P.: The effect of plausibility on eye movements in reading. Journal of Experimental Psychology: Learning, Memory, and Cognition 30(6), 1290 (2004)

[17] Sebastián-Gallés, N., Martí, M.A., Cuetos, F., Carreiras, M.: LEXESP: Léxico informatizado del español. Ediciones de la Universidad de Barcelona, Barcelona (1998)

[18] Taylor, W.L.: "cloze procedure": a new tool for measuring readability. Journalism quarterly (1953)