LEXICAL AND PARTIAL PREDICTION IN A BRAZILIAN PORTUGUESE EYE-TRACKING CORPUS

João Vieira[1], Sidney Leal[2], Érica Rodrigues[3], Sandra Aluísio[2], Denis Drieghe[4], Elisângela Teixeira[1].

[1]Universidade Federal do Ceará, [2]Universidade de São Paulo, [3]PUC-Rio de Janeiro, [4]University of Southampton.

Introduction: Besides predicting exact upcoming words (lexical prediction) [1], readers also predict semantic and morphosyntactic information (partial prediction) [2]. Effects of high levels of predictability are regularly reported in the literature, but could be due to text manipulation, common practice in linguistic experimentation, such as in the use of sentences or contexts that trigger strong anticipation of specific words [3]. It has been argued that, in daily life, linguistic comprehension typically does not mirror such high levels of predictability. Corpora of verbal language usually differ from experimentally constructed materials in the sense that they are not built with manipulated stimuli, but with natural passages taken from books, magazines etc. In one such corpus [4], the authors used predictability norms from a Cloze Task for every word in 55 paragraphs in English and analyzed eye movements of participants who read the same paragraphs. The authors found that predictability was influential on language processing even when it was only partially correct, such as when a grammatical category was predictable, but the exact word was not. The authors also found that function words are generally more predictable than content words.

Materials and Methods: To further investigate the influence of predictability in languages in which nominal and verbal inflections differ from English, we built the first corpus of written language processing in Brazilian Portuguese using the eye movement methodology. We focused the analysis on function and content words, while examining both lexical (exact word prediction) and partial prediction. The corpus consists of predictability norms and reading measures of 50 short paragraphs from three different genres: News, Pop-Science and Literary. To calculate predictability norms, 286 participants answered an on-line word-by-word Cloze Task. Each participant answered five paragraphs, except the first word in each paragraph. Eye movements of different 37 participants were recorded using an EyeLink 1000 Hz while they read all paragraphs in a 19' monitor. Paragraphs were authentic and self-contained in meaning. In total, paragraphs had 2494 words (49 on average), out of which 1237 were unique. Targe words (original words) and words answered in the Cloze Task were tagged for part of speech and divided into eight grammatical categories (nouns, adjectives, verbs, adverbs, determiners, prepositions, conjunctions and pronouns), and two classes (content and function).

Results and Discussion: Lexical predictability was measured by comparing the orthography of target and answered words (OrthographicMatch), and for partial predictability, the part of speech tag was compared (POSMatch). Here, we report two eye movement measures closely related to early processing (Gaze duration and skip rates), expected to be sensitive to predictability effects. Lexical prediction was rare, but higher for function words (0.24) than for content words (0.13). Partial prediction was more common and higher for content words (0.44) than for function words (0.38) (Fig. 1). Lexical prediction was higher on News (0.17) and Pop-Science (0.15) texts than on Literary (0.09) texts. We ran linear mixed model analysis on Gaze Duration and logit linear mixed effects on skip rates (Tables 1 and 2). Predictability was facilitative in general, but lexical prediction was more influential than partial prediction. In Fig. 2, we see how Gaze duration dropped as both lexical and partial prediction increased, while Skip Rates increased as lexical prediction increased. Partial prediction did not influence Skip Rates. Lexical prediction had stronger effects when compared to partial prediction in general. Comparing these findings with previous research in English [4], lexical prediction is lower in BP, inviting further investigation. The Cloze Task results also indicate that predictability is involved in everyday language processing, not only when the context is highly restrictive.

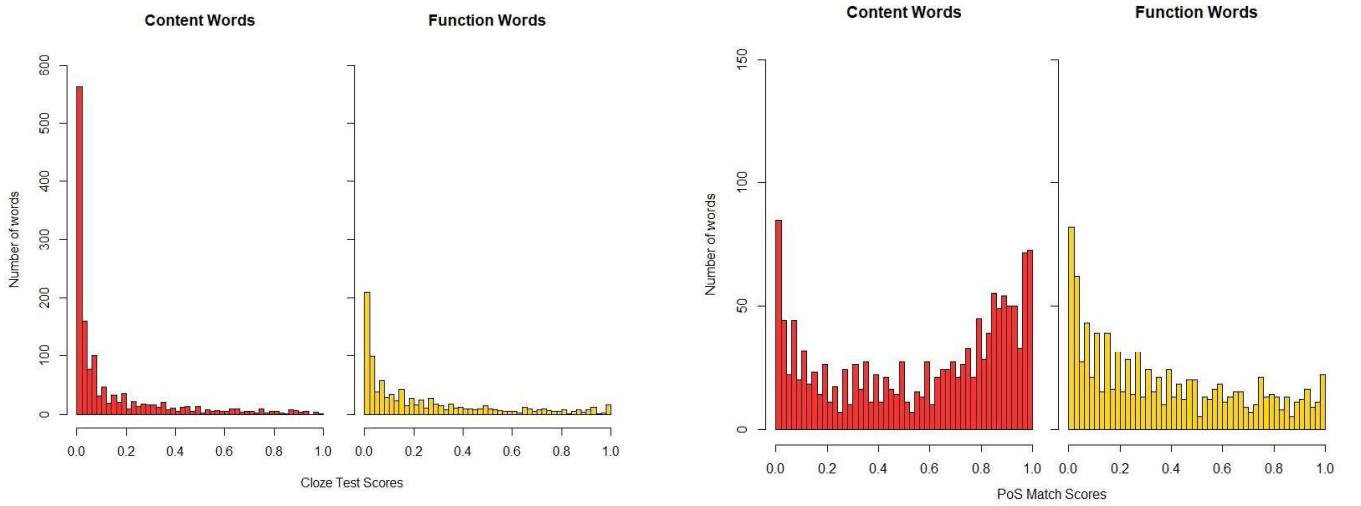

Figure 1. Histogram of lexical (left) and partial (right) predictability of content and function words.

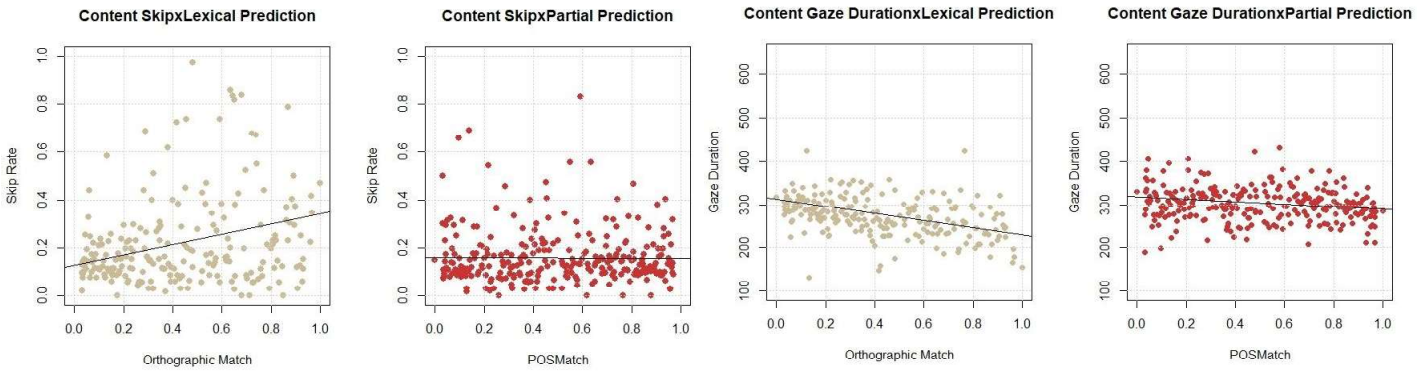

Figure 2. Influence of lexical (beige) and partial (red) predictability on Skip Rates and Gaze Duration.

[1] Calvo, M., & Mesenguer, E. (2002). Eye Movements and processing stages in reading: Relative contribution of Visual, lexical and contextual factors. *The Spanish Journal of Psychology*, 5. 66-77.
[2] Paczynski, M., & Kuperberg, G. R. (2011). Electrophysiological Evidence for Use of the Animacy Hierarchy, but not Thematic Role Assignment, During Verb Argument Processing. *Language and cognitive processes, 26*(9), 1402–1456.
[3] Huettig, F., & Mani, N. (2015) Is prediction necessary to understand language? Probably not. *Language, Cognition And Neuroscience*, 31, n. 1, p.19-31.
[4] Luke, S. G., & Christianson, K. (2016) Limits on lexical prediction during reading. *Cognitive Psychology*, 88, p.22-60.

		b	SE	df	t value	p value
Content words	(Intercept)	5,62	0,02	36,11	264,96	< 0.001
	Lexical Pred.	-0,28	0,01	41450,00	-24,14	< 0.001
	Partial Pred.	-0,04	0,01	41450,00	-5,14	< 0.001
Function words	(Intercept)	5,36	0,02	39,44	281,07	< 0.001
	Lexical Pred.	-0,12	0,02	13660,00	-5,76	< 0.001
	Partial Pred.	-0,04	0,02	13660,00	-2,09	0,03

Table 1 - Linear Mixed Model for Gaze Duration (First Run Dwell Time) on content and function words.

		b	SE	z value	p value
Content words	(Intercept)	-1,68	0,08	-22,07	< 0.001
	Lexical Pred.	1,91	0,1	19,34	< 0.001
	Partial Pred.	-0,67	0,07	-8,84	< 0.001
Function words	(Intercept)	0,11	0,06	1,81	0,07
	Lexical Pred.	1,23	0,07	17,766	< 0.001
	Partial Pred.	-0,05	0,06	-0,77	0,44

Table 2 - Logit Mixed Model for Skip Rates on content and function words.