Lexical Quality as a Proxy for Web Text Understandability

Luz Rello
Web Research & NLP Research Groups
DTIC, Universitat Pompeu Fabra
Barcelona, Spain
luzrello@acm.org

Ricardo Baeza-Yates
Yahoo! Research Barcelona
Barcelona, Spain
rbaeza@acm.org

ABSTRACT

We show that a recently introduced lexical quality measure is also valid to measure textual Web accessibility. Our measure estimates the lexical quality of a site based in the occurrence in English Web pages of a large set of words with errors. We first compute the correlation of our measure with Web popularity measures to show that gives independent information. Second, we carry out a user study using eye tracking to prove that the degree of lexical quality of a text is related to the degree of understandability of a text, one of the factors behind Web accessibility.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: User/Machine Systems.

General Terms

Experimentation, Human Factors, Measurement.

Keywords

Lexical Quality, Understandability, Web Accessibility.

1. INTRODUCTION

Lexical quality broadly refers to the degree of quality of words in a text (spelling errors, typos, etc.) and it is related to the degree of readability of a website [4]. Although lexical quality is not used directly as an accessibility metric, we propose that including text quality and correctness in accessibility metrics could be useful, since the quality of words and language impacts the readers understanding. That is, lexical quality maps to the WCAG principle of content being "understandable" [3].

Our approach is mainly inspired by the work of Gelman and Barletta [5] that apply a spelling error rate as a metric to indicate the degree of quality of websites. They use a set of ten frequently misspelled words and hit counts of a search engine for this set. While they focus on spelling errors, we have previously established an original classification of lexical errors in English motivated by their relationship with textual accessibility, such as the errors made by people with dyslexia, using that to define a new measure of lexical quality (LQ) [1, 2].

In this paper we show that our LQ measure is related to Web text accessibility and hence could be included as an additional measure in quantitative Web accessibility standards.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

2. MEASURING LEXICAL QUALITY

Our error classification for English distinguishes between regular spelling, typographical, non-native speakers, dyslexic and optical character recognition (OCR) errors. Native and non-native misspellings are phonetic errors, typos are behavioral errors, OCR mistakes are visual errors, while dyslexic errors could be phonetic or visual. Detecting different classes of errors provides the possibility of refining the knowledge that we have about Web lexical quality. Besides, the fact that dyslexic errors are discriminated from the rest, makes this study valuable to accessible practices for dyslexic Internet users, which is a relatively large group estimated in 10-17% of the USA population [7].

We selected a sample (W) of 50 target words for English with their corresponding variants with errors (W_E), giving us a total of 1,345 different words. Sample W is bigger than previous related work which used ten words [5, 1]. We show an example in the appendix. Based on this, our lexical quality measure [1, 2] is defined as:

$$LQ = mean_{w_i \in W_M} \left(\frac{df_{\text{misspell } w_i}}{df_{\text{correct } w_i}} \right) .$$

where W_M is the subset of ten misspelled words from W_E that at the same time were frequent and had large relative error, computing the relative ratio of the misspells to the correct spellings averaged over this word sample. In this measure, a lower value of LQ implies a larger lexical quality, zero being perfect quality. Notice that LQ is correlated with the rate of lexical errors but it is not the same because is a ratio against the correct words and takes into account the most frequent misspell for each word. To compute LQ, we estimate df by searching each word in the English pages of a major search engine. Although the lexical quality measured will vary with the set of words W chosen, the relative order of the measure will hardly change as the size of the set grows. Hence, we believe that LQ is a good estimator of the lexical quality of a website.

3. COMPARING WITH WEB POPULARITY

To show the relevance of LQ as an independent variable we computed the Pearson correlation with the following measures of Web popularity for the top 20 sites in English of Alexa.com (all in March 2011): Alexa unique visitors, number of pages, number of in-links, and ComScore Unique Visitors. We see that LQ is mildly correlated with the Alexa ranking (see Table 1) and the size (as expected, more content, more errors). Hence, LQ provides independent infor-

Table 1: Pearson correlation between LQ and several measures of Web popularity.

Measure	Alexa	Pages	Links	ComScore
LQ	**0.4451**	0.4167	0.3966	0.2356
Alexa		**0.7659**	0.6897	**0.6589**
Size			**0.8655**	0.3097
Links				0.1319

mation about Web popularity, which is usually correlated with its quality.

We have assessed LQ in several large Web domains and networks[1] and the major English speaking countries [2]. Although there is a correlation between high LQ and the content of major websites, some domains that should have high LQ, do not have it. Specific LQ results in major and social media websites[2] are given in [8]. We found that many Web 2.0 sites have quite good LQ in spite of their collaborative nature, like Wikipedia or Flickr. Maybe LQ fails to capture some aspects of quality of these sites because they do not contain a lot of structured text. We have done this analysis also in Spanish, obtaining results similar to the case of English [2].

4. UNDERSTANDABILITY

To validate that LQ is related to Web accessibility, we tested its impact in terms of (1) the understandability of the text and (2) its reading speed since readers make longer pauses at points where processing loads are greater [6].

To check the reading speed, we used eye tracking (Tobii T50) where more than 50 participants read two comparable Spanish texts in terms of readability (same length, 77 words, same number of sentences, same genre and topic): one with errors (12 multiple errors) and one without errors. In both cases the participant did not know whether the text had errors or not. The participants were divided in two groups, a group of subjects without reading impairments ($n = 30$) and a dyslexic group ($n = 22$), to see if there were differences when the participant had a disability that affected reading or not. To determine understandability, a comprehension question about the text was asked at the end of it.

The results of our experiment are shown in Table 2. For the non-impaired group (N), we applyied ANOVA tests and found statistical significance between the texts in the duration of the fixations ($p < 0.01$), the number of fixations ($p < 0.01$) and the total duration ($p < 0.02$) when reading the text. Regarding comprehension, the correct text had 50% higher percentage of correct answers than for the text with errors. For the dyslexic group (D), we found statistical significance for the number of the fixations ($p < 0.01$) and the total duration ($p < 0.01$). For the comprehension question, dyslexics got almost the same percentage of correct answers in both cases. This shows that given their reading problems, errors do not make the text more difficult as in most cases they do not see them (we did ask them how many errors they saw in the text and usually the answer was none).

We also found statistical significance among both groups ($p < 0.038$) taking into account the mean fixation time that

[1]For example .com, .edu, .gov, .mil, Wikipedia, Yahoo!, Microsoft, CNN, etc.

[2]Such as Facebook, Flickr, Y! Answers, Twitter, Youtube, Blogger, etc.

Table 2: Experimental results of the eye-tracking user study.

Measure (ave. ± std.dev.)	Correct text	Text with errors
	Group N	
Fixations Duration	0.185 ± 0.032	0.219 ± 0.043
Visit Duration	22.1 ± 8.9	28.4 ± 9.8
Fixations Count	75.8 ± 24.3	94.5 ± 29.6
Correct Answers	90.0%	60.0%
	Group D	
Fixations Duration	0.23 ± 0.13	0.25 ± 0.20
Visit Duration	32.28 ± 14.5	40.4 ± 22.5
Fixations Count	107.5 ± 44.3	121.5 ± 49.7
Correct Answers	77.3%	77.7%

is 0.23 second for dyslexic users while is 0.19 second for non-dyslexic participants. These results show that the lexical quality of a text has an effect on its understandability and its readability, moreover when reading disabilities are involved.

Therefore, although LQ uses a conventionally non-accessibility source and is not an accessibility metric, could be potentially added to Web accessibility mesaures as a proxy measure for Web text understandability.

5. REFERENCES

[1] R. Baeza-Yates and L. Rello. Estimating dyslexia in the Web. In *International Cross Disciplinary Conference on Web Accessibility (W4A 2011)*, pages 1–4, Hyderabad, India, March 2011. ACM Press.

[2] R. Baeza-Yates and L. Rello. On measuring the lexical quality of the web. In *The 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, April 2012.

[3] B. Caldwell, M. Cooper, L. G. Reid, and G. Vanderheiden. Web content accessibility guidelines (WCAG) 2.0. *WWW Consortium (W3C)*, 2008.

[4] M. Cooper, L. G. Reid, G. Vanderheiden, and B. Caldwell. Understanding WCAG 2.0. a guide to understanding and implementing web content accessibility guidelines 2.0., 2010.

[5] I. A. Gelman and A. L. Barletta. A “quick and dirty” website data quality indicator. In *The 2nd ACM workshop on Information credibility on the Web (WICOW '08)*, pages 43–46, 2008.

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

[7] J. E. McCarthy and S. J. Swierenga. What we know about dyslexia and web accessibility: a research review. *Universal Access in the Information Society*, 9:147–152, June 2010.

[8] L. Rello and R. Baeza-Yates. Social media is not that bad! the lexical quality of social media. In *The International AAAI Conference on Weblogs and Social Media (ICWSM).*, June 2012.

Appendix: For the target word *tomorrow* the corresponding errors variants are: **romorrow, *yomorrow, *timorrow, *tpmorrow, *tonorrow, *tomirrow, *tomprrow, *tomoeeow, *tomottow, *tomorriw, *tomorrpw, *tomorroq* and **tomorroe* (typographical errors); *toomorrow* (regular spelling error); **tomorow and *tomorou* (non-native speakers errors); **torromow* (dyslexic error); and **tomorrov, *tamarraw* and **tonorrow* (OCR errors).