On Measuring the Lexical Quality of the Web*

Ricardo Baeza-Yates

Yahoo! Research &
Web Research Group,
Universitat Pompeu Fabra
Barcelona, Spain
rbaeza@acm.org

Luz Rello

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

ABSTRACT

In this paper we propose a measure for estimating the lexical quality of the Web, that is, the representational aspect of the textual web content. Our lexical quality measure is based in a small corpus of spelling errors and we apply it to English and Spanish. We first compute the correlation of our measure with web popularity measures to show that gives independent information and then we apply it to different web segments, including social media. Our results shed a light on the lexical quality of the Web and show that authoritative websites have several orders of magnitude less misspellings than the overall Web. We also present an analysis of the geographical distribution of lexical quality throughout English and Spanish speaking countries as well as how this measure changes in about one year.

Categories and Subject Descriptors

I.7.5 [**Document and Text Processing**]: Document Capture—*Document Analysis*

Keywords

Web quality, lexical quality, social media, English and Spanish domains, geographical distribution

1. INTRODUCTION

Measuring the quality of a web page is one of the key problems for web search engines, as ranking pages is one of the major differentiators in this area. Usually, intrinsic quality depends on semantic quality, which is very hard to measure. Hence, many proxies for the real quality were proposed first in information retrieval based on the use of words and later in the Web, using link analysis and click-through data [2]. Here, we address the lexical quality of a web page.

*This work has been partially funded by the HIPERGRAPH project (TIN2009-14560-C03-01) from the Spanish Economy and Competitiveness Ministry.

WebQuality '12, April 16, 2012, Lyon, France

Lexical quality broadly refers to the degree of excellence of words in a text. This word quality (spelling errors, typos, etc.) impacts the reader's understanding [4] and it is also related to textual accessibility [10]. Previous work had shown that there is a strong correlation between spelling errors and web data content quality [5]. However, to the best of our knowledge, results about the distribution of lexical quality considering the entire Web have not yet been presented.

Our proposed measure was inspired in our previous work to estimate the different types of errors in the Web [1]. Hence, we propose to estimate lexical quality focusing on a small set of misspelled words, carefully chosen. Then, we can use a web search engine to compute this lexical quality measure in any web segment. We have done this for web pages in English and Spanish, applying this measure to 25 major Internet domains and social media websites.[1] Then, we study the geographic distribution of lexical quality for the ten major English and Spanish speaking countries. For English we also study of this measure changed in a period of almost a year.

The rest of this paper is organized as follows. Section 2 describes related work. Section 3 presents our measure to assess lexical quality. The results of our estimation are presented in Section 4, considering different Internet domains, the geographical distribution and their presence in social media. In Section 5, conclusions are drawn and plans for future work are considered.

2. RELATED WORK

Web quality can be related to its contents (highly current, accuracy, source reputation, objectivity, etc.) or to its representation (spelling errors, various typos, sentences with low readability, grammatical errors, etc.). Most of efforts are focused on assessing content quality, e.g. spam detection or source credibility. Ringlstetter *et al.* [11] propose filtering methods to retrieve cleaner corpora from the Web after investigating the distribution of orthographic errors of various types of web pages while Piskorski *et al.* [9] explore certain linguistic features for detecting spam.

Our approach is mainly inspired by the work of Gelman and Barletta [5] that applies 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

[1]A group of Internet-based applications that build on the ideological and technological foundations of the Web 2.0, which allows the creation and exchange of user-generated content [7].

engine for this set, showing that web content quality and lexical quality are related. We followed that idea to estimate the different type of spelling errors in the Web, in particular coming from people with dyslexia [1]. In that work we presented an extended classification of errors which distinguishes between regular spelling errors, typographical errors, errors made by non-native speakers of English, dyslexic errors and optical character recognition (OCR) errors. Then we used 50 words in English to estimate the prevalence of each kind of error with a set of more than 1,500 different spelling variations. Detecting different classes of errors provides the possibility of refining the knowledge we have about lexical quality of the Web and it can be useful for estimating some characteristics of web users. That work inspired our lexical quality measure because we found that there are misspellings that are much more frequent than others.

3. A MEASURE FOR LEXICAL QUALITY

By lexical quality we understand its classic definition taken from the theory of reading acquisition. According to Perfetti [8], a lexical representation has high quality to the extent that (1) it has a fully specified orthographic representation (a spelling) and (2) it has redundant phonological representations (one from spoken language and one recoverable from orthographies-to-phonological mapping).

In this context, a measure of lexical quality for the Web should be independent of the size of the text or the number of pages in a website, to be able to compare this measure across websites or different web segments. One alternative could be to compute the rate of spelling errors, that is, the number of misspellings divided by the total number of words. However, that is hard to compute in the context of the Web. A solution is to use a sample of words and use the rate of spelling errors of those individual words to maintain independence of the text size. However, it is not trivial to find in the Web which are all possible misspells of a word for two reasons: (1) the number of possible variations increases exponentially with the number of errors and (2) there might be more than one correct word at the same distance of errors for a given misspelled word. A possible solution is then to find words that are frequent and that also have a frequent misspell, using that occurrence ratio as a proxy of the exact misspell rate. As the frequency of the most frequent misspell is much less than the correct version,[2] we can approximate the word rate of spelling errors just dividing by the number of correct occurrences instead of the total number of all possible misspells of the word (which as we said earlier is harder to determine).

Hence, we define our measure of lexical quality as the average rate of the most common misspell for a set of words. That is, given a set of words W, we compute the relative ratio of the most common misspell to the correct spelling averaged over this word sample scaled by 100 to obtain values around 1. That is,

$$LQ = 100 \cdot \text{mean}_{w_i \in W} \left(\frac{df_{\text{ misspelled } w_i}}{df_{\text{ correct } w_i}} \right) ,$$

where df is the document frequency of each word as we will measure lexical quality across web pages and not words. Using the term frequency would be better, but that would imply that computing LQ is much harder as then a standard search engine cannot be used.

[2]In fact, the distribution many times follows a power law, as the famous Britney Spears example: http://www.google.com/jobs/britney.html.

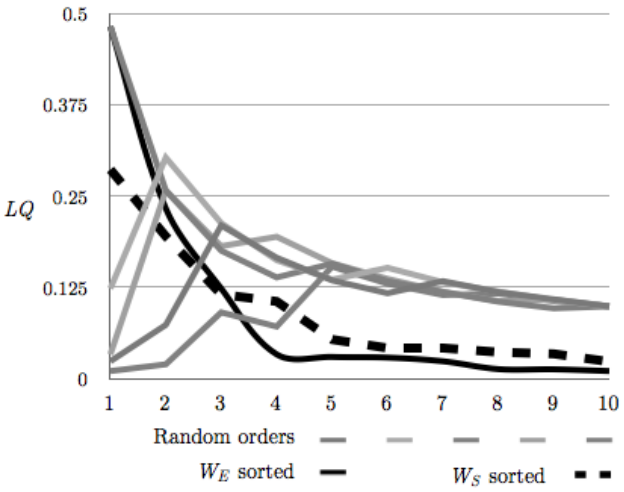

Figure 1: LQ for five variants of W_E in random order for the Web in English, and the sorted individual misspell ratios for English and Spanish.

With this definition, a lower value of LQ (Lexical Quality) implies a larger lexical quality, being 0 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 count of correct words and just takes in account the most frequent misspell for each word.

For W we need to find words that have the following conditions: (1) they are frequent, (2) they have a misspelling with high ratio, and (3) they are non ambiguous, that is, the word or the misspelled word cannot represent other words with the same spelling (e.g. a proper name, acronym or a foreign word). Based on our study of errors in [1] we selected two sets, W_E and W_S, each of ten words, for English and Spanish respectively, that fulfill the conditions stated before (both sets are given in the Appendix).

In Figure 1 we show the convergence of LQ using five different random orders of W_E. We can see that already with half of the words we get values similar to ten words. This shows that the relative order of the measure improves as the size of W_E grows. We also give the sorted order of the individual ratios for W_E and W_S where we can see that the maximum and the minimum misspelling ratios differ by a factor of 50, being the maximum in English for the pair {*becuase, because}. Both curves are quite similar and although LQ is not comparable across languages, this means that in our case the results will be of the same order of magnitude.

The simplest way to measure LQ is by using a web search engine and its frequencies results to compute it. The counts are never exact, but all of them are given by the same estimation algorithm, so the results are still valid to compare different web segments. Using Google we obtained that LQ for the English Web in March of 2011 was 0.047. We also computed LQ for the English Web using (a) *exact* counts for the Yahoo! index in March of 2011 and (b) using the sampling technique of [3] to obtain a set of 28,000 web pages (68% in English). The results for LQ were 0.099 and 0.037,

Pages in English	2011		2012	
	Google	Bing	Google	Bing
.org	0.038	0.075	0.066	0.044
.net	0.08	0.096	0.157	0.099
.com	0.051	0.081	0.149	0.219
Web	0.047	0.099	0.107	0.220

Table 1: LQ for English, two different search engines, two different years, and three major Internet domains.

respectively, which are of the same order of magnitude with respect to 0.047.

Although the lexical quality measured will vary with the search engine, the relative order of the measure among different web segments should not change much. To assess this we present in Table 1 the results for two major search engines across two years for the overall Web and three major Internet domains and English pages. Although the correlation between search engines is not as high as we expected (around 0.5), this is explained with the fact that the correlation among years for the same search engine is not much higher, which may be explained by the intense dynamics of web content.

We can notice that LQ increased almost in all cases from 2011 to 2012 (we include more values in the next section that shows the same). That is, the lexical quality is getting worse. There are a few factors that can explain this trend. First, the expansion of the Web 2.0, which has lower quality. In fact, correct spelling does not seem to be a goal since there are deliberate misspells. Second, most new users are young and they usually do not care much about spelling.

To show the value of LQ as an independent measure, we computed the Pearson correlation for the following measures in the top 13 common websites (see Appendix) of ComScore unique visitors in USA (December 2011) and the Alexa.com reach (February 2012): LQ, Alexa reach, number of pages in websites (as given by Google), number of in-links (as given by Alexa), and ComScore unique visitors. The results are given in Table 2, where we can observe that LQ is partially correlated to all these measures, but at the same time gives additional information. This shows that more content implies a higher misspelling rate and that web traffic does not imply better lexical quality. Therefore, we believe that LQ is a good estimator of the lexical quality of a website.

Measure	Alexa	Pages	Links	ComScore
LQ	0.8029	0.7750	0.6780	0.7785
Alexa		0.8972	0.7937	0.7904
Pages			0.8496	0.6322
Links				**0.4371**

Table 2: Pearson correlation for the top English websites in early 2012.

4. THE LEXICAL QUALITY OF THE WEB

In this section we assess LQ in several large Internet domains including social media sites in English and Spanish. We also measure LQ in the largest English and Spanish speaking countries and we study the evolution in time for English websites.

Pages in English	Range		LQ	
	2011	2012	2011	2012
UK Times	0 –0	0 –0.026	0	0.003
NY Times	0.001–0.117	0.00*–0.054	0.032	0.009
USA Gov.	0.00*–0.286	0.00*–0.379	0.032	0.023
UK Gov.	0.00*–0.033	0.00*–0.048	0.010	0.010
.edu	0.001–0.072	0.001–0.379	0.011	0.064
ac.uk.	0.001–0.026	0.00*–0.526	0.011	0.096
Wikipedia	0.002–0.041	0.00*–0.183	0.018	0.038
ODP	0 –0.277	0 –0.046	0.044	0.008
.mil	0 –0.352	0 –0.096	**0.064**	0.016
.mod.uk	0 –1.231	0 –2.637	**0.163**	**0.279**
Yahoo!	0.002–0.453	0.001–0.419	**0.075**	0.077
Microsoft	0.011–0.524	0.001–0.695	**0.115**	**0.162**
CNN	0.015–0.729	0.00*–4.792	**0.126**	**0.595**
.org	0.002–0.103	0.012–2.906	0.038	**0.484**
.com	0.003–0.139	0.055–5.508	**0.051**	**1.002**
.net	0.004–0.233	0.024–5.807	**0.080**	**1.065**
Web	0.010–0.482	0.010–0.451	**0.047**	**0.107**

Table 3: Range and LQ for a sample of frequent misspellings in English in several Internet domains.

4.1 Major Internet Domains

To assess the correlation of lexical quality with respect to different Internet domains, we apply first our measure to 16 large web segments: the three largest domains (.com,.net,.org), two from the Web 2.0 (Wikipedia and the Open Directory Project, ODP), two major newspapers, a main media player in Internet (CNN), two governmental domains, two academic domains, two military domains and two large Internet companies (Microsoft and Yahoo!).

The results obtained, given in Table 3,[3] where we also include the range of values for the lexical quality of individual words, show several surprises. First, although there is a correlation between high lexical quality and the content of major websites, some domains that should have high lexical quality do not have it. Even worse, the quality is less than the average of the Web, CNN being the worst example. Hence, this means that the correlation is lower than expected. However, maybe the low quality of CNN might be due on the possibility of including user generated content that this website offers.

Second, we find that the .net domain has lower quality than the .com domain, while .org is better than both of them. Third, the edu domain (mainly USA universities) has better quality than UK universities but surprisingly they are not much better than the Web average. On the other hand, UK government has three times better quality than the USA government, while the order is reversed for military domains.

Fourth, Web 2.0 sites have quite good lexical quality in spite of their collaborative nature. On the other hand, the lexical quality of social media shown later, impacts many sites. For example the community section of the NY Times is the main contributor to decrease its lexical quality. A similar effect occurs for almost all large websites like CNN, Microsoft or Yahoo!. Last, for our arbitrary small sample, we did find a perfect website in 2011, lexically speaking.

We also measured the lexical quality for Spanish in a similar way to English. For this we searched in web pages in Spanish for major Internet domains and two major news-

[3]In all tables, unless indicated, the values over the Web average are highlighted and 0.00* represents a number larger than 0 but less than 0.0005.

Pages in Spanish	Range 2012	Average 2012
.edu	0.00*–0.004	0.001
CNN	0 –0.011	0.002
El Pais (Sp)	0.00*–0.052	0.012
El Universal (Mx)	0.004–0.109	0.047
Wikipedia	0.003–0.194	0.039
Yahoo!	0.029–0.254	0.115
Microsoft	0.00*–0.965	0.116
.org	0.009–0.234	0.070
.com	0.055–0.570	**0.222**
.net	0.039–0.790	**0.236**
Web	0.029–0.300	**0.147**

Table 4: Range and *LQ* for a sample of frequent misspellings for the Spanish text in several Internet domains.

papers, one in Spain and another one in Mexico. Some results are different from English, in particular all major domains have better lexical quality as well as CNN and Yahoo!. Other cases are strikingly similar like Microsoft and Wikipedia. On the other hand, although the lexical quality is worse for Spanish, less major web segments are worse than the Web average.

4.2 Social Media

Lexical quality results in major social media websites are shown in Table 5. In Flickr the lexical quality is better than in the Web. An explanation of this could be that texts in Flickr are short (e.g. tags) and our words are long. On the other hand, all the other websites have worse lexical quality than Web average.

We also measured the lexical quality of the pages in Spanish in the social media same sites as shown in Table 6. The order is a bit different, probably due that some of those sites are more popular in English than in Spanish. On the other hand the quality seems to be better, but as we pointed out earlier, our results are not comparable across languages.

Pages in English	Range 2011	Range 2012	*LQ* 2011	*LQ* 2012
Flickr	0.001–0.358	0.00*–0.219	0.073	**0.045**
Y! Answers	0.020–4.680	0.005–0.744	0.707	0.149
Twitter	0.002–0.439	0.00*–0.859	0.068	0.154
MySpace	0.002–0.590	0.015–0.613	0.144	0.159
Youtube	0.007–0.578	0.001–1.534	0.137	0.192
Blogger	0.003–1.715	0.001–1.403	0.225	0.258
Facebook	0.040–1.551	0.004–3.155	0.309	0.479

Table 5: Range and *LQ* for a sample of frequent misspellings in several social media sites in English. In this case the values *below* the Web average are highlighted.

4.3 Geographical Distribution

Now we study the difference in lexical quality across countries. We have taken into account the countries which have the highest populations of native English speakers. These are, in descending order: United States (215 M), United Kingdom (58.1 M), Canada (17.7 M), Australia (15.6 M), Nigeria (4 M), Ireland (3.8 M), South Africa (3.7 M) and New Zealand (3.6 M) [12].

Pages in Spanish	Range 2012	Average 2012
Youtube	0.004–0.080	0.022
Blogger	0.004–0.162	0.038
Flickr	0.009–0.208	0.059
MySpace	0.011–0.307	0.092
Twitter	0.015–0.944	**0.161**
Y! Answers	0.038–0.496	**0.217**
Facebook	0.030–2.358	**0.375**

Table 6: Range and *LQ* for a sample of frequent misspellings in several social media sites in Spanish in 2012.

There are more non-native speakers of English than English native speakers. It is estimated that non-native speakers now outnumber native speakers by a ratio of 3 to 1. This estimation depends on how literacy or mastery of a language is defined and measured, so we have also added to our group of countries, India (86.1 M) and Philippines (44 M), where English as a second language is widespread [12]. In India and Philippines only 0.2 and 3.4 millions of speakers have English as a first language, respectively.

Surprisingly, looking at the results in Table 7, web pages from USA, Nigeria and India have the highest lexical quality. This can be explained by the high education level of users, as in India and Nigeria only 6.9% and 28.9% of their respective populations have Internet access [6]. In addition, websites written in English in these countries tend to be official websites since English is an official language used in education, government and business, but is not the most common language. In the USA, the domain .us is less frequent than .com or .net, but USA has the highest number

Domain	Range 2011	Range 2012	*LQ* 2011	*LQ* 2012
.USA (.us)	0.00*–0.113	0.001–0.194	0.028	0.058
.com.ng	0.00*–0.705	0.00*–0.405	0.022	0.090
Nigeria	0.00*–0.100	0.00*–0.399	0.023	0.072
.co.in	0.003–0.230	0.001–0.917	**0.075**	**0.168**
.net.in	0 – 4.802	0 – 0.553	**0.636**	**0.215**
India	0.003–0.077	0.001–0.275	0.022	0.081
.co.za	0.002–0.154	0.001–0.319	0.040	0.060
South Africa	0.003–0.198	0.004–0.324	0.046	0.098
Ireland	0.003–0.495	0.004–0.657	**0.088**	**0.116**
.com.ph	0.005–0.150	0.001–0.206	0.038	0.069
Philipines	0.001–0.147	0.001–0.333	0.047	**0.118**
Canada	0.002–0.208	0.005–0.635	**0.060**	**0.123**
.com.au	0.007–1.264	0.005–0.548	**0.186**	**0.127**
Australia	0.006–0.190	0.005–0.420	**0.075**	**0.134**
.co.uk	0.007–0.282	0.007–0.340	**0.058**	0.090
.ac.uk	0.001–0.025	0.001–0.744	0.012	**0.112**
.gov.uk	0.001–0.124	0.00*–0.180	0.010	0.026
UK	0.001–0.191	0.007–0.609	**0.075**	**0.135**
.co.nz	0.004–0.356	0.006–0.681	**0.102**	**0.244**
New Zealand	0.005–0.193	0.006–0.596	**0.069**	**0.202**

Table 7: Range and *LQ* for a sample of frequent misspellings in several countries' domains.

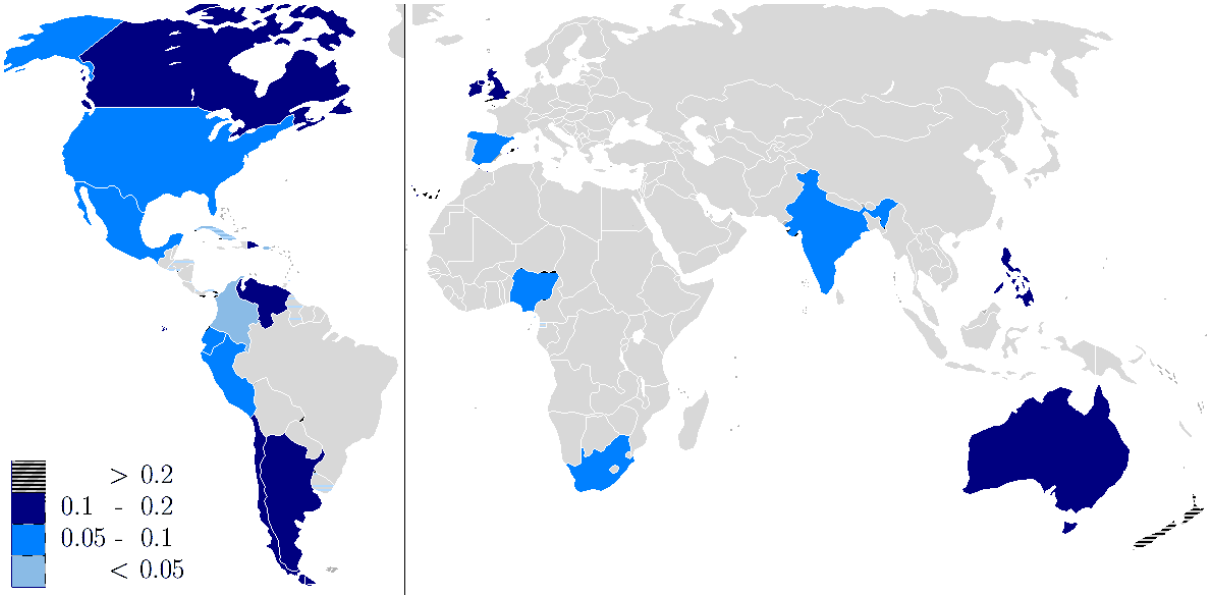

Figure 2: Geographical distribution of LQ in 2012 for the major English and Spanish speaking countries.

of Internet users [6].

South Africa and Philippines have also a considerably high lexical quality considering the co-existing varieties or dialects of English in those countries. After finding that some of the UK domains where the ones with the highest lexical quality it is curious how the average of the overall lexical quality drops due to other subdomains. It is worth noting that .co.uk in 2011 was five times worse than {gov,ac}.uk which changed in 2012 where .ac.uk was four times better.

We observe a common trend between lower lexical quality and higher Internet access rate in Canada, Australia, United Kingdom, and New Zealand. An explanation of this could be the impact of social media in countries where Internet penetration is higher [6], since on average social media presents lower lexical quality than the Web, as shown in the previous subsection.

We also studied LQ in the ten countries with largest population of native Spanish as official language (that is, USA is not included): Mexico (104.1 M), Colombia (45.9 M), Spain (42.0 M), Argentina (36.3 M), Venezuela (28.4 M), Peru (25.3 M), Chile (17.1 M), Ecuador (11.9 M), Cuba (11.2 M) and Dominican Republic (10.0 M) [12]. The results are given in Table 8, where we can notice that the lexical quality in all countries is better than the Web average in Spanish.

Pages in Spanish	Range 2012	Average 2012
Cuba	0.00*–0.029	0.005
Colombia	0.003–0.250	0.042
Ecuador	0.006–0.304	0.053
Mexico	0.005–0.254	0.055
Spain	0.010–0.197	0.057
Peru	0.006–0.295	0.068
Chile	0.005–0.416	0.100
Dominican Rep.	0.004–0.684	0.113
Venezuela	0.005–0.339	0.119
Argentina	0.005–0.394	0.119

Table 8: Range and LQ for the ten largest Spanish speaking countries in 2012.

5. CONCLUDING REMARKS

Our results show that the correlation between lexical quality and domain quality is high, and that the geographical distribution of lexical quality show the impact of business web pages and number of users among English speaking countries. We speculate that the low lexical quality in countries where social media has a greater impact is related to a greater amount of user generated content in their websites. On the other hand, it is possible that as we use a small number of misspells, we are not able to capture the real lexical quality in those websites, for instance, tweets containing intentionally misspelled words. Hence, a tailored set of words might be needed for some social media sites.

Lexical quality is a useful measure, as also can be used as a feature to assess web content quality or it could help to estimate the understandability of a text in accessibility practices [10].

Finally, our results show that it is important to analyze periodically the impact of the lexical quality of the Web. Hence, future work will include to validate further our results regarding our lexical quality measure, as well as improving the measure itself.

Acknowledgements

We thank Berkant Barla Cambazoglu for his help getting some of the document frequencies of the words considered in this study.

6. 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 B. Ribeiro-Neto. *Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition)*. Addison Wesley, Harlow, UK, second edition, 2011.

[3] Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine's index. In *Proc. WWW*, pages 367–376. ACM Press, 2006.

[4] S. F. Ehrlich and K. Rayner. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641 – 655, 1981.

[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] Internet World Stats. Usage and population statistics, April 2011. `http://www.internetworldstats.com`.

[7] A. M. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53:59–68, January-February 2010.

[8] C. Perfetti and L. Hart. *Precursors of functional literacy*, chapter The lexical quality hypothesis, pages 189–213. Amsterdam/Philadelphia: John Benjamins, 2002.

[9] J. Piskorski, M. Sydow, and D. Weiss. Exploring linguistic features for web spam detection: a preliminary study. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, AIRWeb '08, pages 25–28, New York, NY, USA, 2008. ACM.

[10] L. Rello and R. Baeza-Yates. Lexical quality as a proxy for web text understandability. In *The 21st International World Wide Web Conference (WWW 2012)*, April 2012.

[11] C. Ringlstetter, K. U. Schulz, and S. Mihov. Orthographic errors in web pages: Towards cleaner web corpora. *Computational Linguistics*, 2006.

[12] Wikipedia. Wikipedia, the free encyclopedia, April 2011. `http://www.wikipedia.org`.

APPENDIX

The sample W_E of ten frequently misspelled words in English is:

> **albun (album), *alwasy (always), *arround (around), *becuase (because), *enoguh (enough), *everyhting (everything), *haveing (having), *problen (problem), *remenber (remember)*, and **workig (working)*.

The sample W_S of ten frequent misspelled Spanish words is:

> **entocnes (entonces)*, **haceindo (haciendo)*, **honbre (hombre)*, **momemto (momento)*, **pefecto (perfecto)*, **porqeu (porque)*, **peuden (pueden)*, **siemrpe (siempre)*, **tenog (tengo)* and **vamso (vamos)*.

The top 13 websites used in the correlation study were:

> *amazon.com, aol.com, craigslist.org, ebay.com, espn.go.com, facebook.com, google.com, linkedin.com, msn.com, netflix.com, twitter.com, wikipedia.org,* and *yahoo.com.*