Estimating Dyslexia in the Web

Ricardo Baeza-Yates
Yahoo! Research &
Web Research Group, UPF
Barcelona, Spain

Luz Rello
Web Research and NLP Groups
Univ. Pompeu Fabra
Barcelona, Spain

ABSTRACT

In this study we present an estimation of texts containing English dyslexic errors in the Web. A classification of lexical errors is proposed and unique dyslexic errors are distinguished from other kind of errors due to spelling and grammatical errors, typos, OCR errors and errors produced when English is used as a foreign language. A representative sample of each kind of error is used to calculate a lower bound for the prevalence of dyslexia in the English Web. Although dyslexia has been studied in the context of Web accessibility, to the best of our knowledge, an estimation of Web texts containing dyslexic errors was unknown. Our results are useful to tackle future work in Web accessibility among dyslexic users focusing not only in the interface but also in the text content.

1. INTRODUCTION

Dyslexia is a neurologically-based disorder which interferes with the acquisition and processing of language. Varying in degrees of severity, it manifests itself with difficulties in receptive and expressive language, including phonological processing, in reading, writing, spelling and handwriting and sometimes in arithmetic [8]. Although in some literature, dyslexia is only referred to reading and dysgraphia to writing, this study takes into account a broader definition of dyslexia and its manifestations in writing. Following the Boder's diagnostic screening tool for developmental dyslexia (The Boder' Test of Reading-Spelling Patterns [4]), our study takes into consideration the dysphonetic dyslexia which is the largest of the three subtypes of dyslexia that the author presents. Dysphonetic dyslexia is viewed as a disability in associating symbols with sounds. The misspellings typical of this disorder are due to phonetic inaccuracy.

There is a universal neuro-cognitive basis for dyslexia, nevertheless, it manifestations are culture-specific due to different orthographies. Therefore, spelling errors vary considerably between languages [1]. This work focuses on English and since it is a language with deep orthography, the mapping between letters, speech sounds, and whole-word sounds is often highly ambiguous and therefore dyslexics examples are more widespread than in other languages with transparent or shallow orthography [18]. Researchers estimate that 10-17 % of the population in the U.S.A. has dyslexia and only 30 % of dyslexics have trouble with reversing letters and numbers [16]. On the other hand, the level of dyslexia in other regions such as Europe or China is lower.

Regarding this important and relatively large group of users, various studies take into account dyslexia from the Web accessibility point of view. They focus mainly on designing guidelines for authoring dyslexic-accessible interfaces [15] and on producing special text formats for dyslexic users [11]. However, there is a common agreement in these studies that the application of dyslexic-accessible practices benefits also the readability for non-dyslexic users [17, 15] as well as other users with disabilities such as low vision [6, 11, 5].

There is a considerable body of knowledge on dyslexia and its relationship with computers. Among others, dyslexia has been considered for e-learning [23], for the creation of tools to diagnose [24] and correct dyslexic errors [19]. Many softwares for assisting dyslexic users has already been developed [14, 22, 12]. However, to the extent of our knowledge, this is the first attempt to estimate the amount of texts containing English dyslexic errors in the Web.

Detecting the presence of dyslexic texts in the Web helps us to know the real impact of dyslexia in the Web as well as to value dyslexic-accessible practices. Moreover, spelling error rates has proven a useful index for website content quality [13].

This paper is organized as follows. Section 2 presents our error classification. Then, Section 3 explains the methodology used in our study, including the two steps used for the estimation of errors in the Web and the word samples used. The results of our estimation of the dyslexic Web is presented in Section 4. In Section 5 conclusions are drawn and plans for future work are considered, as well as a discussion about the amount of different types of dyslexic errors from previous related work.

2. ERROR CLASSIFICATION

In order to detect lexical errors produced by dyslexic individuals, it is required to distinguish dyslexic errors from the rest of lexical errors. To this purpose, we establish five classes of errors, taking into consideration the user disability, the user mother tongue and the source of the text:

1. Dyslexic errors: Among the different kinds of errors commonly made made by dyslexics (i.e. unfinished

W4A 2011, March 28–29, 2011, Hyderabad, India.

words or letters, omitted words, inconsistent spaces between words and letters [21]), we only consider multiple repetitions, additions, transpositions, omissions, substitutions, and reversals in letters inside words. For instance, **reiecve* instead of *receive*[1].

2. Regular spelling errors produced by non-impaired native English individuals, such as the transposition error **recieve*.

3. Regular typos caused by the adjacency of letters in the keyboard, i.e. **teceive*.

4. Optical character recognition (OCR) errors, due to letters of similar shape, such as **ieceive*.

5. Errors made by non-native speakers who use English as a foreign language. For example, **receibe* is a typical error made by Spanish learners of English, since the graphemes 'b' and 'v' are pronounced as /b/, and the phoneme /v/ does not exist in the standard Spanish phonemic system.

Notice that depending on the word, one error might belong to more than one class. Hence, for our purposes we will need to find words where these cases are not ambiguous.

3. METHODOLOGY

3.1 Estimating the Presence of Errors

Assume that the fraction of Web pages with lexical errors is f and that the relative fraction of dyslexic errors among all lexical errors is d. Then, the fraction of Web pages with dyslexia is $f \times d$. Sampling the Web is a difficult problem in general [3] and even more in our particular case. Hence, we just do a rough estimation for a lower bound of f and d, and hence we obtain a lower bound for the fraction of dyslexic pages in the Web.

We use the three major search engines (Bing, Google and Yahoo!) to estimate the document frequency of a word. Each of the words in our list is searched only in English web pages to avoid cases of wrong words that may have a meaning in other languages.

We estimate the relative fraction of wrongly written documents, f, by using a sample of frequent words that appear in most documents, usually called stopwords in information retrieval [2]. Then we use the largest relative fraction of misspells for all these words to estimate f, as we cannot assume that all of them appear in different pages.

To estimate d we do the same frequency search with a sample of non-frequent words where we can distinguish the different types of errors without ambiguity. In this case we use a small sample of words as finding words where all possible cases of errors are non ambiguous is not trivial. Here we assume that these strange misspells appear in different Web documents. We cover this case next.

3.2 Dyslexic Words

All the dyslexic spelling errors are taken from samples of text written by adults with diagnosed dyslexia and from literature [20]. Since some of the lexical errors made by dyslexics overlap with other existing words (i.e. **was* and *saw*, **form* and *from*) or with other kind of errors such as misspells and regular typos, i.e. **remeber*, the following criteria for the sample selection is used.

Among the dyslexic errors, we take in account the ones which include the letters that produce more confusion among dyslexic individuals, such as 'b', 'd', 'p', 'm', 'n', 'u' and 'w' together with other similar looking letters. For instance, it is specially frequent to find reversals of similar letters, such as 'b' and 'd' [10].

To avoid the overlap of dyslexic errors and regular typos, simple errors [9] are not taken in account. We consider only words written by dyslexics containing multi-errors [20], that is, the dyslexic word differs from the intended correct word by more than one letter. For example, the dyslexic word **konwlegde* from *knowledge*.

In our set of multi-errors dyslexic words not all the kind of possible errors are taken into account. Errors due to homophone confusion, that is words which have a similar pronunciation [20], are not selected even though 15 % of the dyslexic errors presented homophone confusion in a corpus of dyslexic texts. We also avoid taking into account errors which produce a syntactic anomaly, i.e. words that have no part of speech tags in common, such as the error **from (form)* or inflection errors, i.e., **story (stories)*, which are non-correctable.

Errors which coincide with other existing words in English (real world errors in the literature [20]) are omitted, i.e. **trust* being the intended word *truth* [7]. Errors which give as a result a proper name are also filtered, for instance the typo **wirries* from *worries* is also a proper name.

For each of the dyslexic words selected we check its uniqueness by distinguishing different versions of the word belonging to the other kind of errors, so their source can be predicted without any ambiguity. Due to this, the selected words are usually long (9 letters per word on average).

For example, the word *comparison* has the corresponding types of errors:

1. Dyslexic error: **comaprsion*.

2. Regular spelling errors: **comparision, *conparison* and **coparison*.

3. Regular typos: **vomparison, *xomparison, *cimparison, *cpmparison, *conparison, *co,parison, *comoarison, *com[arison, *comprison, *compsrison, *compaeison, *compatison, *comparuson, *comparoson, *compariaon, *comparidon, *comparisin, *comparispn, *comparisob* and **comparisom*.

4. Optical character recognition (OCR) errors: **compaison* and **comparisom*.

5. Errors made by non-native speakers who use English as a foreign language: **comparition* and **comparizon*.

There are other possibilities, but their frequency is negligible and we can disregard them as we are computing a lower bound.

Regular typos are generated by substituting each of the letter that appears in the word by its adjacent letter (left or right) in the keyboard. Other cases have much smaller frequency (keys above or below). The sample D has a group of ten dyslexic-prone words[2] together with their corresponding

[1] In this work, examples with errors are preceded by an asterisk "*".

[2] *comparison, understanding, knowledge, impossible, tomorrow, worries, explain, interesting, situation* and *confusion*.

variants with errors, giving as a result a total of 260 different words.

4. DYSLEXIA IN THE WEB

Before estimating d we need a short digression. In our study only multi-errors are taken into consideration, which are less frequent than simple errors in dyslexic texts. In previous research [20], the error rate from a corpus of 12,000 words written by dyslexics was calculated. In the study, 17% of the errors are real-word errors while 83% are non-word errors. Over all the errors, 39% of them were multi-errors while 53% were simple errors. The rest of the cases (8%) were due to word boundary errors (i.e. **alot* from *a lot*). Hence, our method underestimates the real value of d and then we can safely use a factor of 3 to correct this fact.

From the sample D, the absolute percentage of dyslexic errors is very low with an average of approximately 0.7%. Then, we can estimate d as 2.01%. Table 1 shows the ranges for all our error classes. We use the real document frequencies of the terms from one of the search engines to validate the results obtained, finding very similar results.

Error Class	Range (%)	Average (%)
Spelling	47.390 – 91.571	63.732
Typo	11.309 – 47.747	28.184
Foreign	1.568 – 10.218	6.615
OCR	0.003 – 3.648	0.799
Dyslexia	0.007 – 3.100	0.669

Table 1: Range of percentages and average for the different error classes.

Following our methodology, we have $f \geq 0.27\%$ obtained with the common word *because*, taking the maximum over all search engines. Then, using the previous value of d, a lower bound for the percentage of dyslexic text documents in the Web is 0.005%. This value is much lower that the corresponding number of dyslexic users (say 10%). On the other hand it is of the same order of magnitude of OCR errors and then, most probably, our lower bound is very conservative.

Although this is a small percentage, for each 20 billion Web pages, there are at least one million pages containing dyslexic errors. If we consider all spelling errors as dyslexic errors, the lower bound would increase to close to 0.2% and for each 10 billion pages, 20 million Web pages would contain dyslexic errors.

5. CONCLUDING REMARKS

Our main conclusions are that:

- The amount of dyslexic texts in the Web is not as large as it could be. This suggests the idea that the widespread use of spell checkers ameliorates dyslexia in the Web.
- Particular words can be used to detect dyslexic texts, and hence dyslexic users. This can be used to improve Web accessibility as well as future spell checkers targeted to dyslexic users.

Since this is the first attempt to estimate text written by dyslexics individuals in the Web, a comparison with previous work is not possible. Moreover, previous research on dyslexia reveals that error frequency is related with word length [20]. Short words such as *there, where, form, etc.* are misspelled much more frequently in dyslexic texts than long words like the ones used in our experiments. Hence, we can do a better estimation by using a larger sample of stopwords as well as long dyslexic words.

As a byproduct we have found that other types of errors are much more frequent in the Web and this can be used to assess the quality of Web text. In addition, the methodology applied in this study to measure different types of errors will be used in future work to develop natural language processing tools to improve Web accessibility for dyslexic users.

6. REFERENCES

[1] J. Alegria. Support for a psycholinguistic approach to reading acquisition and reading difficulties: Twenty years later. *Infancia y Aprendizaje*, 29(1):93–111, 2006.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition)*. Addison Wesley, 2010.

[3] Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine's index. *J. ACM*, 55(5), 2008.

[4] E. Boder and S. Jarrico. *The Boder' Test of Reading-Spelling Patterns: A Diagnostic Test for Subtypes of Reading Disability*. Grune and Stratton, New York, 1982.

[5] P. Brophy and J. Craven. Web accessibility. *Library Trends*, 55(4):950–972, 1964.

[6] B. Caldwell, M. Cooper, L. Guarino Reid, and T. Vanderheiden. *Web Content Accessibility Guidelines (WCAG) 2.0*. MIT, ERCIM, Keio, 2008.

[7] M. Coltheart, K. Patterson, and J. C. Marshall. *Deep dyslexia*. Routledge and Kegan Paul, London, 1980.

[8] Committee of Members Orton Dyslexia Society. Definition of Dyslexia, 1994. See Thomson, P. and Gilchrist, P., *Dyslexia: a multidisciplinary approach*, 1996, p. 5.

[9] F. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the A.C.M.*, 7:171–176, 1964.

[10] G. Deloche, E. Andreewsky, and M. Desi. Surface dyslexia: A case report and some theoretical implications to reading models. *Brain and Language*, 15(1):12–31, 1982.

[11] L. Evett and D. Brown. Text formats and web design for visually impaired and dyslexic readers-clear text for all. *Interacting with Computers*, 17:453–472, July 2005.

[12] Y. Feng and M. Lapata. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99, 2010.

[13] I. A. Gelman and A. L. Barletta. A "quick and dirty" website data quality indicator. In *WICOW '08 Proceeding of the 2nd ACM workshop on Information credibility on the web*, pages 43–46, 2008.

[14] Y. Marchand and R. I. Damper. A multistrategy approach to improving pronunciation by analogy. *Computational Linguistics*, 26:195–219, June 2000.

[15] 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.

[16] H. Meng, S. Smith, K. Hager, M. Held, J. Liu, R. Olson, B. Pennington, J. DeFries, J. Gelernter, T. O'Reilly-Pol, S. Somlo, P. Skudlarski, S. Shaywitz, B. Shaywitz, K. Marchione, Y. Wang, P. Murugan, J. LoTurco, P. Grier, and J. Gruen. DCDC2 is associated with reading disability and modulates neuronal development in the brain. In *Proceedings of the National Academy of Sciences*, volume 102, pages 17053–17058, November 2005.

[17] M. G. Paciello. *Web accessibility for people with disabilities*. CMP Books, Lawrence, Kansas, 2000.

[18] E. Paulesu, J.-F. Démonet, F. Fazio, E. McCrory, V. Chanoine, N. Brunswick, S. F. Cappa, G. Cossu, M. Habib, C. D. Frith, and U. Frith. Dyslexia: Cultural diversity and biological unity. *Science*, 291(5511):2165–2167, November 2001.

[19] J. Pedler. The detection and correction of real-word spelling errors in dyslexic text. In *Proceedings of the 4th Annual CLUK Colloquium*, 2001.

[20] J. Pedler. *Computer Correction of Real-word Spelling Errors in Dyslexic Text*. PhD thesis, Birkbeck, London University, 2007.

[21] F. R. Vellutino. *Dyslexia. Theory and Research*. The MIT Press, Cambridge, MA, 1979.

[22] S. Williams and E. Reiter. Generating basic skills reports for low-skilled readers. *Natural Language Engineering*, 14:495–535, October 2008.

[23] B. P. Woodfine, M. B. Nunes, and D. J. Wright. Text-based synchronous e-learning and dyslexia: Not necessarily the perfect match! *Comput. Educ.*, 50:703–717, April 2008.

[24] T.-K. Wu, S.-C. Huang, and Y.-R. Meng. Evaluation of ANN and SVM classifiers as predictors to the diagnosis of students with learning disabilities. *Expert Systems with Applications: An International Journal*, 34(3):1846–1856, 2008.