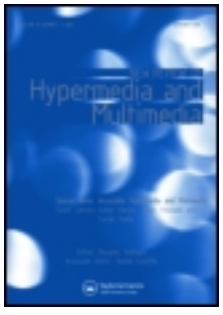

New Review of Hypermedia and Multimedia

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tham20

The presence of English and Spanish dyslexia in the Web

Luz Rello [a b] & Ricardo Baeza-Yates [a c]

[a] Web Research Group, Universitat Pompeu Fabra, Barcelona, Spain

[b] Natural Language Processing Group, Centre for Autonomous Systems and Neuro-Robotics, Universitat Pompeu Fabra, Barcelona, Spain

[c] Yahoo! Research, Barcelona, Spain

Version of record first published: 21 Sep 2012.

To cite this article: Luz Rello & Ricardo Baeza-Yates (): The presence of English and Spanish dyslexia in the Web, New Review of Hypermedia and Multimedia, DOI:10.1080/13614568.2012.721401

To link to this article: http://dx.doi.org/10.1080/13614568.2012.721401

New Review of Hypermedia and Multimedia,
2012, 1–28, iFirst article

Taylor & Francis
Taylor & Francis Group

The presence of English and Spanish dyslexia in the Web

LUZ RELLO†‡* and RICARDO BAEZA-YATES†§
†Web Research Group, Universitat Pompeu Fabra, Barcelona, Spain
‡Natural Language Processing Group, Centre for Autonomous Systems and Neuro-Robotics, Universitat Pompeu Fabra, Barcelona, Spain
§Yahoo! Research, Barcelona, Spain

(Received 20 November 2011; final version received 13 August 2012)

In this study we present a lower bound of the prevalence of dyslexia in the Web for English and Spanish. On the basis of analysis of corpora written by dyslexic people, we propose a classification of the different kinds of dyslexic errors. A representative data set of dyslexic words is used to calculate this lower bound in web pages containing English and Spanish dyslexic errors. We also present an analysis of dyslexic errors in major Internet domains, social media sites, and throughout English- and Spanish-speaking countries. To show the independence of our estimations from the presence of other kinds of errors, we compare them with the overall lexical quality of the Web and with the error rate of noncorrected corpora. The presence of dyslexic errors in the Web motivates work in web accessibility for dyslexic users.

Keywords: Dyslexia; Spelling errors; Error rate; Web mining; Lexical quality

1. Introduction

Worldwide, around 15–20% of the population has a language-based learning disability; where 70–80% of it is likely dyslexic (International Dyslexia Association 2011b). Dyslexia is a specific learning disability with a neurological origin but, despite its universal basis, its prevalence estimates vary (Vellutino *et al.* 2004). This variation has mostly been ascribed to variations in orthographic depth among languages (Brunswick 2010).

Regarding this substantial group of users, various studies take into account dyslexia from the Web accessibility point of view. They mainly focus on tools (Pedler 2001a, Gregor *et al.* 2003, Wu *et al.* 2008) and guidelines for dyslexic-accessible practices (Gregor and Newell 2000, McCarthy and Swierenga 2010, Rello *et al.* 2012c). There is a common agreement in these studies that the application of dyslexic-accessible practices also benefits the readability for nondyslexic users (Kurniawan and Conroy 2006, Dixon 2007) as well as other users with disabilities such as low vision (Evett and Brown 2005).

*Corresponding author. Email: luzrello@acm.org,

New Review of Hypermedia and Multimedia
ISSN 1361-4568 print/ISSN 1740-7842 online
http://www.tandfonline.com
http://dx.doi.org/10.1080/13614568.2012.721401

Detecting the presence of dyslexic texts helps us to know the real impact of dyslexia in the Web as well as to motivate dyslexic-accessible practices. We were the first to estimate a lower bound of the fraction of pages containing some kind of English dyslexic errors in the Web (Baeza-Yates and Rello 2011).

The goal of this paper is to present a complete and improved lower bound of the prevalence of dyslexic errors in the Web for English and Spanish. In this paper, we have extended our work with the following new contributions:

- The addition of Spanish language in our estimation.
- A larger sample of errors from a broader range of sources covering a large percentage of Web pages.
- An improved methodology to estimate the impact of dyslexic errors in the Web.
- The lower bound is extended to the major Internet domains, social media sites and throughout English and Spanish-speaking countries.
- A validation of our results by comparing them with large corpora and the lexical quality of the Web.

This paper is organized as follows. In Section 2 we define dyslexia and explain why we chose to use dyslexic errors as a source of knowledge for English and Spanish. Section 3 explains the methodology used in this study: we define our measure and explain the data sets. In Section 4 we present the results of our estimations and draw conclusions in Section 5.

2. Dyslexia

2.1. Definition

Dyslexia is a specific learning disability with a neurological origin. It is characterized by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities. These difficulties typically result from a deficit in the phonological component of language that is often unexpected in relation to other cognitive abilities. Secondary consequences may include problems in reading comprehension and reduced reading experience that can impede growth of vocabulary and background knowledge (Lyon 1995, Lyon *et al.* 2003, International Dyslexia Association 2011a).

In some literature, dyslexia is referred to as a specific reading disability (Vellutino *et al.* 2004) and dysgraphia as its writing manifestation only (Romani *et al.* 1999).[1] However, our study follows the standard definitions of the *International Statistical Classification of Diseases* (*ICD-10*) (World Health Organization 1993) and the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-IV*) (American Psychiatric Association 2000) where dyslexia is listed as a reading and spelling disorder (*ICD-10*) or a reading disorder and a disorder of written expression (*DSM-IV*).

2.2. Types of dyslexia

Dyslexia is more frequently developmental. However, acquired dyslexia also exists when specific disorders of reading or writing occur after a brain injury. Researchers broadly agree on three different kinds of dyslexia:[2] phonological, surface, and deep dyslexia. However, the delimitation of these three types is not clear, and symptoms of different types of dyslexia overlap (Friedman 1996).

Phonological dyslexia[3] is a reading disorder characterized by impairment in nonword[4] reading ability (Coltheart 1996). Surface dyslexia[5] is characterized by poor reading of low-frequency irregular words, coupled with accurate reading of nonwords. Errors made in reading irregular words tend to be regularizations[6] (Patterson *et al.* 1995). While phonological dyslexics use a visual reading route (read words at once), surface dyslexics use a phonological reading route (grapheme to phoneme conversion rules) (Coltheart 1981). While phonological dyslexics remember an orthographic and phonological vocabulary, surface dyslexics encounter problems when using the grapheme to phoneme conversion rules. People with deep dyslexia present semantic errors related to a lack of semantic representation retrieval, such as substitutions of entire words among others (Coltheart *et al.* 1980).

Although the classification of the types of dyslexia is based on reading models (Coltheart 1981), all dyslexias involve spelling errors (Romani *et al.* 1999, Lyon *et al.* 2003).

2.3. Universality and specificity: dyslexia among languages

Brain structure, brain function, and genetics studies confirm the biological foundations of dyslexia with the exception of acquired dyslexias (Vellutino *et al.* 2004). However, despite its universal neuro-cognitive basis, dyslexia manifestations are variable and culture-specific (Goulandris 2003).

This variability is due to the different language orthographies depending on their grade of consistency and regularity. English has an opaque—or deep—orthography in which the relationships between letters and sounds are inconsistent and many exceptions are permitted. English presents to the beginning reader a significantly greater challenge compared with other languages, such as Spanish. Spanish has a more regular alphabetic system that contains consistent mappings between letters and sounds, that is, a transparent—or shallow—orthography. For instance, in Paulesu *et al.* (2001), Italian dyslexics—shallow orthography—performed better on reading tasks than English and French dyslexics did—deep orthographies. Dyslexia has been called a *hidden* disability due to the difficulty of its diagnosis in languages with shallow orthographies (Vellutino *et al.* 2004).

Our study focuses on English and Spanish. These languages are archetypes of deep and shallow orthographies, respectively. In an orthographic transparency scale for European languages, English appears as the language with the deepest orthography and Spanish as the second most shallow after

Finnish (Seymour *et al.* 2003). We chose English and Spanish because they are the most frequent languages in the world after Mandarin Chinese.[7]

2.4. The prevalence of dyslexia in the population

Depending on the language, the estimations on the prevalence of dyslexia differ. The Interagency Commission on Learning Disabilities (1987) states that 10–17.5% of the population in the USA has dyslexia. The model of Shaywitz *et al.* (1992) predicts that 10.8% of English-speaking children have dyslexia, while in Katusic *et al.* (2001) the rates varied from 5.3% to 11.8% depending on the formula used. Brunswick (2010) estimates 10% for English and 3.5% for Italian. Data on the prevalence of dyslexia in Spanish speakers are much more scarce: Galván Gómez (2010) reports a 7.5% prevalence among school children in Perú; Carrillo *et al.* (2011) found that 11.8% of the school children examined in Murcia (Spain) exhibited difficulties associated with dyslexia, and Jiménez *et al.* (2009) report an 8.6% for a similar population in the Canary Islands (Spain).

However, these estimations of the presence of dyslexia in the population are based on reading tests not on writing misspells. Sterling *et al.* (1998) present evidence that dyslexic adults have a chronic problem in the lexical domain, manifested in poor spelling ability.

2.5. Dyslexic errors

In general terms, spelling errors could be used as a signal of information. For instance, the presence of errors in the textual Web has been used for detecting spam (Piskorski *et al.* 2008), measuring quality (Gelman and Barletta 2008), and understandability of web content (Rello and Baeza-Yates 2012).

As the kind of errors that dyslexic people make are related to the kind of difficulties that they have (Sterling *et al.* 1998), their written errors have been used for various purposes such as (1) the study of dyslexia, (2) its diagnosis, or (3) for accessibility-related purposes.

First, the analyses of writing errors made by dyslexics were used in previous literature to study different aspects of dyslexia (Connelly *et al.* 2006). For instance, the specific types of dyslexic errors highlight different aspects of dyslexia (Treiman 1997) such as a phonological processing deficit (Moats 1996, Lindgrén and Laine 2011). Although the dyslexic error rates (DERs) vary depending on the language writing system (Lindgrén and Laine 2011), errors attributable to phonological impairment, spelling knowledge, and lexical mistakes are more frequent in dyslexics than in nondyslexics (Sterling *et al.* 1998).

Second, as people with dyslexia exhibit higher spelling error rates than nondyslexic people (Coleman *et al.* 2009), there are diagnoses of dyslexia based on the spelling score (Schulte-Körne *et al.* 1996). Also, the spelling error rate is being used as a diagnosing factor in the current official Catalonian protocols (Col legi de Logopedes de Catalunya 2011).

Third, the exploration of corpora of dyslexic errors (Pedler 2007, Rello *et al.* 2012a) was used for various accessibility-related purposes such as the development of tools like spellcheckers (Pedler 2007), text prediction software,[8] games for children with dyslexia (Rello *et al.* 2012b), or word processors which perform text customization taking into account frequent writing errors (Gregor *et al.* 2003).

3. Methodology

In this section we present our measure for sampling dyslexic errors in the Web and a description of the data sets we created, their sources and the criteria used for the selection of the content. On the methodological side, this work was inspired by Gelman and Barletta (2008) who apply the spelling error rate as a metric to indicate the degree of content quality of websites. That work uses a carefully chosen set of 10 frequently misspelled words and hit counts of a search engine for this set.

3.1. Measuring dyslexic errors in the Web

Lexical quality refers to the degree of excellence of words in a text, including all kinds of spelling errors (Perfetti and Hart 2002). In our previous study (Baeza-Yates and Rello 2011), for measuring the impact of the different kind of errors in the Web (including dyslexic errors), we presented an extended classification of errors which distinguishes between regular spelling errors, typographical errors, errors made by nonnative 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. In Baeza-Yates and Rello (2012), we defined a measure of lexical quality and computed it for English and Spanish. In this work we take into consideration what we learned from both approaches to create a measure for estimating the impact of dyslexic errors in the Web.

A measure for estimating the impact of dyslexic errors in 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 dyslexic 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 edit distance for a given misspelled word.

A possible solution is then to find words that are frequent and that also have a frequent dyslexic misspell, using that occurrence ratio as a lower bound of the exact dyslexic misspell rate. As the frequency of the most frequent misspell is

much lower than the correct version,[9] we can approximate the word rate of spelling errors just by dividing the most frequent misspell by the number of correct occurrences instead of using the total number of all possible misspells of the word (which as we said earlier is harder to determine).

To estimate a lower bound of the presence of dyslexia in the Web, we define a measure of the DER as the average rate of the dyslexic misspells. That is, given a set of words W, we compute the relative ratio of the dyslexic misspell to the correct spelling averaged over this word sample scaled by 100 to obtain values that can be interpreted as a percentage. That is,

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

where df is the document frequency[10] of each word as we will measure lexical quality across web pages and not the number of words, because web pages have different number of words. Using the term frequency would be better, but that would imply that computing DER cannot be done using a standard search engine such as we propose here.

As there is no reasonable way to know the overall frequency of words in the Web outside a search engine, for the frequencies we use the hit counts of the Google search engine using the Advanced Search option to search only in English or Spanish websites depending on each case. Then, we compare our results using more than one search engine (Bing and Yahoo!) and validate them with real document frequencies computed from the Yahoo! web search index.[11] The similarity of DER among search engines varied depending on the moment when the queries were submitted and on data set used. While most of the differences were insignificant, we observed the greatest difference using English dyslexic simple errors data set between Google (DER = 0.1023) and Bing (DER = 0.1559) in November 2011.

For W we need to find words that have the following properties: (1) they are frequent, (2) they have a frequent dyslexic misspelling, and (3) they are nonambiguous, that is, the word or the misspelled word cannot represent another word with the same spelling (e.g. a proper name, acronym, or a foreign word).

Using this measure we can compute the impact of different kinds of errors depending on the data sets used for W that fulfill the conditions stated before.

To pursue our goal, we created two new data sets. First, we created W_D (for English, $W_D en$, and Spanish, $W_D sp$) composed of only dyslexic misspellings to compute the impact of different kinds of dyslexic errors. Second, we expanded W_D to W_E (for English, $W_E en$, and Spanish, $W_E sp$) including all types of errors to estimate the impact of dyslexic errors in comparison with other kinds of errors.

3.2. Types of dyslexic errors

The dyslexic errors for our samples were selected from different sources: (1) a corpus of dyslexic errors in English (Pedler 2007), (2) a corpus we composed

of Spanish texts written by children with diagnosed dyslexia (Rello *et al.* 2012a), and (3) literature about dyslexic errors (Sterling *et al.* 1998, Silva Rodríguez and Aragón Borja 2000, Pedler 2001b, 2007).

Following Pedler's classification there are the following kinds of dyslexic errors (Pedler 2007):

(1) Dyslexic errors based on the degree of difference to the intended or target word:
   (a) Simple errors. They differ from the intended word by only a single letter. They can be due to (i) substitution, **reelly* (*really*),[12] (ii) insertion, **situartion* (*situation*), (iii) omission, **approch* (*approach*), and (iv) transposition, **artcile* (*article*). According to Meng *et al.* (2005) only 30% of dyslexics have trouble with reversing letters and numbers. In Damerau (1964), 80% of the misspellings in his corpus (nondyslexic errors) were simple errors.[13]
   (b) Multierrors. They differ in more than one letter from the target word. Some errors, such as **guapoisismo* (*guapísimo, ''*gorgeous*''*), closely resemble the intended word, while others are not so obvious, **lignsuitc* (*linguistics*).
   (c) Word boundary errors. They are mistakes (run-ons and split words) which are special cases of omission and insertion errors. A run-on is the result of omitting a space, such as **alot* (*a lot*). A split word occurs when a space is inserted in the middle of a word, such as **sub marine* (*submarine*).
(2) Dyslexic errors based on their correspondence with existing words:
   (a) Real word errors. Misspellings that result in another valid word. For instance, **witch* being the intended word *which*.
   (b) Nonword errors. Misspellings that do not result in another correct word.
(3) First letter dyslexic errors:
   (a) Only 5% of the dyslexic errors are first letter errors, like **no* (*know*). This is consistent with Yannakoudakis and Fawthrop (1983) whose findings report that the first letter of a misspelling is correct in the majority of cases.

The percentages of dyslexic errors found in the English and Spanish corpora (Rello *et al.* 2012a) are given in Table 1. The error ratio computed as the fraction of errors over the correctly spelled words is higher in the texts written in English (20%) than in Spanish (15%). This might be due to their different orthographies. However, the corpora used are not large enough (3,134 words for English and 1,075 words for Spanish) to draw such conclusion (Rello *et al.* 2012a).

3.3. Selection criteria for words with dyslexic errors

Sample W_D (see Appendix 1) is composed of English and Spanish dyslexic errors extracted from texts produced by people with dyslexia.

Table 1. Error distribution in English and Spanish corpora of dyslexic errors.

	English		Spanish	
Category	Number	%	Number	%
Simple errors	307	53	96	67
Multierrors	227	39	33	23
Word boundary errors	47	8	15	10
Real word errors	100	17	30	21
Nonword errors	477	83	114	79
First letter errors	30	5	16	11
Total	577	100	144	100

First, we extracted all the misspellings from the sources described in the previous section. Second, we selected the errors which are related to a phonological and orthographical processing deficit. For instance, errors due to the similarity of sounds, **vidreo* (*vidrio*), "glass", or the graphemes, **pabre* (*padre*), "father."

Sample W_D includes nonword dyslexic errors from both simple—**childern* (*children*)—and multierrors—**felicdidad* (*felicidad,* "happiness"). Sample W_D includes all kinds of simple errors. However, each error type has a different frequency. For instance, simple omissions are the most frequent kind (Bustamante and Díaz 2006).

We do not take into account first letter errors because in Pedler (2007) a quarter of them are capitalization errors, another quarter are real word errors, and overall they present low frequency. We also discard word boundary errors because they sometimes involve more than one lexical unit.

To reduce the overlap of the errors produced by people with dyslexia in our data set with other kinds of errors, we selected them according to this principle: the errors related to the target word need to be unique and not ambiguous. For example, errors which coincide with other existing words in other languages are omitted, that is **couver* (*cover*). Similarly, errors which give as a result a proper name are also filtered **klene* (*clean*) (Bourassa and Treiman 2003).

To avoid a significant overlap of dyslexic simple errors with other kind of errors such as regular typos, we define additional criteria. First, we consider the cases which include letters with similar pronunciation that produce more confusion among dyslexic individuals than nondyslexic individuals, such as $<m|n>$, $<m|p>$, $<b|p>$, or $<b|m>$. Second, during the selection process, we pay special attention to examples with similar looking or symmetric letters, such as $<d|b>$, $<p|q>$, and $<d|p>$, among others. These criteria are consistent with literature. For instance, it is specially frequent to find substitutions of orthographically similar letters, such as $<b>$ and $<d>$ (Deloche *et al.* 1982) as well as substitutions in letters with similar sounds in Spanish, such as /g/ by /d/ in **piegra* (*piedra,* "stone") or /t/ and /k/ (written $<qu>$ in the example) **pateque* (*paquete,* "packet") and

**arrugatido* (*arrugadito,* "wrinkled") (Silva Rodríguez and Aragón Borja 2000).

The reduced size of W_D (40 words, 20 for English and 20 for Spanish covering simple and multiple dyslexic errors) is explained by: (1) the difficulty of finding texts written by people diagnosed with dyslexia; (2) the lack of corpora of dyslexic errors, with the exception of Pedler's (2007) corpus for English, and ours for Spanish texts produced by dyslexics; and (3) the strict criteria that we establish for selecting the misspelled words. These criteria aim to guarantee, as much as possible, the uniqueness and unambiguity of the dyslexic word, constraining their selection.

However our estimations of dyslexia would not vary much using a larger sample of words. In Figure 1 we show the convergence of DER using the average of k words (k from 1 to 10) for five different random orderings of the simple dyslexic errors in Spanish W_Dsp, that is, 10 words. We can see that already with seven of the words we get values similar to 10 words, so the results are already reliable above that. This shows that the accuracy of the measure improves as the size of W_D grows. We also give the sorted order of the ratios for W_Den and W_Dsp where we can see that the maximum and the minimum misspelling ratios differ by a factor of 4 for English and 20 for Spanish, being the maximum in Spanish for the pair **necestio* (*necesito*) "need". Both curves are similar and although the DER is not comparable across languages, this means that in our case the results will differ within one order of magnitude.

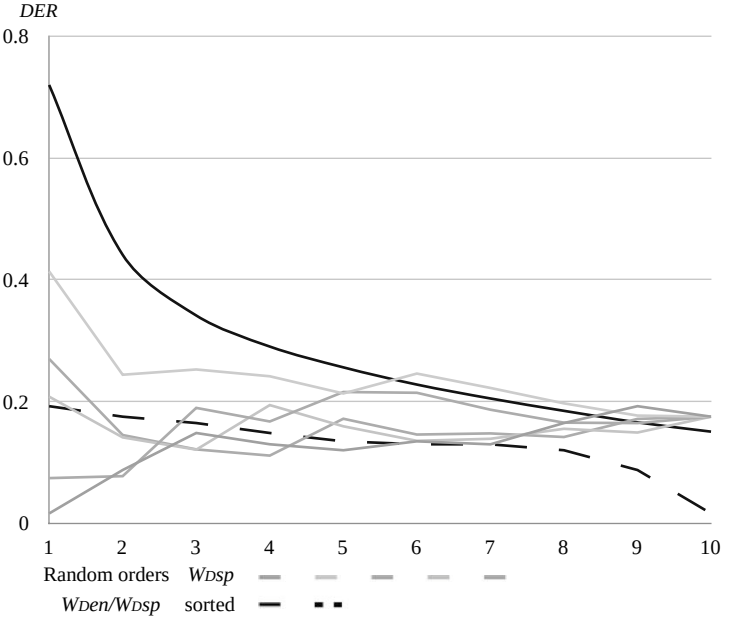

Figure 1. Partial DER for five random orderings of W_Dsp for the Web in Spanish, and the sorted individual misspell ratios for English and Spanish.

3.4. Adding Different kind of errors

Sample W_E (see Appendix 2) is composed of 479 words in English ($W_E en$) and Spanish ($W_E sp$). They are divided into 20 subgroups composed of the target word and the different type of errors related to the intended word. There are no frequent words and the words are relatively long (an average length of 9.3 letters per word). Our data sets are composed of long words compared with the average of letters per lemma (8.78) for the Royal Spanish Academy Dictionary (Real Academia Española 2001) and the average number of letters per word (4.5) occurring in English texts (Barnard 1955).

Sample W_E contains all types of errors. In order to detect lexical errors produced by dyslexic individuals, it is required to distinguish pure dyslexic errors within all lexical errors. Therefore, 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 by dyslexics, that is, unfinished words or letters, omitted words or inconsistent spaces between words and letters (Vellutino 1979), we only consider lexical errors, that is errors inside words such as multiple additions, transpositions, omissions, or substitutions of letters. For instance, **unforchanely* instead of *unfortunately* (Pedler 2007).
(2) Regular spelling errors produced by nonimpaired native speakers in English or Spanish, such as the transposition error **recieve* instead of *receive*.
(3) Regular typos caused by the adjacency of letters in the keyboard, i.e. **dituation* (*situation*).
(4) OCR errors, due to letters of similar shape, such as **tornorrow* (*tomorrow*).
(5) Errors made by nonnative speakers who use English or Spanish as a foreign language. For example, **receibe* (*receive*) is a typical error made by Spanish learners of English.

The other possibilities of errors related to the target with negligible frequency were discarded according to the conditions that W should fulfill (Section 3.1). Note that typos are behavioral errors, native and non-native misspellings are phonetic errors, OCR mistakes are visual errors, while dyslexic errors could be phonetic or visual.

3.5. Criteria for other errors

As dyslexic errors are the most difficult to find, our starting point was W_D. After identifying the dyslexic errors and their corresponding target word, we examined manually each of the different error types related to the target word. Then, we decided to include them or not in our sample according to this principle: the errors related to the target word need to be unique and not

ambiguous. For instance, the real word *worried* could also be a typo from the intended word *worries* because *s* and *d* are adjacent in the keyboard. Similarly, the typo **dxplain* (*explain*) is also a proper name. Hence, named entities and real word errors were dismissed, as well as target words with more than three ambiguous errors. The great majority of the candidates were ambiguous and did not match the criteria; as a consequence, our sample is composed of a reduced but reliable group of words.

The dyslexic errors contained in sample W_E are the multierrors from W_D, a subgroup of all the possible errors made by dyslexic people. However, in the set of dyslexic multierrors, not all the kinds of possible errors are taken into account. We also avoid taking into account errors which produce a syntactic anomaly, that is words that have no part of speech tags in common, such as the error **from* (*form*) or inflection errors, that is **storys* (*stories*).

Regular spelling errors were created taking into account their high frequency in query logs and also general spelling error patterns taken from literature (Yannakoudakis and Fawthrop 1983, Bustamante and Díaz 2006).

Regular typos, caused by the adjacency of letters in the keyboard, were generated by substituting each letter of the target word with the letter situated immediately left and right from the intended letter. Other cases have much smaller frequency (keys above or below). We discarded the cases in which the adjacent key was not a letter, that is **co,parison* (*comparison*).

For generating the OCR errors we substituted the typical letters which are usually mistaken, for instance, $c \rightarrow e$, $rn \rightarrow m$, or $cl \rightarrow d$ (Taghva and Stofsky 2001).

To find the typical errors made by nonnative speakers who use English or Spanish as foreign languages, we have taken into account errors caused by phonological transference from English or Spanish. For instance, **gobernment* is a typical error made by Spanish learners of English, because the graphemes $< b >$ and $< v >$ are pronounced as /b/, and the phoneme /v/ does not exist in the standard Spanish phonemic system. Besides, its translation in Spanish is written with $< b >$ ("gobierno").[14]

4. Results

In this section we present the results of using the sample W_D estimating the lower bound of dyslexic errors in the Web and diverse domains and websites (Sections 4.1 and 4.2). Later we use the sample W_E to estimate the percentages of different kinds of errors in the Web (Section 4.3), to end with a validation of our measure (Section 4.4).

4.1. Dyslexia in the Web

To compute DER we use Google to estimate the document frequency of each word in data set W_D. To validate our results we use exact frequencies from Yahoo!'s web search index. In Table 2 we present the prevalence of dyslexia of the Web for English and Spanish.

Table 2. Range and DER of different dyslexic errors in the Web.

	Multierrors		Simple errors		W_D
Dyslexic errors	Range	DER	Range	DER	DER
English	0.00*–0.007	0.0019	0.004–0.72	0.1374	0.0697
Spanish	0.00*–0.008	0.0018	0.016–0.414	0.1753	0.0886

A number larger than 0 but less than 0.0005 is represented by 0.00*.

To compare them with the rest of the Web, we chose six social media sites belonging to five different classes: blogs (Blogger) including micro-blogs (Twitter), social networks (Facebook), collaboration sites (Wikipedia), multimedia sites (YouTube) and opinions, including community question-answering systems (Y! Answers). To be able to estimate the overall impact of each site, we need to estimate the relative size of each of them. For this we use the total number of pages in the public content of each website according to Google's search engine. A search engine identifies this number by restricting the search to the pages on that site (this option is given in the Advanced Search page).

In Table 3 we compare each site and social media as a whole with other important web domains and the Web. For each site we also give the relative size of their (public) content.

Social media written in English has lower DER than the overall Web. However, compared with high-quality sites (*.org* and *.edu*), the presence of dyslexia in social media is higher in English. This should not be a surprise considering the diversity and sheer volume of social media content. It seems that the higher DER of *.com* may be due to the fact that most Web spam and social media content is part of *.com*.

Wikipedia values vary depending on the language and the kind of errors. We believe that the main contributor to this variability is the community section, because many examples were found in user, discussion, and project pages of Wikipedia. Websites with *.edu* domain have the lowest presence of dyslexic errors for both languages and, among the social media sites, Blogger had the lowest occurrence of dyslexic errors.

4.2. Geographical distribution

There are around 329 and 328 millions of Spanish and English speakers as first language, respectively. As a second language, English has the highest number of speakers reaching 1.4 billion, while Spanish is spoken as a second language by around 500 million people (Lewis 2009).

To compute the geographical distribution of dyslexic spelling errors among the countries where English and Spanish is spoken, we have taken into account the countries which have the highest populations of native English and Spanish speakers.

Table 3. Relative size, range, and DER for English and Spanish dyslexic errors in web domains and social media sites.

Domain/site	Size (%)	Multierrors		Simple errors		W_D
		Range	DER	Range	DER	DER
English						
.edu	1.66	0.00*–0.00*	0.00*	0.00*–0.074	0.0333	0.0167
.org	27.60	0.00*–0.001	0.0002	0.002–0.310	0.0614	0.0308
.com	70.74	0.00*–0.006	0.0011	0.010–0.793	0.1705	**0.0858**
Overall		0–0.005	0.0004	0.00*–0.793	0.0884	0.0444
Blogger	12.23	0–0.001	0.0001	0.001–0.142	0.0438	0.0220
Youtube	17.60	0.00*–0.00*	0.0001	0.001–0.277	0.0724	0.0363
Facebook	57.64	0.00*–0.001	0.0003	0.003–0.267	0.0742	0.0373
Twitter	9.98	0–0.00*	0.0001	0.00*–0.619	0.0804	0.0403
Wikipedia	0.40	0.00*–0.006	0.0008	0.008–0.502	0.1244	0.0626
Y! Answers	2.21	0.00*–0.002	0.0004	0.002–1.619	**0.2814**	**0.1409**
Overall		0.00*–0.006	0.0003	0.00*–1.619	0.1128	0.0566
English Web		0.00*–0.007	0.0012	0.006–0.749	0.1290	0.0651
Spanish						
.edu	0.90	0–0.00*	0.00*	0–0.036	0.0041	0.0021
.org	10.19	0–0.00*	0.0001	0.005–0.071	0.0348	0.0175
.com	88.91	0.00*–0.006	0.0002	0.032–0.525	**0.1927**	**0.0965**
Overall		0–0.006	0.0001	0–0.528	0.0772	0.0387
Blogger	15.05	0.00*–0.001	0.0002	0.007–0.075	0.0202	0.0102
Facebook	71.06	0–0.002	0.0004	0.001–0.068	0.0231	0.0118
Twitter	5.24	0–0.008	**0.0012**	0.004–0.068	0.0253	0.0133
Wikipedia	0.18	0.00*–0.002	**0.0078**	0–0.095	0.0315	0.0197
Y! Answers	2.81	0.00*–0.014	**0.0023**	0.036–0.153	0.0744	0.0384
Youtube	5.70	0.00*–0.005	0.0005	0.005–0.661	0.0846	0.0426
Overall		0.00*–0.014	**0.0011**	0.00*–0.661	0.0432	0.0222
Spanish Web		0.00*–0.008	0.0010	0.010–0.607	0.0996	0.0503

The values over the DER average are highlighted in bold face and 0.00* represents a number larger than 0 but less than 0.0005.

As it is not possible to distinguish countries among the *.com, .edu, .net* and *.org* domains, the websites were geographically identified by the country domain (see Tables 4 and 5). For instance, we consider US websites the ones with *.us* domain. According to the Internet Systems Consortium Domain Survey[15] there are 2.1 millions of computers in *.us* and according to a major search engine[16] there are more than 463 million web pages in that domain. In fact, many websites have both, the *.com* and the country domains.

For English, we consider countries where it is an official or *de facto* official language, or national language. These are, in descending order of native speakers (in parenthesis): USA (215 million (M)), United Kingdom (61 M), Canada (18.2 M), Australia (15.5 M), Nigeria (4 M), Ireland (3.8 M), South Africa (3.7 M), New Zealand (3.6 M), and Guyana (<1 M) (Wikipedia 2011).

Crystal (2003) estimates that nonnative speakers now outnumber native speakers by a ratio of 3 to 1. Despite this fact, this estimation depends on how

literacy or mastery of a language is defined and measured. So, we have added India (86.1 M) and Philippines (44 M), where English as a second language is widespread (Wikipedia 2011). However, in India and Philippines, only 0.2 and 3.4 millions of speakers have English as a first language, respectively. The fact that these two countries mainly use English as a second language may have an influence on a higher rate of spelling errors in general, India having the highest rate of errors. Philippines is the fourth place after United Kingdom and Ireland. The results are given in Table 4.

For Spanish we considered the countries where it is an official language. These countries also present the highest percentage of Spanish native speakers of their populations. They are, in descending order of native speakers (in parenthesis): Mexico (104.1 M), Colombia (45.7 M), Spain (42 M), Argentina (36.3 M), Venezuela (28.3 M), Peru (25.0 M), Chile (17.0 M), Ecuador (11.9 M), Cuba (11.2 M), Dominican Republic (10.0 M), Guatemala (8.6 M), Honduras (8.0 M), Bolivia (6.0 M), El Salvador (6.2 M), Nicaragua (5.3 M), Costa Rica (4.5 M), Puerto Rico (3.8 M), Paraguay (3.7 M), Uruguay (3.2 M), Panama (3.0 M), and Equatorial Guinea (1.7 M) (Wikipedia 2011). The results are given in Table 5.

As our percentages are relative, the size of the country domain shall not have a great influence in the error rate. In countries with small sizes such as Guyana and Equatorial Guinea, the low rate could be due to other reasons. For instance, even though English or Spanish are official languages in those countries, other native languages are spoken by the population as well; however official websites used English and Spanish. Surprisingly, *.us* has a relatively low rate of dyslexic errors maybe because of the fact that in the USA, the domain *.us* is less frequent than *.com* or *.net*, but the USA has the highest number of Internet users (Internet World Stats 2011). Notably, India,

Table 4. Relative size (%), range and DER for a sample of frequent dyslexic errors in several English speaking countries' domains.

Country, *Domain*	Size	Range	DER
India, *.in*	13.83	0.00*–0.004	**0.066**
UK, *.uk*	39.15	0.00*–0.004	**0.050**
Ireland, *.ie*	2.51	0.00*–0.002	**0.040**
Philippines, *.ph*	3.20	0.00*–0.001	0.034
Canada, *.ca*	10.08	0.00*–0.001	0.034
New Zealand, *.nz*	5.82	0.00*–0.001	0.032
Australia, *.au*	11.97	0.00*–0.001	0.028
USA, *.us*	5.13	0.00*–0.001	0.023
South Africa, *.za*	7.90	0.00*–0.001	0.022
Nigeria, *.ng*	0.29	0.00*–0.001	0.008
Guyana, *.gy*	0.12	0.00*–0.00*	0.006
Overall		0 –0.367	0.034

The values over the DER average are highlighted in bold face and 0.00* represents a number larger than 0 but less than 0.0005.

Table 5. Relative size (%), range and DER for a sample of frequent dyslexic errors in several Spanish speaking countries' domains.

Country, *Domain*	Size	Range	DER
Ecuador, *.ec*	4.85	0.00*–0.350	**0.046**
Spain, *.es*	38.56	0.00*–0.121	**0.041**
Chile, *.cl*	7.90	0.00*–0.138	**0.039**
Guatemala, *.gt*	0.24	0 –0.316	**0.039**
Argentina, *.ar*	14.70	0.00*–0.128	**0.034**
Peru, *.pe*	4.66	0–0.127	**0.028**
Bolivia, *.bo*	0.46	0–0.120	**0.026**
Puerto Rico, *.pr*	0.69	0–0.183	**0.021**
Costa Rica, *.cr*	0.84	0–0.067	0.019
Paraguay, *.py*	0.27	0–0.076	0.018
Venezuela, *.ve*	2.61	0.00*–0.058	0.018
Mexico, *.mx*	6.61	0.00*–0.039	0.015
Colombia, *.co*	5.88	0.00*–0.056	0.015
Dominic. Rep., *.do*	1.55	0.00*–0.055	0.014
Honduras, *.hn*	0.12	0–0.035	0.007
Nicaragua, *.ni*	0.88	0–0.031	0.007
Panama, *.pa*	0.44	0–0.020	0.006
Uruguay, *.uy*	5.50	0–0.010	0.005
Cuba, *.cu*	1.47	0–0.018	0.005
El Salvador, *.sv*	1.75	0–0.016	0.004
Eq. Guinea, *.gq*	0.01	0–0	0
Overall		0–0.350	0.019

The values over the DER average are highlighted in bold face and 0.00* represents a number larger than 0 but less than 0.0005.

United Kingdom, Ecuador, and Spain have the highest rate of dyslexic errors. Figure 2 compares the results for all the countries studied.

4.3. Dyslexic errors among all errors

Using W_E we computed the percentages of the different kinds of errors in the Web for English and Spanish (see Tables 6 and 7). Compared with other kinds of errors, the percentage of dyslexic errors is very low with an average of approximately 0.63% for English and 0.43% for Spanish. This percentage is very conservative because DER is a lower bound for dyslexia by definition and W_E only has dyslexic errors of the multi-error type, which are less frequent than simple errors. Simple dyslexic errors were not taken into consideration for W_E to avoid overlapping with other kind of errors.

We observe that the order of magnitude for all kinds of errors is the same in both languages, but dyslexic errors are more frequent in English than in Spanish compared with the other misspelling types.

4.4. Validating our measure

To validate DER we have: (1) corroborated that DER rates have different order of magnitude in the Web than in noncorrected corpora written by

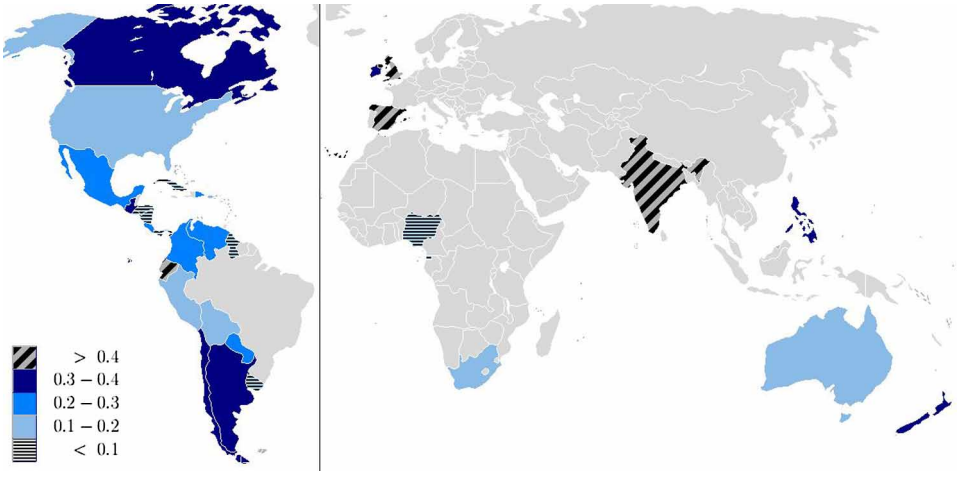

Figure 2. Geographical distribution of dyslexic errors in English- and Spanish-speaking countries' domains.

native and nonnative speakers and (2) checked that DER is not correlated with the general lexical quality of the Web, that is, web pages with a high DER do not have a higher spelling error rate due to a greater presence of misspells, not necessarily dyslexic errors.

For the first validation we took into consideration the largest corpora available for English and Spanish: the Collins Word banks Online[17] with 550 million words, the British National Corpus (BNC)[18] with 100 million words and the Royal Spanish Academy Corpus or Current Spanish (CREA)[19] with 3.5 million words. These corpora are made of written and spoken language (nonaphasic) from various sources and with no corrected errors, although as the sources are of high quality we would expect to have a much lower DER. We also took into account the only available corpus we found composed of English essays written by students who use English as a foreign language, the Janus Pannonius University Corpus (JPC).[20] In these corpora we only found examples of simple errors such as **poeple* but no examples of multierrors. We computed DER for these corpora (see Table 8) and for both English and Spanish the DER was negligible (DER = 0.001 for Spanish and

Table 6. Range, percentages, and average for the different error classes in English.

Error Class	Range		Average		Percentage	
	2010	2011	2010	2011	2010	2011
Spelling	1.100–91.570	2.579–92.985	0.2072	0.3515	51.99	55.77
Foreign	1.025–92.957	0.710–92.575	0.2733	0.2948	24.53	27.02
Typo	4.249–39.175	2.464–40.764	0.0602	0.0662	21.55	15.56
OCR	0.007–3.648	0.005–2.460	0.0070	0.0080	1.51	1.02
Dyslexia	0.004–3.400	0.008–3.563	0.0010	0.0019	0.42	0.63

Table 7. Range, percentages, and average for the different error classes in Spanish.

Error Class	Range 2011	Average 2011	Percentage 2011
Spelling	21.849–93.976	0.4010	54.11
Foreign	0.195–47.626	0.1001	26.51
Typo	2.874–43.772	0.0506	17.17
OCR	0.244–7.669	0.0051	1.79
Dyslexia	0.019–1.838	0.0018	0.43

DER = 0.002 for all English corpora). However note that as expected, DER in the corpus of less quality, JPC, was higher.

For the second validation we considered the general spelling errors rate (not only dyslexic errors) for web pages. To compute the spelling error rate we use the lexical quality (*LQ*) of web pages. *LQ* is a measure similar to DER (Baeza-Yates and Rello 2012) which takes into consideration all kind of possible errors in the Web and gives as a result an estimation of the web page error rate.

Intuitively, DER could be correlated to the overall *LQ* of a website, because when the general error rate grows, DER should grow too. To show that this is not the case we took the results related to the lexical quality of the Web presented in Baeza-Yates and Rello (2012) and computed the Pearson correlation between measures (DER and *LQ*) for the English- and Spanish-speaking countries and the major social media sites. The data used are normally distributed (Shapiro–Wilk test). In Table 9, we can observe that DER is not correlated with the LQ measure. This shows that a higher misspelling rate does not imply a higher DER.

Note that in all the corpora used, we assume that the majority of the authors involved are not dyslexics because: (1) the corpora are composed by written and spoken language made by nonaphasic from various sources, without mention to dyslexia or other aphasias in the descriptions of the corpora (BNC, Collins, CREA); plus the description of the corpus of essays written by students who use English as a foreign language (JPC) nothing is mentioned about the possible learning difficulties of the students; and (2) the measure of lexical quality uses the most frequent errors found in the Web, for instance, **becuase* (*because*), and most people sometimes make this kind of errors, not only dyslexics.

Table 8. DER in noncorrected corpora.

	English			Spanish
Corpus	BNC	Collins	JPC	CREA
DER	0.000*	0.001	0.005	0.001

The value 0.00* represents a number larger than 0 but less than 0.0005.

Table 9. Pearson correlation for DER and *LQ* measures in different web domains.

	English		Spanish	
Pearson correlation	Countries	Social Media	Countries	Social Media
LQ/DER	0.0967	−0.3626	−0.2003	−0.2048

Hence, although there is no conclusive evidence that these errors found in the Web were produced by dyslexic people, it is highly probable because of: (1) the strict criteria in the selection of the dyslexic errors (see Section 3.5); (2) the validation performed using noncorrected and nonnative speaker corpora for both languages; and (3) the fact that DER is not correlated with general lexical quality of the Web. Therefore, we believe that DER is a good estimation of the lower bound of the impact of dyslexic errors in the Web.

5. Concluding remarks

Our lower bound shows that at least 0.07% and 0.09% of the web pages contain dyslexic errors for the English and Spanish Web, respectively. These results could be surprising considering that the estimations of dyslexia among population are higher for English—from 5.3% (Katusic *et al.* 2001) to 17.5% (Interagency Commission on Learning Disabilities 1987)—than for Spanish—from 7.5% (Galván Gómez 2010) to 11.8% (Carrillo *et al.* 2011). However, these estimations are based on reading tests but not on writing misspells. If we take into consideration the error rates found in the corpora written by dyslexic people, then our results might be expected because there is a higher rate of multierrors in English (39%) than in Spanish (23%) and a higher rate of simple errors in Spanish (67%) than in English (53%). At the same time, multierrors are less frequent in the Web. Therefore, the high presence of simple errors in Spanish has an effect on the final estimation of DER being higher for Spanish than for English.

These results are conservative due to two reasons: (1) the fact that DER was designed as a lower bound for making the estimation in the Web feasible (see Section 3.1) and (2) the strict conditions that the words for the data sets must meet to assure as far as possible their dyslexic origin (see Section 3.3). For example, the errors in our data sets are long words, and previous research on dyslexia reveals that error frequency is related to word length (Sterling *et al.* 1998, Pedler 2007), errors in shorter words being more frequent than in longer ones.

Our results should be taken with care, because DER is a lower bound and there is no consensus on the definition of dyslexia, and previous user studies with dyslexics pointed out that dyslexia is highly variable and there is no typical nor universal profile of a dyslexic Internet user (Gregor *et al.* 2003, Pollak 2005). However, our estimations are useful to consider the prevalence of dyslexia in the Web as well as to motivate dyslexic-accessible practices.

Our main conclusions are that:

- The amount of dyslexic texts in the Web is not as large as it could be. This suggests that the widespread use of spell checkers ameliorates dyslexia in the Web so the prevalence of content with dyslexic errors is a function of both people and technology.
- The rate of dyslexic errors is independent from the rate of spelling errors in web pages. A comparison with our previous estimations made one year ago shows that the error percentages are growing. Hence, the increase of the Web does not correspond with the improvement of its lexical quality, which can be explained by the fact that each year we have a large number of new users.
- Spanish has a higher DER for simple dyslexic errors than English. However, if we compare the percentages of the different error kinds, there are more dyslexic errors in the English Web than in the Spanish Web.
- Even though Spanish orthography is shallower than English, the difference between these two languages in terms of DERs in the Web is not as substantial as expected.

As a byproduct we have found that other types of errors are much more frequent in the Web and this can be used to approach the quality of web text (Baeza-Yates and Rello 2012). In the future, we plan to create new corpora of dyslexic errors using the Web as corpus.

Acknowledgements

We thank Jennifer Pedler for sharing her English corpus of dyslexic errors and Berkant Barla Cambazoglu for his help getting part of the document frequencies of the words considered in this study.

Notes

[1] Dysgraphia refers to a writing disorder associated with the motor skills involved in not only writing, handwriting, and sequencing, but also orthographic coding (Berninger and Wolf 2009). It is comorbid with dyslexia (Nicolson and Fawcett 2011).
[2] These kinds are divided into acquired and developmental dyslexia, except for deep dyslexia, which is mostly acquired (Coltheart 1987).
[3] Phonological dyslexia was first reported—and coined—by Beauvois and Derouesne (1979).
[4] A nonword is a word that has no meaning, is not known to exist, or is disapproved, e.g. **approch* (*approach*) *or* **happisfaction* (Coltheart 1996).
[5] Surface dyslexia was fist modeled by Patterson *et al.* (1989).
[6] An example of a regularization would be saying /væs/ for the word $<$*vase*$>$/vaz/ (decorative container).
[7] In Kanji-based languages, such as Chinese, dyslexia is associated with multiple deficits, rather than with a core phonological deficit, because this writing system contains a large number of visual symbols or characters that represent units of meaning rather than phonemes as in an alphabet (Ho *et al.* 2002).
[8] *Penfriend XL* (http://www.penfriend.biz/).

[9] In fact, the distribution many times follows a power law, as the famous Britney Spears example: http://www.google.com/jobs/britney.html.
[10] Document frequency is the number of documents where a term appears (Baeza-Yates and Ribeiro-Neto 2011).
[11] These frequencies were obtained before Yahoo! started to use Bing.
[12] In this work, examples with errors are preceded by an asterisk "*".
[13] In the standard definition of edit distance, Levenshtein (1965) considered transpositions as two errors, while Damerau defined it as a single error.
[14] Phonemes are marked with "//" and graphemes with symbols " <>".
[15] www.isc.org/ds/ Consulted in April 2012.
[16] http://www.google.ca/advanced_search Accessed in April 2012.
[17] http://wordbanks.harpercollins.co.uk/auth/
[18] http://www.natcorp.ox.ac.uk/
[19] http://corpus.rae.es/creanet.html
[20] http://joeandco.blogspot.com.es/ Unfortunately, the rest of the corpora we found based on written essays of students are not available, such as, International Corpus of Learner English (ICLE), Written Corpus of Learner English (WriCLE), Corpus Escrito del Español L2 (CEDEL2), Uppsala Student English Corpus (USE), the Catalan-English Barcelona Corpus and Spencer Corpus.

References

American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR*, Washington, DC: American Psychiatric Publishing, Inc., 2000.

R. Baeza-Yates and L. Rello, "Estimating dyslexia in the web", in *International Cross Disciplinary Conference on Web Accessibility (W4A 2011)*, 28–29 March, Hyderabad, India: ACM Press, pp. 1–4, 2011.

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, Lyon, France: ACM Press, 2012.

R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search* (2nd edn), Second Harlow, UK: Addison Wesley, 2011.

G. Barnard, "Statistical calculation of word entropies for four western languages", *Transactions on Information Theory*, 1(1), pp. 49–53, 1955.

M. Beauvois and J. Derouesne, "Phonological alexia: Three dissociations", *Journal of Neurology, Neurosurgery and Psychiatry*, 42, pp. 1115–1124, 1979.

V. Berninger and B. Wolf, *Teaching Students With Dyslexia and Dysgraphia: Lessons From Teaching and Science*, Baltimore: Paul H Brookes Pub Co., 2009.

D. Bourassa and R. Treiman, "Spelling in children with dyslexia: Analyses from the Treiman–Bourassa early spelling test", *Scientific Studies of Reading*, 7(4), pp. 309–333, 2003.

N. Brunswick, "Unimpaired reading development and dyslexia across different languages", in *Reading and Dyslexia in Different Orthographies*, S. McDougall and P. de Mornay Davies (Eds), Hove: Psychology Press, pp. 131–154, 2010.

F.R. Bustamante and E. Díaz, "Spelling error patterns in Spanish for word processing applications", in *Proceedings of the International Conference on Language Resources and Evaluation* ELRA, Genoa, Italy, pp. 93–98, 2006.

M.S. Carrillo, J. Alegría, P. Miranda and N. Sánchez Pérez, "Evaluación de la dislexia en la escuela primaria: Prevalencia en español", *Escritos de Psicología*, 4(2), pp. 35–44, 2011.

C. Coleman, N. Gregg, L. McLain and L.W. Bellair, "A comparison of spelling performance across young adults with and without dyslexia", *Assessment for Effective Intervention*, 34(2), pp. 94–105, 2009.

Col legi de Logopedes de Catalunya, *PRODISCAT Protocol de detecció i actuació en la dislèxia. Àmbit Educativo* (Protocol of detection and action of dyslexia. Education scope). Barcelona: Departament d'Ensenyament de la Generalitat de Catalunya, 2011.

M. Coltheart, "Disorders of reading and their implications for models of normal reading", *Visible Language*, XV, pp. 245–286, 1981.

M. Coltheart, "Varieties of developmental dyslexia: A comment on Bryant and Impey", *Cognition*, 27, pp. 97–101, 1987.
M. Coltheart, "Phonological dyslexia: Past and future issues", *Cognitive Neuropsychology*, 13(6), pp. 749–762, 1996.
M. Coltheart, K. Patterson and J.C. Marshall, *Deep Dyslexia*, London: Routledge and Kegan Paul, 1980.
V. Connelly, S. Campbell, M. MacLean and J. Barnes, "Contribution of lower order skills to the written composition of college students with and without dyslexia", *Developmental Neuropsychology*, 29(1), pp. 175–196, 2006.
D. Crystal, *English as a Global Language*, Cambridge: Cambridge University Press, 2003.
F. Damerau, "A technique for computer detection and correction of spelling errors", *Communications of the A.C.M.*, 7, pp. 171–176, 1964.
G. Deloche, E. Andreewsky and M. Desi, "Surface dyslexia: A case report and some theoretical implications to reading models", *Brain and Language*, 15(1), pp. 12–31, 1982.
M. Dixon, "Comparative study of disabled vs. non-disabled evaluators in user-testing: Dyslexia and first year students learning computer programming", *Universal Access in Human Computer Interaction. Coping with Diversity*, 4554, pp. 647–656, 2007.
L. Evett and D. Brown, "Text formats and web design for visually impaired and dyslexic readers-clear text for all", *Interacting with Computers*, 17, pp. 453–472, 2005.
R. Friedman, "Recovery from deep alexia to phonological alexia: Points on a continuum", *Brain and Language*, 52(1), pp. 114–128, 1996.
R. Galván Gómez, "Dislexia va en aumento", *El Peruano*, 8th January, p. 22, 2010.
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), Napa Valley, CA: ACM Press, pp. 43–46, 2008.
N. Goulandris, *Dyslexia in Different Languages: Cross-linguistic Comparisons*, London, UK: Whurr Publishers, 2003.
P. Gregor, A. Dickinson, A. Macaffer and P. Andreasen, "See Word: A personal word processing environment for dyslexic computer users", *British Journal of Educational Technology*, 34(3), pp. 341–355, 2003.
P. Gregor and A.F. Newell, "An empirical investigation of ways in which some of the problems encountered by some dyslexics may be alleviated using computer techniques", in *Proceedings of the Fourth International ACM Conference on Assistive Technologies*, ASSETS 2000, Arlington, Virginia, New York, NY: ACM, pp. 85–91, 2000.
C. Ho, D.W.O. Chan, S.M. Tsang and S.H. Lee, "The cognitive profile and multiple-deficit hypothesis in Chinese developmental dyslexia", *Developmental Psychology*, 38(4), p. 543, 2002.
Interagency Commission on Learning Disabilities, *Learning Disabilities: A Report to the U.S. Congress*, Washington DC: Government Printing Office, 1987.
International Dyslexia Association, 2011a. Definition of dyslexia. Available online at: http://interdys.org/DyslexiaDefinition.htm. Based in the initial definition of the Research Committee of the Orton Dyslexia Society, former name of the IDA, done in 1994 (accessed 12 September 2012).
International Dyslexia Association, 2011b. Frequently asked questions about dyslexia. Available online at: http://www.interdys.org/FAQ.htm (accessed 12 September 2012).
Internet World Stats, 2011. Usage and population statistics. Available online at: http://www.internetworldstats.com (accessed 11 November 2011).
J.E. Jiménez, R. Guzmán, C. Rodríguez and C. Artiles, "Prevalencia de las dificultades específicas de aprendizaje: La dislexia en español", *Anales de Psicología*, 25(1), pp. 78–85, 2009.
S. Katusic, R.C. Colligan, W.J. Barbaresi, D.J. Schaid and S.J. Jacobsen, "Incidence of reading disability in a population-based birth cohort, 1976–1982, Rochester, MN", *Mayo Clinic Proceedings*, 76(11), p. 1081, 2001.
S. Kurniawan and G. Conroy, "Comparing comprehension speed and accuracy of online information in students with and without dyslexia", in *Advances in Universal Web Design and Evaluation: Research, Trends and Opportunities*, Sri Kurniawan (Ed.) Hershey, PA: Idea Group Publishing, pp. 257–270, 2006.
V. Levenshtein, "Binary codes capable of correcting spurious insertions and deletions of ones", *Problems of Information Transmission*, 1, pp. 8–17, 1965.
M.P. Lewis, *Ethnologue: Languages of the World*, Sixteenth Dallas, TX: SIL International, 2009.

S. Lindgrén and M. Laine, "Multilingual dyslexia in university students: Reading and writing patterns in three languages", *Clinical Linguistics and Phonetics*, 25(9), pp. 753–766, 2011.
G. Lyon, "Toward a definition of dyslexia", *Annals of Dyslexia*, 45(1), pp. 3–27, 1995.
G. Lyon, S. Shaywitz and B. Shaywitz, "A definition of dyslexia", *Annals of Dyslexia*, 53(1), pp. 1–14, 2003.
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, pp. 147–152, 2010.
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", *Proceedings of the National Academy of Sciences*, 102, pp. 17053–17058, 2005.
L. Moats, "Phonological spelling errors in the writing of dyslexic adolescents", *Reading and Writing*, 8(1), pp. 105–119, 1996.
R. Nicolson and A. Fawcett, "Dyslexia, dysgraphia, procedural learning and the cerebellum", *Cortex*, 47(1), pp. 117–127, 2011.
K. Patterson, M.S. Seidenberg and J.L. McClelland, "Connections and disconnections: Acquired dyslexia in a computational model of reading processes", Oxford Psychology Series, in *Parallel Distributed Processing: Implications for Psychology and Neurobiology*, R.G.M. Morris (Ed.), Oxford: Clarendon Press/ Oxford University Press, pp. 131–181, 1989.
K. Patterson, T. Suzuki, T. Wydell and S. Sasanuma, "Progressive aphasia and surface alexia in Japanese", *Neurocase*, I, pp. 155–165, 1995.
E. Paulesu, J.-F. Démonet, F. Fazio, E. McCrory, V. Chanoine, N. Brunswick, S.F. Cappa, G. Cossu, G.M. Habib, C.D. Frith and U. Frith, "Dyslexia: Cultural diversity and biological unity", *Science*, 291(5511), pp. 2165–2167, 2001.
J. Pedler, "The detection and correction of real-word spelling errors in dyslexic text", in *Proceedings of the 4th Annual CLUK Colloquium*, Sheffield: CLUK, pp. 115–119, 2001a.
J. Pedler, "Computer spellcheckers and dyslexics – A performance survey", *British Journal of Educational Technology*, 32(1), pp. 23–37, 2001b.
J. Pedler, *Computer correction of real-word spelling errors in dyslexic text*, Thesis (PhD), Birkbeck College, London University, UK, 2007.
C. Perfetti and L. Hart, "The lexical quality hypothesis", in L. Verhoeven, C. Elbro and P. Reitsma (Eds.) *Precursors of Functional Literacy*, pp. 189–213, Amsterdam/Philadelphia: John Benjamins, 2002.
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, Beijing, China, New York, NY: ACM, pp. 25–28, 2008.
D. Pollak, *Dyslexia, the Self and Higher Education: Learning life Histories of Students Identified as Dyslexic*, Stoke-on-Trent: Trentham Books Ltd., 2005.
Real Academia Española, *Diccionario de la lengua española* (Royal Spanish Academy Dictionary) (22nd edn), Madrid: Espasa-Calpe, 2001.
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, Lyon, France: ACM Press 2012.
L. Rello, R. Baeza-Yates, H. Saggion and J. Pedler, "A first approach to the creation of a Spanish corpus of dyslexic texts", in LREC Workshop Natural Language Processing for Improving Textual Accessibility (NLP4ITA), May, Istanbul, Turkey: Eighth international conference on Language Resources and Evaluation, 2012a.
L. Rello, C. Bayarri and A. Gorriz, "What is wrong with this Word? Dyseggxia: A game for children with dyslexia", in *The 14th International ACM SIGACCESS Conference of Computers and Accessibility (ASSETS 2012)*, October, Boulder, CO: ACM Press, 2012b.
L. Rello, G. Kanvinde and R. Baeza-Yates, "Layout guidelines for web text and a web service to improve accessibility for dyslexics", in *International Cross Disciplinary Conference on Web Accessibility (W4A 2012)*, April, Lyon, France: ACM Press, 2012c.
C. Romani, J. Ward and A. Olson, "Developmental surface dysgraphia: What is the underlying cognitive impairment?", *The Quarterly Journal of Experimental Psychology*, 52(1), pp. 97–128, 1999.
G. Schulte-Körne, W. Deimel, K. Müller, C. Gutenbrunner and H. Remschmidt, "Familial aggregation of spelling disability", *Journal of Child Psychology and Psychiatry*, 37(7), pp. 817–822, 1996.
P. Seymour, M. Aro and J. Erskine, "Foundation literacy acquisition in European orthographies", *British Journal of Psychology*, 94(2), pp. 143–174, 2003.

S. Shaywitz, M.D. Escobar, B.A. Shaywitz, J.M. Fletcher and R. Makuch, "Evidence that dyslexia may represent the lower tail of a normal distribution of reading ability", *New England Journal of Medicine*, 326(3), pp. 145–150, 1992.
A. Silva Rodríguez and L. Aragón Borja, "Análisis cualitativo de un instrumento para detectar errores de tipo disléxico (IDETID-LEA)", *Psicothema*, 12(2), pp. 35–38, 2000.
C. Sterling, M. Farmer, B. Riddick, S. Morgan and C. Matthews, "Adult dyslexic writing", *Dyslexia*, 4(1), pp. 1–15, 1998.
K. Taghva and E. Stofsky, "OCRSpell: An interactive spelling correction system for OCR errors in text", *International Journal on Document Analysis and Recognition*, 3(3), pp. 125–137, 2001.
R. Treiman, "Spelling in normal children and dyslexics", in B. Blachman (Ed.), *Foundations of Reading Acquisition and Dyslexia: Implications for Early Intervention*, Mahwah, NJ: Lawrence Erlbaum Associates, pp. 191–218, 1997.
F.R. Vellutino, *Dyslexia. Theory and Research*, Cambridge, MA: The MIT Press, 1979.
F. Vellutino, J.M. Fletcher, M.J. Snowling and D.M. Scanlon, "Specific reading disability (dyslexia): What have we learned in the past four decades?", *Journal of Child Psychology and Psychiatry*, 45(1), pp. 2–40, 2004.
Wikipedia, 2011. Wikipedia, the free encyclopedia. Available online at: http://www.wikipedia.org (accessed 11 November 2011).
World Health Organization, *International Statistical Classification of Diseases, Injuries and Causes of Death (ICD-10)*, Tenth World Health Organization, Geneva: World Health Organization (WHO), 1993.
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), pp. 1846–1856, 2008.
E. Yannakoudakis and D. Fawthrop, "The rules of spelling errors", *Information Processing and Management*, 19(2), pp. 87–99, 1983.

Appendix 1: Sample W_D

For the Sample W_D show the error kind, the error, the target word and the source (in parenthesis) from where the error was extracted.

English Sample $W_D en$:

(1) Simple errors:
  (a) Substitution:
    **studends (students)* (Sterling *et al.* 1998).
  (b) Insertion:
    **promblem (problem)* and **deleteing (deleting)* (Pedler 2007).
  (c) Omission:
    **approch (approach)* (Pedler 2007); **carful (careful)* (Pedler 2001b); **construtions (constructions)* (Sterling *et al.* 1998).
  (d) Transposition:
    **worng (wrong)*, **artcile (article)* (Pedler 2007); **childern (children)* and **poeple (people)* (Pedler 2007).

(2) Multierrors:
    **situartion (situation)* (Pedler 2007); **exaplin (explain)*, **confusetion (confusion)*, **torromow (tomorrow)*, **knwolegde (knowledge)*, **comaprsion (comparison)*, **intersenting (interesting)*, **worires (worries)*, **understangind (understanding)* and **inpossbile (impossible)* (dyslexic subject).

Spanish Sample $W_D sp$:

(1) Simple errors:
  (a) Substitution:
    **probrema (problema)* (dyslexic subject).
  (b) Insertion:
    **docotorado (doctorado)* and **escribies (escribes)* (dyslexic subject).
  (c) Omission:
    **gande (grande)*, **hombes (hombres)* and **pegunta (pregunta)* (Silva Rodríguez and Aragón Borja 2000).
  (d) Transposition:
    **cambaido (cambiado)*, **hablamso (hablamos)*, **necestio (necesito)*, **tmabin (tambin)* (dyslexic subject).
(2) Multierrors:
  **entenmiento (entendimiento)* and **sechora (señora)* (Silva Rodríguez and Aragón Borja 2000); **felicdidad (felicidad)*, **incleibre (increíble)*, **respondodido (respondido)*, **cominucaion (comunicación)*, **contimigo (conmigo)*, **sufieicnte (suficiente)*, **tambpo (tampoco)* and **temriando (terminando)* (dyslexic subject).

Appendix 2: Sample W_E

For the Sample W_E show the target word, the error kind and the variants of each error type.

English Sample $W_E en$:

(1) Target word: *comparison*.
  (a) Dyslexic: **comaprsion*.
  (b) Spelling: **comparision, *conparison* and **comprison*.
  (c) Typo: **vomparison, *xomparison, *cimparison, *cpmparison, *conparison, *comoarison, *comprison, *compsrison, *compaeison, *compatison, *comparuson, *comparoson, *compariaon, *comparidon, *comparisin, *comparispn, *comparisob* and **comparisom*.
  (d) OCR: **compaiison* and **comparisom*.
  (e) Foreign: **comparition* and **comparizon*.
(2) Target word: *confusion*.
  (a) Dyslexic: **confusetion*.
  (b) Spelling: **confussion*.
  (c) Typo: **xonfusion, *vonfusion, *cinfusion, *cpnfusion, *cobfusion, *comfusion, *condusion, *congusion, *confysion, *confision, *confuaion, *confudion, *confusuon, *confusoon, *confusiin, *confusipn, *confusiob* and **confusiom*.
  (d) OCR: **coniusion* and **confuslon*.
  (e) Foreign: **confution*.
(3) Target word: *explain*.
  (a) Dyslexic: **exaplin*.
  (b) Spelling: **explane*.

(c) Typo: **wxplain, *rxplain, *ezplain, *ecplain, *exolain, *expkain, *explin, *explsin, *explaun, *explaon, *explaib* and **explaim*.
(d) OCR: **explaln*.
(e) Foreign: **esplain*.

(4) Target word: *impossible*.
(a) Dyslexic: **inpossbile*.
(b) Spelling: **umpossible, inpossible* and **anpossible*.
(c) Typo: **umpossible, *ompossible, *imoossible, *impissible, *imppssible, *impoaaible, *impoddible, *impossuble, *impossoble, *impossivle, *impossinle, *impossibke, *impossiblw* and **impossiblr*.
(d) OCR: **imposlble, *imposslble, *inpossible* and **impossibie*.
(e) Foreign: **imposible* and **impozible*.

(5) Target word: *interesting*.
(a) Dyslexic: **intersenting*.
(b) Spelling: **intresting*.
(c) Typo: **unteresting, *onteresting, *ibteresting, *imteresting, *inreresting, *inyeresting, *intwresting, *intrresting, *inteeesting, *intetesting, *interwsting, *interrsting, *intereating, *interedting, *interesring, *interesying, *interestung, *interestong, *interestibg, *interestimg, *interestinf* and **interestinh*.
(d) OCR: **interesiing, *inieresting* and **inleresting*.
(e) Foreign: **intrestin*.

(6) Target word: *knowledge*.
(a) Dyslexic: **knwolegde*.
(b) Spelling: **nowledge*.
(c) Typo: **jnowledge, *lnowledge, *kbowledge, *kmowledge, *kniwledge, *knpwledge, *knoqledge, *knoeledge, *knowkedge, *knowlwdge, *knowlrdge, *knowlesge, *knowlefge, *knowledfe, *knowledhe, *knowledgw* and **knowledgr*.
(d) OCR: **knowiedge* and **knowledqe*.
(e) Foreign: **knowlegde* and **noledge*.

(7) Target word: *situation*.
(a) Dyslexic: **situartion*.
(b) Spelling: **situacion*.
(c) Typo: **aituation, *dituation, *sutuation, *sotuation, *siruation, *siyuation, *sityation, *sitiation, *sitution, *situstion, *situarion, *situayion, *situatuon, *situatoon, *situatiin, *situatipn, *situatiob* and **situatiom*.
(d) OCR: **situaiion* and **siluation*.
(e) Foreign: **situasion*.

(8) Target word: *tomorrow*.
(a) Dyslexic: **torromow*.
(b) Spelling: **toomorrow*.
(c) Typo: **romorrow, *yomorrow, *timorrow, *tpmorrow, *tonorrow, *tomirrow, *tomprrow, *tomoeeow, *tomottow, *tomorriw, *tomorrpw, *tomorroq* and **tomorroe*.
(d) OCR: **tomorrov, *tamarraw* and **tonorrow*.
(e) Foreign: **tomorow* and **tomorou*.

(9) Target word: *understanding*.
  (a) Dyslexic: **understangind*.
  (b) Spelling: **understend* and **understandin*.
  (c) Typo: **ynderstanding, *inderstanding, *ubderstanding, *umderstanding, *unserstanding, *unferstanding, *undwrstanding, *undrrstanding, *undeestanding, *undetstanding, *undertanding, *underatanding, *undersranding, *undersyanding, *understnding, *understsnding, *understabding, *understamding, *understansing, *understanfing, *understandung, *understandong, *understandibg, *understandimg, *understandinf* and **understandinh*.
  (d) OCR: **undersianding* and **understandinq*.
  (e) Foreign: **underestanding* and **underestandin*.

(10) Target word: *worries.*
  (a) Dyslexic: **worires.*
  (b) Spelling: **worrys.*
  (c) Typo: **qorries, *eorries, *wirries, *wprries, *woeeies, *wotties, *worrues, *worroes, *worriws, *worrirs, *worriea* and **worried.*
  (d) OCR: **woiiies.*
  (e) Foreign: **worryes.*

Spanish Sample W_{ESP}:

(1) Target word: *comunicación*.
  (a) Dyslexic: **cominucaion*.
  (b) Spelling: **comunicasion*.
  (c) Typo: **xomunicación, *vomunicación, *cimunicación, *cpmunicación, *conunicación, *comynicación, *comumicación, *comunocación, *comunixación, *comunivación, *comunicsción, *comunicaxión, *comunicavión, *comunicacuón, *comunicacoón, *comunicaciin, *comunicacipn, *comunicaciób* and **comunicacióm*.
  (d) OCR: **comunicadón, *cornunicación* and **comunlcación*.
  (e) Foreign: **conmunicación, *cominicación* and **comunucación*.

(2) Target word: *conmigo*.
  (a) Dyslexic: **contimigo*.
  (b) Spelling: **cinmigo, *comego* and **commigo*.
  (c) Typo: **xonmigo, *vonmigo, *cinmigo, *cpnmigo, *cobmigo, *conmugo, *conmofo, *conmigi* and **conmigp.*
  (d) OCR: **conmlgo.*
  (e) Foreign: **conmiguo, *connigo* and **conmiho*.

(3) Target word: *entendimiento*.
  (a) Dyslexic: **entenmiento*.
  (b) Spelling: **entendimento* and **entindimiento*.
  (c) Typo: **wntendimiento, *rntendimiento, *ebtendimiento, *emtendimiento, *enrendimiento, *enyendimiento, *entwndimiento, *entrndimiento, *entebdimiento, *entemdimiento, *entensimiento, *entenfimiento, *entendumiento, *entendomiento, *entendiniento, *entendimuento, *entendimoento, *entendimiwnto, *entendimirnto, *en-*

*tendimiebto, *entendimiemto, *entendimienro, *entendimienyo, *entendimienti* and **entendimientp.*

(d) OCR: **entcndimiento* and **entendimicnto.*

(e) Foreign: **intendimiento* and **entendimient.*

(4) Target word: *felicidad.*

(a) Dyslexic: **felicdidad.*

(b) Spelling: **felizidad.*

(c) Typo: **delicidad, *gelicidad, *fwlicidad, *frlicidad, *fekicidad, *felucidad, *felocidad, *felividad, *felicudad, *felicodad, *felicisad, *felicifad, *felicidsd* and **felicidaf.*

(d) OCR: **fellcidad* and **felieidad.*

(e) Foreign: **felisidad, *felicidaz* and **felicidas.*

(5) Target word: *increíble.*

(a) Dyslexic: **incleibre.*

(b) Spelling: **increible.*

(c) Typo: **uncreíble, *oncreíble, *ibcreíble, *inxreíble, *invreíble, *inceeíble, *incteíble, *incrwíble, *incrríble, *increuble, *increívle, *increínle, *increíbke, *increíblw* and **increíblr.*

(d) OCR: **increíbie* and **lncreíble.*

(e) Foreign: **hincreible, *imcreíble* and **incraible.*

(6) Target word: *respondido.*

(a) Dyslexic: **respondodido.*

(b) Spelling: **repondio.*

(c) Typo: **eespondido, *tespondido, *rwspondido, *rrapondido, *reapondido, *redpondido, *resoondido, *respindido, *resppndido, *respobdido, *respomdido, *responsido, *responfido, *respondudo, *respondodo, *respondiso, *respondifo, *respondidi* and **respondidp.*

(d) OCR: **rcspondido* and **respondiclo.*

(e) Foreign: **respodido.*

(7) Target word: *señora.*

(a) Dyslexic: **sechora.*

(b) Spelling: **siñora.*

(c) Typo: **aeñora, *deñora, *swñora, *srñora, *sebora, *señira, *señpra, *señoea* and **señota.*

(d) OCR: **scñora* and **señom.*

(e) Foreign: **segnora* and **sennora.*

(8) Target word: *suficiente.*

(a) Dyslexic: **sufieicnte.*

(b) Spelling: **sufuciente* and **sificiente.*

(c) Typo: **eespondido, *tespondido, *rwspondido, *rrapondido, *reapondido, *redpondido, *resoondido, *respindido, *resppndido, *respobdido, *respomdido, *responsido, *responfido, *respondudo, *respondodo, *respondiso, *respondifo, *respondidi* and **respondidp.*

(d) OCR: **sufidente* and **suficlente.*

(e) Foreign: **sificiente.*

(9) Target word: *tampoco*.
   (a) Dyslexic: **tambpo*.
   (b) Spelling: **tanpoco*.
   (c) Typo: **rampoco, *yampoco, *tsmpoco, *tamooco, *tamppco, *tampoxo, *tampovo, *tampoci* and **tampocp*.
   (d) OCR: **tampoeo*.
   (e) Foreign: **tanpoko*.

(10) Target word: *terminando*.
   (a) Dyslexic: **temriando*.
   (b) Spelling: **terminao*.
   (c) Typo: **rerminando, *yerminando, *twrminando, *trrminando, *teeminando, *tetminando, *terninando, *termunando, *termonando, *termibando, *termimando, *terminsbdo, *terminamdo, *terminanso, *terminanfo, *terminandi* and **terminandp*.
   (d) OCR: **terrninando, *termlnando, *terminanclo* and **tcrminando*.
   (e) Foreign: **termenando* and **tirminando*.