A First Approach to the Creation of a Spanish Corpus of Dyslexic Texts

Luz Rello[1,2], Ricardo Baeza-Yates[3,2], Horacio Saggion[1], Jennifer Pedler[4]

[1]NLP and [2]Web Research Groups, Universitat Pompeu Fabra, Barcelona, Spain
[3]Yahoo! Research, Barcelona, Spain
[4]Dept. of Computer Science, Birkbeck, University of London
luzrello@acm.org, rbaeza@acm.org, horacio.saggion@upf.edu, jenny@dcs.bbk.ac.uk

Abstract

Corpora of dyslexic texts are valuable for studying dyslexia and addressing accessibility practices, among others. However, due to the difficulty of finding texts written by dyslexics, these kind of resources are scarce. In this paper, we introduce a small Spanish corpus of dyslexic texts with annotated errors. Since these errors require non-standard annotation, we present the annotation criteria established for the different types of dyslexic errors. We compare our preliminary findings with a similar corpus in English. This comparison suggests that the corpus shall be enlarged in future work.

Keywords: Corpus, Non-standard annotation, Errors, Dyslexia.

1. Introduction

Worldwide, around 15-20% of the population has a language-based learning disability; where 70-80% of them are likely dyslexic (International Dyslexia Association, 2011).

Regarding this substantial group of people, various accessibility studies take dyslexia into account. They mainly focus on tools (Pedler, 2007; Gregor et al., 2003) and guidelines for dyslexic-accessible practices (McCarthy and Swierenga, 2010). There is a common agreement in these studies that the application of dyslexic-accessible practices benefits also the readability for non-dyslexic users as well as other users with disabilities such as low vision (Evett and Brown, 2005).

Although the use of corpora of dyslexic errors have been used for various purposes such as diagnosing dyslexia (Schulte-Körne et al., 1996) and developing tools, i.e. spell checkers (Pedler, 2007), their existence is scarce.

In this paper we present the following contributions:

- The first approach to create a corpus of dyslexic errors in Spanish,
- guidelines for the annotation of dyslexic errors and,
- a comparison of our corpus with a similar corpus in English.

In the next section we make a brief explanation of dyslexia and explain in Section 3 how dyslexic errors have been used for different purposes. In Section 4 we describe our related work, Pedler's corpus of dyslexic texts in English (Pedler, 2007), and in Section 5 we present a classification of the dyslexic errors. Sections 6 and 7 detail the characteristics of our corpus and its annotation guidelines. In Section 8 we compare the distribution of dyslexic errors in English and Spanish. Conclusions and future work are drawn in Section 9.

2. What is Dyslexia?

Dyslexia is a specific learning disability which is neurological in 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 et al., 2003; Orton Dyslexia Society Research Committee, 1994).

In some literature, dyslexia is referred to as a specific reading disability (Vellutino et al., 2004) and dysgraphia its writing manifestation (Romani et al., 1999).[1] However, our study follows the standard definitions of ICD-10 and DSM-IV (World Health Organization, 1993; American Psychiatric Association, 2000) where dyslexia is listed as a reading and spelling disorder.

Despite its universal neurocognitive basis, dyslexia manifestations are variable and culture-specific (Goulandris, 2003). This variability is due to the different language orthographies concerning their grade of consistency and regularity (Brunswick, 2010). English has an opaque –or deep– orthography in which the relationships between letters and sounds are inconsistent and many exceptions are permitted. English presents a significantly greater challenge to the beginning reader than other languages, such as Spanish, with a more regular alphabetic system that contains consistent mappings between letters and sounds, that is, a transparent –or shallow– orthography.

Depending on the language, the estimations on the prevalence of dyslexia varies. The (Interagency Commission on Learning Disabilities, 1987) states that 10-17.5% of the population in the U.S.A. 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.

[1]Dysgraphia refers to a writing disorder associated with the motor skills involved in writing, handwriting and sequencing, but also orthographic coding (Romani et al., 1999). It is comorbid with dyslexia, that is, it is a medical condition that co-occurs with dyslexia (Nicolson and Fawcett, 2011).

3. The Use of Dyslexic Errors

In general terms, errors could be used as a source of knowledge. For instance, the presence of errors in the textual Web have been used for detecting spam (Piskorski et al., 2008), measuring quality (Gelman and Barletta, 2008) and understandability (Rello and Baeza-Yates, 2012) of web content. Among the different kind of errors found in the Web, at least 0.67% errors are only made by dyslexic users (Baeza-Yates and Rello, 2011). In the case of people with dyslexia, their written errors have been used for various accessibility related purposes such as the development of tools like spell checkers (Pedler, 2007) or word processors (Gregor et al., 2003).

Besides the accessibility practices, analyses of writing errors made by dyslexics have been used in previous literature to study different aspects of dyslexia. 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). People with dyslexia exhibit higher spelling error rates than non-dyslexic people (Coleman et al., 2009) and, due to this fact, there are diagnosis of dyslexia based on the spelling score (Schulte-Körne et al., 1996). According to (Meng et al., 2005) only 30% of dyslexics have trouble with reversing letters and numbers. However, errors attributable to phonological impairment, spelling knowledge, and lexical mistakes are more frequent in dyslexics than in non-dyslexics (Sterling et al., 1998). Nonetheless, the dyslexic error rate vary depending on the language writing system (Lindgrén and Laine, 2011).

4. Related Work

To the best of our knowledge, there is only one corpus of dyslexic texts, the corpus used by Pedler (2007) for the creation of a spell checker of real-word errors made by dyslexic people.

This corpus in English is composed of 3,134 words and 363 errors (Pedler, 2007). This corpus is made of: (1) word-processed homework (saved before it was spellchecked) produced by a third year secondary school student; (2) two error samples used for a comparative test of spellcheckers (Mitton, 1996); and (3) short passages of creative writing produced by secondary school children of low academic ability in the 1960s (Holbrook, 1964).

To develop a program designed to correct actual errors made by dyslexics, this initial corpus was enlarged to 12,000 words containing just over 800 real-word errors.[2] The additional sources for that corpus were: texts from a dyslexic student, texts from an online typing experiment (Spooner, 1998), samples from dyslexic bulletin boards and mailing lists and stories written by dyslexic children.

All the errors in this corpus were annotated in the format illustrated next, where **pituwer* is the dyslexic error from the intended work *picture*.[3]

```
<ERR targ=picture> pituwer </ERR>
```

Our current annotation method is inspired by Pedler's work (2007) and is described in Section 7.

5. Types of Dyslexic Errors

Pedler (2007) found the following kinds of dyslexic errors in her corpus and proposed the following classification of dyslexic errors:

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)*,
      ii. insertion, **situartion (situation)*,
      iii. omission, **approch (approach)* and
      iv. transposition, **artcile (article)*.

      In (Damerau, 1964), 80% of the misspellings in his corpus (non-dyslexic errors) were simple errors.[4]
   (b) Multi-errors. They differ in more than one letter from the target word. Some errors, such as **queraba (quedara, 'stayed')*, 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)* while 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) Non-word errors. Misspellings that do not result in another correct word, such as **conmitigo (contigo, 'with you')*
3. First letter dyslexic errors:
   (a) First letter errors, like **no (know)*.

6. Spanish Corpus of Dyslexic Texts

Manifestations of dyslexia varies among languages (Goulandris, 2003) but also among subjects and among ages (Vellutino et al., 2004). For instance misspelling rate in dyslexic children is higher than in adults (Sterling et al.,

[2] A corpus containing 833 dyslexic real-word errors in context is available at: `http://www.dcs.bbk.ac.uk/~jenny/resources.html`

[3] Dyslexic errors are preceded by * while the intended target word follows in parenthesis.

[4] The standard definition of edit distance (Levenshtein, 1965) consider transpositions as two errors, while Damerau defined it as a single error.

1998). However, experiments evidence that adult dyslexics have a continuing problem in the lexical domain, manifested in poor spelling ability (Sterling et al., 1998).
Due to this variability, we pursued to collect texts written by a similar population in terms of age, education, native language and diagnosed dyslexia. We collected 16 Spanish texts written by dyslexic children from 13 to 15 years old. The texts are composed of homework writing exercises and were written by children who had Spanish as native language. The texts were all handwritten and we transcribed them manually. The words that we were not able to transcript due to the illegibility of the hand writing were marked. One example of a fragment of our texts is given in Figure 1.

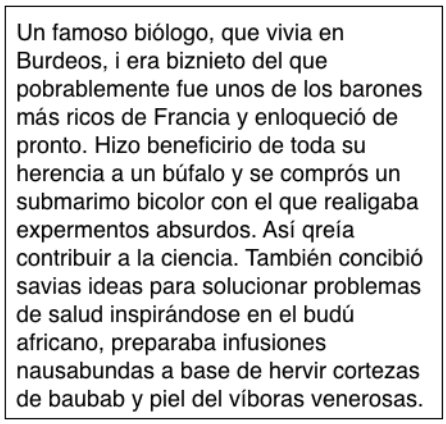
Un famoso biólogo, que vivia en Burdeos, i era biznieto del que pobrablemente fue unos de los barones más ricos de Francia y enloqueció de pronto. Hizo beneficirio de toda su herencia a un búfalo y se comprós un submarimo bicolor con el que realigaba expermentos absurdos. Así qreía contribuir a la ciencia. También concibió savias ideas para solucionar problemas de salud inspirándose en el budú africano, preparaba infusiones nausabundas a base de hervir cortezas de baubab y piel del víboras venerosas.

Figure 1: Example of one story of the texts written by a dyslexic child (14 years).

In the example in Figure 1[5] we have errors of all possible kinds, most of them simple: (i) substitution: **i (y), *realigaba (realizaba), *qreía (creía), *savias (sabias), *budú (vudú), *venerosas (venenosas)* and **baubab (baobab).*; (ii) insertion: **comprós (compró),* ; (iii) omission: **expermentos (experimentos), *unos (uno), *beneficirio (beneficiario), *nausabundas (nauseabundas)* and **del (de)*; and a double (iv) transposition **pobrablemente (probablemente)*.
We observe that there are errors that might not be attributed to dyslexia, for instance **i (y)* could be easily attributed as a transference from Catalan language (bilingual writer) and two others are concordance errors (**unos* and **del*). There is also one accentuation error: **vivia (vivía)*.
Since dyslexic errors overlap with other kind of errors found in documents, it is challenging to determine which errors are more likely to be only done by dyslexics. However, non-word multi-errors are more likely to be produced by a person with dyslexia (Baeza-Yates and Rello, 2011).

[5]Approximated literal translation: A famous biologist, who lived in Bordeaux, and was great-grandson of who probably was one of the wealthiest barons of France and suddenly went mad. He chose a buffalo as the beneficiary of his inheritance and bought a bicolor submarine in which he made absurd experiments. So he thought that he contributed to science. He also conceived wise ideas to solve health problems inspired by the African voodoo, preparing nauseating infusions based on boiled baobab barks and poisonous snakes.

The length average per text is 67 words and the total corpus size is 1,057 words. The reduced size of the corpus is explained by the difficulty of finding texts written by people diagnosed with dyslexia and the lack of a previous Spanish corpus of dyslexic errors. However, we believe that a corpus of this characteristics is valuable to analyze Spanish dyslexic errors and provide insight in where they appear or which is their distribution in Spanish. To the best of our knowledge, lists but not texts of dyslexic errors were used in previous work (Silva Rodríguez and Aragón Borja, 2000; Baeza-Yates and Rello, 2011).

7. Annotation of Dyslexic Errors

Following Pedler's annotation tag for errors, we marked-up all the errors in XML format. This kind of simple annotation gives the possibility, using regular expressions, to extract the errors and their corresponding target word from the corpus, as well as computing statistics.
We manually annotated the errors and added several tag attributes to typify each dyslexic error. Following we present the attributes and their possible values.

- Targ: the correct word(s).
- Type: this attribute refers to the error type depending on their edit distance. Its possible values are: "simple", "multi" and "boundary". Boundary specifies the case when one word is slit or two words are joined.
- Real: this attribute records if the error produced another real word. These errors are the most difficult to find automatically.
- First Letter: if the error is in the first letter or not.
- Edit Distance: The edit distance to the correct word(s).

Below we show an example for the error **pobrablemente (probablemente)* ('maybe').

```
<ERR targ = "probablemente"
type = "multi"
real = "no"
first_letter = "no"
ed = "2" >
pobrablemente </ERR>
```

In the case that there were two kind of errors we annotated as a multi-error, for instance, in **devidreo (de vidrio)* ('of glass') a boundary error is combined with a simple substitution error.
We did not annotate capitalization errors and accentuation errors since children among that age are still learning how to accentuate in Spanish. If the handwriting word was illegible an empty tag `<ILLEGIBLE/>` was added.

8. Comparing English and Spanish Errors

The corpora that we compare in this paper are in English and Spanish. These languages are archetypes of deep and shallow orthographies, respectively. Along 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).
In Tables 1 and 2 we compare the data of the corpus described in (Pedler, 2007) with our corpus. We compute the error ratio as the fraction of errors over the correctly spelt words we observe. As expected, Spanish dyslexics make less spelling errors (15%) than English dyslexics (20%) due to their different orthographies. On the other hand the percentage of unique errors is almost the same.

Category	English	Spanish
Total words	3,134	1,075
Total errors	636	157
Error ratio	0.20	0.15
Distinct errors	577	144
Percentage	90.7	91.7

Table 1: Error ratio and percentage in English and Spanish corpora of dyslexic errors.

Table 2 presents the distribution the different types of dyslexic errors for both corpus. To determine if an error was a real world error we checked its existence in the Royal Spanish Academy Dictionary (Real Academia Española, 2001).

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

Table 2: Error distribution in English an Spanish corpora of dyslexic errors.

As expected, there is a greater percentage of multi errors in a language with deep orthography as English than in Spanish, i.e. **qreía (creía)* ('thought'). However, the first letter errors are double in Spanish, i.e.: **tula (ruta)* ('way'). This is surprising according to (Yannakoudakis and Fawthrop, 1983) whose findings report that the first letter of a misspelling is correct in the majority of cases.
The rest of the dyslexic error types are similar in both languages. There are slightly more real word errors in Spanish, **dijo (digo)* ('said') or **llegada (llegaba)* ('said').
Simple errors are the most frequent ones in both languages. However, each error type has a different frequency. For instance, in our corpus substitution errors, **dertro (dentro)* ('in') are the most frequent ones (65% of the simple errors) while (Bustamante and Díaz, 2006) states that simple omissions are the most frequent kind.

9. Conclusions and Future Work

The comparisons presented in this works among different kind of dyslexic errors shed light on how dyslexia manifestations varies among languages and suggest that dyslexic accessible practices and tools are partially language dependent. This corpus is available for the research community.[6] Due to the difficulty of collecting texts of diagnosed dyslexics our Spanish corpus is still small but enough to present the distribution of the dyslexic errors and to settle the annotation criteria. In future work we plan to enlarge this corpus with more texts written by dyslexics and also using the Web as corpus. Also we plan to improve its annotation by separating the number of errors (simple or multi) from the case of happening at the boundaries of a word as simple and multi errors overlap with word boundary errors.

Acknowledgements

We deeply thank Yolanda Otal de la Torre for helping us to collect the Spanish texts written by dyslexics.

10. References

American Psychiatric Association. 2000. *Diagnostic and statistical manual of mental disorders: DSM-IV-TR*. American Psychiatric Publishing, Inc.

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

Nicola Brunswick. 2010. Unimpaired reading development and dyslexia across different languages. In Sine McDougall and Paul de Mornay Davies, editors, *Reading and dyslexia in different orthographies*, pages 131–154. Psychology Press, Hove.

F. R. Bustamante and E.L. Díaz. 2006. Spelling error patterns in spanish for word processing applications. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 93–98. ELRA.

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

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

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

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

N.E. Goulandris. 2003. *Dyslexia in different languages: Cross-linguistic comparisons*. Whurr Publishers.

P. Gregor, A. Dickinson, A. Macaffer, and P. Andreasen. 2003. Seeworda personal word processing environment for dyslexic computer users. *British Journal of Educational Technology*, 34(3):341–355.

D. Holbrook. 1964. English for the rejected: Training literacy in the lower streams of the secondary school.

Interagency Commission on Learning Disabilities. 1987. *Learning Disabilities: A Report to the U.S. Congress*. Government Printing Office, Washington DC, U.S.

[6] http://www.luzrello.com/Dyswebxia.html

International Dyslexia Association. 2011. Frequently Asked Questions About Dyslexia. http://www.interdys.org/.

S.K. Katusic, R.C. Colligan, W.J. Barbaresi, D.J. Schaid, and S.J. Jacobsen. 2001. Incidence of reading disability in a population-based birth cohort, 1976-1982, rochester, mn. *Mayo Clinic Proceedings*, 76(11):1081.

V. Levenshtein. 1965. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1:8–17.

S.A. Lindgrén and M. Laine. 2011. Multilingual dyslexia in university students: Reading and writing patterns in three languages. *Clinical Linguistics & Phonetics*, 25(9):753–766.

G.R. Lyon, S.E. Shaywitz, and B.A. Shaywitz. 2003. A definition of dyslexia. *Annals of Dyslexia*, 53(1):1–14.

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

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. 2005. DCDC2 is associated with reading disability and modulates neuronal development in the brain. *Proceedings of the National Academy of Sciences*, 102:17053–17058, November.

R. Mitton. 1996. *English spelling and the computer*. Longman Group.

L.C. Moats. 1996. Phonological spelling errors in the writing of dyslexic adolescents. *Reading and Writing*, 8(1):105–119.

R.I. Nicolson and A.J. Fawcett. 2011. Dyslexia, dysgraphia, procedural learning and the cerebellum. *Cortex*, 47(1):117–127.

Orton Dyslexia Society Research Committee. 1994. Definition of dyslexia. Former name of the International Dyslexia Association.

J. Pedler. 2007. *Computer Correction of Real-word Spelling Errors in Dyslexic Text*. Ph.D. thesis, Birkbeck College, London University.

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

Real Academia Española. 2001. *Diccionario de la lengua española*. Espasa-Calpe, Madrid, 22 edition.

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

C. Romani, J. Ward, and A. Olson. 1999. Developmental surface dysgraphia: What is the underlying cognitive impairment? *The Quarterly Journal of Experimental Psychology*, 52(1):97–128.

G. Schulte-Körne, W. Deimel, K. Müller, C. Gutenbrunner, and H. Remschmidt. 1996. Familial aggregation of spelling disability. *Journal of Child Psychology and Psychiatry*, 37(7):817–822.

P.H.K. Seymour, M. Aro, and J.M. Erskine. 2003. Foundation literacy acquisition in european orthographies. *British Journal of psychology*, 94(2):143–174.

A. Silva Rodríguez and L.E. Aragón Borja. 2000. Análisis cualitativo de un instrumento para detectar errores de tipo disléxico (IDETID-LEA). *Psicothema*, 12(2):35–38.

R. Spooner. 1998. *A spelling aid for dyslexic writers*. Ph.D. thesis, PhD thesis, University of York.

C. Sterling, M. Farmer, B. Riddick, S. Morgan, and C. Matthews. 1998. Adult dyslexic writing. *Dyslexia*, 4(1):1–15.

R. Treiman. 1997. Spelling in normal children and dyslexics. *Foundations of reading acquisition and dyslexia: Implications for early intervention*, pages 191–218.

F.R. Vellutino, J.M. Fletcher, M.J. Snowling, and D.M. Scanlon. 2004. Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of child psychology and psychiatry*, 45(1):2–40.

World Health Organization. 1993. *International statistical classification of diseases, injuries and causes of death (ICD-10)*. World Health Organization, tenth edition.

E.J. Yannakoudakis and D. Fawthrop. 1983. The rules of spelling errors. *Information Processing & Management*, 19(2):87–99.