DysList: An Annotated Resource of Dyslexic Errors

Luz Rello,[1] Ricardo Baeza-Yates,[2] and Joaquim Llisterri[3]
[1] Natural Language Processing and Web Research Groups, Universitat Pompeu Fabra, Barcelona
[2] Web Research Group, Universitat Pompeu Fabra & Yahoo Labs, Barcelona
[3] Departament de Filologia Espanyola, Universitat Autònoma de Barcelona
{luzrello,rbaeza}@acm.org, Joaquim.Llisterri@uab.cat

Abstract

We introduce a language resource for Spanish, *DysList*, composed of a list of unique errors extracted from a collection of texts written by people with dyslexia. Each of the errors was annotated with a set of characteristics as well as visual and phonetic features. To the best of our knowledge this is the largest resource of this kind, especially given the difficulty of finding texts written by people with dyslexia.

Keywords: Errors, Dyslexia, Visual, Phonetics, Resource

1. Introduction

Dyslexia is a reading and spelling disorder with neurological origin (World Health Organization, 1993; American Psychiatric Association, 2000).[1] It is characterized by difficulties with accurate and 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 comparison to other cognitive abilities (Lyon et al., 2003). Secondary consequences may include problems in reading comprehension and reduced reading experience that can impede growth of vocabulary and background knowledge (Orton Dyslexia Society Research Committee, 1994). Although dyslexia is universal, its prevalence varies depending on the language, from 10-17.5% of the population of the USA (Interagency Commission on Learning Disabilities, 1987) to 7.5-11% of the Spanish speaking population (Carrillo et al., 2011).

The errors that people with dyslexia write are very valuable and have been used for various purposes as shown in the next section, ranging from diagnosing dyslexia to software applications targeted to people with dyslexia. However, the existence of resources, such as corpora or lists of dyslexic errors are scarce. Therefore, in this paper we present the first list of Spanish dyslexic errors which has been annotated with different types of linguistic information. This resource is valuable and can serve as a basis to develop more tools to help this target group.

The rest of this paper is organized as follows. In the next section we show how dyslexic errors have been used for different purposes. In Section 3 we cover related work while in Section 4 we show how we extracted the errors. In Section 5 we present the annotation criteria used for *DysList* and in Section 6 we present the characteristics of the resource. Some conclusions and future work are drawn in Section 7.

2. Dyslexic Errors as a Source of Knowledge

In general terms, errors can be used as a source of knowledge. 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 comprehensibility of web content (Rello and Baeza-Yates, 2012a).

Since the kinds of errors that people with dyslexia make are related to the types of difficulties that they have (Sterling et al., 1998), their written errors have been used for various purposes such as (1) studying dyslexia, (2) diagnosing dyslexia, or (3) for accessibility related purposes.

First, the analyses of writing errors made by people with dyslexia were used in previous literature to study different aspects of dyslexia (Connelly et al., 2006; Silva Rodríguez and Aragón Borja, 2000). For instance, the specific types of dyslexic errors highlight different aspects of dyslexia (Treiman, 1997), such as the phonological processing deficit (Moats, 1996; Lindgrén and Laine, 2011). The dyslexic error rates vary depending on the language writing system (Lindgrén and Laine, 2011). However, compared to non-dyslexics, people with dyslexia present more errors attributable to phonological impairment, spelling knowledge, and lexical mistakes (Sterling et al., 1998). Even if dyslexia is popularly identified by the letter reversals, according to (Meng et al., 2005) only 30% of people with dyslexia have trouble with reversing letters and numbers.

Second, since people with dyslexia exhibit higher spelling error rates than non-dyslexic people (Coleman et al., 2009), there are diagnoses of dyslexia based on the spelling score (Schulte-Körne et al., 1996; Toro and Cervera, 1984). Also, the spelling error rate is being used as a diagnosing factor in the current official Catalonian protocols (Speech Therapy Association of Catalonia, 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 (Korhonen, 2008; Li et al., 2013; Pedler, 2007), text prediction software,[2] 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).

[1] In some literature, dyslexia is referred to as a specific reading disability only (Vellutino et al., 2004) and dysgraphia as its writing manifestation (Romani et al., 1999).

[2] *Penfriend XL* (`http://www.penfriend.biz/`).

3. Related Work

To the best of our knowledge, there is only one corpus of dyslexic texts in English, the corpus used by Pedler (Pedler, 2007) for the creation of a spell checker of real-word errors made by people with dyslexia. This corpus has 3,134 words and 363 errors (Pedler, 2007). It is composed 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 the spellchecker, that initial corpus was enlarged to 21,524 words containing 2,654 errors, with over 800 real-word errors. The additional sources for that corpus were: texts from a student with dyslexia, texts from an online typing experiment (Spooner, 1998), samples from dyslexic bulletin boards and mailing lists, and stories written by children with dyslexia.

For Spanish, *DysCorpus* is composed of texts written by children with dyslexia, containing 16 texts (1,057 words) with 157 unique errors. The texts are school essays from children with dyslexia between 6 and 15 years old. In (Rello et al., 2012a) we described the this corpus, comparing the frequency and the types of errors with Pedler's corpus (Pedler, 2007).

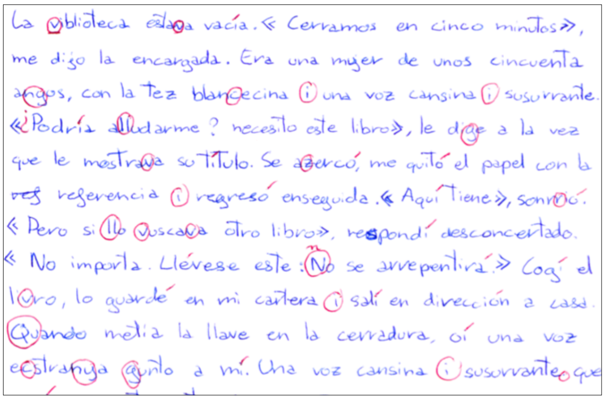
Regarding lists of dyslexic errors, the only similar resource is the list of English confusion sets compiled by Pedler (Pedler, 2007),[3] extracted from the corpus of text written by people with dyslexia mentioned before (Pedler, 2007). This list is composed of 833 confusion sets. A confusion set is a small group of words that are likely to be confused with one another, such as *weather* and *whether*.

4. Extracting Errors from Dyslexic Texts

Manifestations of dyslexia varies among languages (Goulandris, 2003) but also among subjects and ages (Vellutino et al., 2004). For instance, misspelling rate in dyslexic children is higher than in adults (Sterling et al., 1998). However, experiments evidence that adult with dyslexia have a continuing problem in the lexical domain, manifested in a poor spelling ability (Sterling et al., 1998). Hence, we collected texts written by a similar population in terms of age, education, native language (Spanish), and that have been diagnosed with dyslexia. These 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.

We used a total of 83 texts composed of (a) 54 school essays and homework exercises provided by teachers from children and teenagers with dyslexia between 6 and 15 years old (Figure 1), and (b) 29 texts provided by parents with children with dyslexia. The school essays include the ones from *DysCorpus* (Rello et al., 2012a).

From our text collection we manually extracted a list of 887 misspelled words, without taking into account illegible handwritten words. We did not extracted capitalization or accentuation errors since most children among that age are still learning how to capitalize and accentuate in Spanish. This list has 678 different target words, where *sigilosamente (‘stealthily’)* is the word with more misspelling variants (7). From this set of words we extracted 894 different correct-misspelled pairs with a total of 1,171 errors. For instance, the words *accesibilidad (‘accessibility’)* and *sigilosamente* are the ones that have more different errors (12). That is, there is more than one way to correct the mistake.



La viblioteca estava vacía. « Cerramos en cinco minutos»,
me dijo la encargada. Era una mujer de unos cincuenta
angos, con la tez blancecina i una voz cansina i susurrante.
«¿Podría alludarme ? necesito este libro», le dige a la vez
que le mostrava su titulo. Se azerco, me quito el papel con la
~~vez~~ referencia i regreso enseguida. «Aqui tiene», sonrrio.
« Pero si llo vuscava otro libro», respondi desconcertado.
« No importa. Llévese este: No se arrepentira.» Cogi el
livro, lo guarde en mi cartera i sali en dirección a casa.
Quando metia la llave en la cerradura, oi una voz
ecstranya qunto a mi. Una voz cansina i susurrante que

Figure 1: Example of a (corrected) handwritten text of a teenager with dyslexia (15 years old).

5. Annotation of Dyslexic Errors

We annotated each of the word-error pairs to create *DysList* with the following information:

- **Target word**: the intended word the person aimed to write.
- **Misspelled word**: the wrongly written word or tokens.[4]
- **Damerau-Levenshtein distance**: the minimum number of edits (insertion, deletion, substitution, transposition) required to change the misspelled error into the (target) correct word (Damerau, 1964; Levenshtein, 1965).
- **Target and misspelled word frequencies**: defined as the number of hit counts in a major search engine for Web pages written in Spanish.
- **Target and misspelled length**: number of characters.
- **Error position**: the position in the target word where the error occurs.
- **Target word syllables**: number of syllables.
- **Target syllable**: the structure of the syllable where the error occurs, such as CV, CVC, or CCV.
- **Type of error**: a detail analyses of the different kind of dyslexic errors is given in (Rello and Baeza-Yates, 2012b).

[3] http://www.dcs.bbk.ac.uk/~jenny/resources.html

[4] We use ‘*’ to denote this word in the examples given.

Visual Feature	Values	Letter(s)
Mirror (digital)	**V** = vertical, **H** = horizontal, **B** = both, **N** = none	**H** = <n, u>, **B** = <b, d, p, q>
Mirror (handwriting)	**V** = vertical, **H** = horizontal, **B** = both, **N** = none	**Y** = <g, h, m, n, u, v, w, y>, **B** = <b, d, p, q>
Box (digital)	**U** = upper, **L** = lower, **B** = both, **N** = none	**U** = <b, d, f, h, k, l, t>, **L**= <g, j, p, q, y>
Box (handwriting)	**U** = upper, **L** = lower, **B** = both, **N** = none	**U** = <b, d, h, k, l, t>, **L**= <g, j, p, q, y, z>, **B**= <f>
Line (digital)	**V** = vertical, **H** = horizontal, **B** = both, **N** = none	**H** = <a, e, f, s>, **V** = <m, w>, **B** = <k>
Line (handwriting)	**V** = vertical, **H** = horizontal, **B** = both, **N** = none	**H** = <k, z>, **V** = <m, w>
Rotation (digital)	**Y** = yes, **N** = no	**Y** = <a, e, d, b, p, q, n, u>
Rotation (handwriting)	**Y** = yes, **N** = no	**Y** = <a, b, d, e, h, m, n, p, q, u, w, y>
Fuzzy (digital)	**Y** = yes, **N** = no	**Y** = <b, c, d, f, g, i, j, l, n, ñ, o, p, q, t, u, v>
Fuzzy (handwriting)	**Y** = yes, **N** = no	**Y** = <b, d, g, h, m, n, ñ, p, q, s, r, u, v, w, y, z>

Table 1: Visual features of the annotated target and error letters.

S **Substitution**: change one letter for another, for example **reelly (really)*.

I **Insertion**: insert one letter, like in **situartion (situation)*. A word that has been split in two different tokens is counted as an insertion, like in **sub marine (submarine)*.

D **Deletion**: omit one letter, as in **approch (approach)*. Run-on word boundary errors, like in **alot (a lot)*, are counted as one deletion.[5]

T **Transposition**: reversing the order of two adjacent letters, for example **artcile (article)*.

- **Real word**: this Boolean attribute records if the error produced another real word. For instance, *witch* being *which* the intended word.

- **Visual information**: each of the target and the error graphemes we annotate the letters involved in the error with the following visual information, considering both, handwritten text and digital typography (*sans serif*). See Table 5.

  - **Mirror letter (handwriting/digital)** such as <d> and <b> or <m> and <w>, with three possible values: vertical, horizontal, and none.[6]
  - **Box (handwriting/digital)**: lower box (*e.g.* <p, q>, or <g>), upper box (*e.g.* <t>, or <b>), both (*e.g.* <f>), and none (*e.g.* <n, m>, or <s>).
  - **Line (handwriting/digital)**: vertical (*e.g.* <m>), horizontal (*e.g.* <e>), and none (*e.g.* <o>).
  - **Rotation (handwriting/digital)**: boolean attribute that indicates if the rotation of a letter produces another letter, such as <d> and <p>.
  - **Fuzzy letters (handwriting/digital)**: boolean attribute that indicates if the letter have similar visual letters (not due to rotate or mirror) such as, such as <s> and <z>.

- **Phonetic information**: each of the target and the error phones associated to the graphemes in the text are annotated using traditional articulatory phonetic features (International Phonetic Association, 1999):

  - **Phone type**: vowel (*e.g.* [a]) or consonant (*e.g.* [p]); combinations of vowels forming a diphthong (*e.g.* [i̯a]) and consonant clusters in syllabic onsets (*e.g.* [pl]) have also been annotated as specific phone types.
  - For consonants:
    - ∗ **Voicing**: voiced (*e.g.* [b]) or voiceless (*e.g.* [p]).
    - ∗ **Manner of articulation**: plosive (*e.g.* [p]), nasal (*e.g.* [m]), trill (*e.g.* [r]), tap or flap (*e.g.* [ɾ]), fricative (*e.g.* [f]), lateral (*e.g.* [l]), approximant (*e.g.* [β̞]), and affricate (*e.g.* [tʃ]).
    - ∗ **Place of articulation**: bilabial (*e.g.* [p]), labiodental (*e.g.* [f]), interdental (*e.g.* [θ̟]), dental (*e.g.* d̪]), alveolar (*e.g.* [s]), palatal (*e.g.* [tʃ]), and velar (*e.g.* [k]).
  - For vowels:
    - ∗ **Height**: open (*e.g.* [a]), mid (*e.g.* [e]), and close (*e.g.* [i]).
    - ∗ **Place of articulation**: front (*e.g.* [i]), central (*e.g.* [a]), and back (*e.g.* [u]).
    - ∗ **Lip rounding**: rounded (*e.g.* [u]) or unrounded (*e.g.* [i]).

- **Language transfer**: some of the errors in the list were due to transference from Catalan to Spanish.[7] Hence we tagged the error caused by transference from Catalan. For instance, **accessiblidad (accesiblilidad)* may be due to the existence of the word *accessibilitat* in Catalan.

6. DysList Characteristics

Frequency: The target word web frequency ranged from 190, *arbolazo, ('big tree')*, to 1,389,717,667 *en ('in')*. The errors words frequency ranged from 0, *aczecibilidad (accesibilidad, 'accessibility')*, to 1,178,165,310 in the real word error **ha (a, 'to')*. On average correct words were 4.63 more frequent than words with errors.

[5] Notice that a deletion in the target word is an insertion in the misspelled word and vice versa.

[6] Allophones are marked with '[]' and graphemes with '<>'.

[7] Most of the texts come from Catalan schools where the rate of bilingual students (Catalan-Spanish) is high.

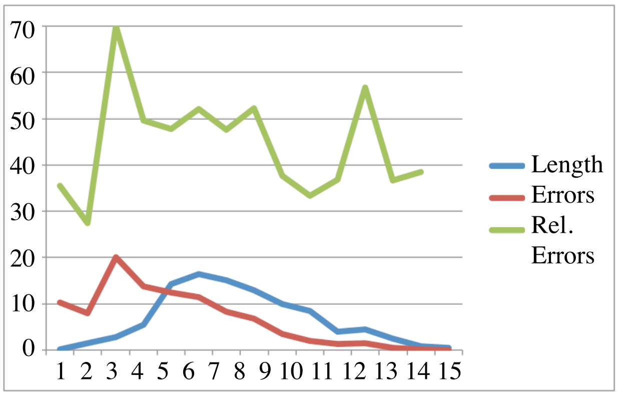


Figure 2: Percentage distribution of *DysList* word lengths, error positions and relative percentage of errors in the position.

Syllable Type	Percentage	No. Syllables	Percentage
CV	37.40	3	33.30
CVC	21.35	2	26.30
none	13.15	4	17.68
CCV	8.20	1	11.87
CVV	7.77	5	7.51
CVVC	6.06	6	3.25
VC	3.67	7	0.09
CCVC	1.54		
V	0.60		
VV	0.09		
CCVV	0.09		
CCVCC	0.09		

Table 2: Distribution of syllable types and errors. None refers to the boundary errors such as **a drede (adrede), 'in purpose'*.

Length and error position: The length of the target words range from 1 to 20, with the mode at length 6 and an average length of 7.47 letters. Figure 2 gives the percentage distribution of target word lengths, the percentage distribution of the word positions where the errors appear, and the relative percentage of errors in the position (that is, 100 times the number of errors in that position divided by the total number of words that have that position).

Syllables: The target words range from one to seven and we observed eleven types of syllables with the distributions shown in Table 2.

Damerau-Levenshtein distance In most cases the distance is just 1 (73.3%), with 21.6% of the cases at distance 2 and only 5.1% at distance 3 or greater.

Type of error: In Table 3 we give the percentages of every error type. As we can notice, substitution errors are the most frequent ones (near 60%) while (Bustamante and Díaz, 2006) states that simple omissions (deletions) are the most frequent kind for Spanish. Even if dyslexia is popularly known for the transposition errors, less than 1% of the errors where of this type. This is consistent with (Meng et al., 2005) which states that only 30% of dyslexics have trouble with reversing letters and numbers.

In our analysis we consider some special phonetic errors coming from double letters that have a single sound in Spanish (such as <ll> and <rr>). We found 229 different errors (without considering the phonetics there are only 186 different errors). The most frequent errors (down to 2%) are shown in Table 4. These nine errors represent more than 40% of all errors found, showing the extreme bias of them (*i.e.* less than 4% of the unique errors cover more than 40% of the cases). The most frequent case produces more than 11% of the errors and involve two letters that in Spanish have the same pronunciation, <b> and <v> (which is not the case in English). Analyzing this and other frequent cases, we found three big groups of errors:

1. Inserting or deleting a consonant represent 37.9% of the errors, excluding <h> and <y>, which are included in the next cases.

2. Deleting or inserting a vowel, including <y> that can have the same phonetic values as <i> in certain contexts, represent 37.5% of the errors.

3. Substituting two letters of similar sound (*e.g.* <g> and <j>) or deleting/inserting an <h>, a letter that in Spanish most of the time has no sound, represent 15.4% of the errors.

Notice that these three groups of errors cover more than 80% of the errors.

We also studied the position of the errors without finding any important preference, although most errors occur inside the target word. The four most frequent cases were inserting an <h> at the beginning of the word (3.7%), substituting <b> by <v> at the first (2.8%) or third (2.1%) positions, and inserting an <e> in the second position (2.8%). Finally, only 8.97% of the errors were real word errors.

Visual Analysis: To access the analyses of the visual features we used Chi-Square goodness of fit to establish

Error Type	Percentage
Substitution	58.84
Insertion	13.40
Deletion	26.30
Transposition	1.45

Table 3: Percentages of dyslexic error types.

Error Type	Letter(s)	Percentage
S	<b, v>	11.36
D	*space*	6.75
S	<g, j>	5.46
D	<h>	4.53
I	*space*	3.07
S	<c, z>	2.82
S	<c, s>	2.22
D	<r>	2.22
I	<r>	2.13

Table 4: Percentages of frequent specific errors.

Visual Feature	Correct Letters (%)	Error Letters (%)
Mirror (digital)	**none** = 26.81, **N** = 57.90, **H** = 3.93, **B** = 11.36	**none** = 33.39, **N** = 54.74, **H** = 4.01, **B** = 7.86
Mirror (hand)	**none** = 26.81, **N** = 47.65, **H** = 14.18, **B** = 11.36	**none** = 33.39, **N** = 39.28, **H** = 19.47, **B** = 7.86
Box (digital)	**none** = 26.81, **U** = 19.04, **N** = 43.81, **L** = 10.33,	**none** = 33.39, **U** = 11.44, **N** = 44.41, **L** = 10.76
Box (hand)	**none** = 26.81, **U** = 18.53, **N** = 42.70, **L** = 11.44, **B** = 0.51	**none** = 33.39, **U** = 11.44, **N** = 41.33, **L** = 13.83
Line (digital)	**none** = 33.39, **V** = 0.85, **N** = 58.67, **H** = 13.66	**none** = 26.81, **V** = 1.11, **N** = 54.48, **H** = 11.02
Line (hand)	**none** = 26.81, **V** = 0.85, **N** = 71.22, **H** = 1.11	**none** = 33.39, **V** = 1.11, **N** = 62.43, **H** = 3.07
Rotation (dig.)	**none** = 26.81, **Y** = 22.63, **N** = 50.56	**none** = 33.39, **Y** = 18.19, **N** = 48.42
Rotation (hand)	**none** = 26.81, **Y** = 30.57, **N** = 42.61	**none** = 33.39, **Y** = 20.58, **N** = 46.03
Fuzzy (digital)	**none** = 26.81, **Y** = 44.41, **N** = 28.78	**none** = 33.39, **Y** = 43.47, **N** = 23.14
Fuzzy (hand)	**none** = 26.81, **Y** = 44.66, **N** = 28.52	**none** = 33.39, **Y** = 41.59, **N** = 25.02

Table 5: Visual features of the annotated target and error letters.

whether or not an observed frequency distribution (in the error letters) differs from a theoretical distribution (the one of the correct letters). The percentage of error letters differ from the correct letters by digital visual features ($\chi^2(9) = 97.67, p < 0.001$) as well as handwriting visual features ($\chi^2(9) = 377.59, p < 0.001$). The percentage of correct fuzzy letters differ from the percentage of error fuzzy letters taking into account both, digital ($\chi^2(4) = 76.36, p < 0.001$) and handwriting typographies ($\chi^2(4) = 41.10, p < 0.001$). The percentage of error letters differ from the correct letters by box visual features ($\chi^2(9) = 324.56, p < 0.001$) as well as handwriting visual features ($\chi^2(12) = 244.13, p < 0.001$). The percentage of error letters differ from the correct letters by line visual features ($\chi^2(9) = 73.29, p < 0.001$) as well as handwriting visual features ($\chi^2(9) = 34.21, p < 0.001$). The percentage of error letters differ from the correct letters by rotation visual features ($\chi^2(4) = 23.13, p < 0.001$) as well as handwriting visual features ($\chi^2(4) = 32.59, p < 0.001$).

Phonetic Analysis: Vowel substitutions account for 5.38% (*N* = 63) of the total number of errors in the corpus. After the transcription of the vowel graphemes according to their phonetic realization in Spanish, the percentage of substitutions errors concerning single vowels has been computed, as shown in Table 6.

	Error							**Total**
	a	e	i	i̯	o	u	u̯	(%)
Target								
a	0	20.63	3.17	0	9.52	0	0	33.33
e	15.87	0	4.76	1.59	6.35	0	0	28.57
i	0	7.94	6.35	0	0	0	0	14.29
i̯	0	0	0	6.35	0	0	0	6.35
o	3.17	4.76	0	0	0	1.59	1.59	11.11
u	0	0	1.59	0	3.17	0	0	4.76
u̯	0	0	0	1.59	0	0	0	1.59
Total (%)	19.05	33.33	15.87	9.52	19.05	1.59	1.59	

Table 6: Percentage of vowel substitutions.

The analysis of the phonetic features associated to each vowel shows that the most frequent substitution errors involve target unrounded vowels ([i], [e], [a]) when lip rounding is considered, target mid vowels ([e], [o]) in errors related to vowel height, and target front vowels ([i], [e]) when place of articulation is studied.

In terms of shared features, the most frequent types of substitution errors involve one phonetic feature, lip rounding being the most frequent one. It is interesting to note that only 15.87% (*N* = 10) of the vowel substitution errors correspond to phones that do not have any feature in common. Errors occur most frequently in unrounded vowels ([i], [e], [a]) as far as lip rounding is concerned, mid vowels ([e], [o]) if the degree of opening is considered and front vowels ([i], [e]) when place of articulation is taken into account. This confirms the findings in (Rello and Llisterri, 2012) in a smaller sample.

The pattern arising from the study of the phonetic features involved in substitution errors is consistent with the most frequent substitutions found in the corpus (Table 6):

[a] ([unrounded]) → [e] ([unrounded] [mid] [front])

[e] ([unrounded] [mid] [front]) → [a] ([unrounded])

[i] ([unrounded] [front]) → [e] ([unrounded] [mid] [front])

[o] ([mid]) → [e] ([unrounded] [mid] [front])

Substitutions in vowel combinations forming a diphthong account for the 0.94% (*N* = 11) of the errors found in the corpus. The most frequent errors in this category –2 cases of each in the corpus– are found in the substitution of [i̯a] by [ea] and of [i̯o] by [eo]. The highest proportion of errors is observed in target [i̯a] and [o̯e] combinations. In terms of the result of the substitutions, [ea] and [eo] are the two most frequent errors. Given the small size of the sample, no further analysis have been performed, but the trend is coherent with the prevalence of substitutions involving [e] and [a] described for vowels.

Substitution errors in single consonants correspond to the 46.37% (*N* = 543) of the total number of errors in the corpus. They represent, then, the largest category of errors present in *DysList* and are summarized in Table 7.

It can be observed that the most frequent errors in consonants are related to the cases in which a one-to-one correspondence between graphemes and phones is not maintained. This results in two different graphemes having the same phonetic value:

<b> and <v>: both realized as a bilabial plosive [b] or a bilabial approximant [β̞] according to the phonetic context.

	Error																						Total
	b	β̞	d	ð̞	f	g	ɣ̞	j	ɲ	k	ks	l	λ	m	n	p	ɾ	rr	s	t	θ	x	(%)
Target																							
b	7.73	0	0.37	0	0	0	0	0	0	0	0	0	0	0.18	0	0.18	0	0	0	0	0	0	8.47
β̞	0.18	16.57	0	0.74	0	0	0	0	0	0	0	0.37	0	0	0.18	0	0.18	0	0	0	0	0	18.23
d	0.37	0	0	0	0	0.18	0	0	0	0	0	0	0	0	0	0	0	0	0	0.18	0	0	0.74
ð̞	0	0.74	0	0	0	0	0	0	0	0.18	0	0	0	0	0	0	0.74	0	0	0.37	0	0	2.03
f	0	0	0	0	0	0	0	0	0	0	0	0	0	0.18	0	0	0	0	0	0	0.92	0	1.10
g	0	0	0	0	0	0	0	0	0	0.55	0	0	0	0	0	0	0	0	0	0.18	0	0.37	1.10
ɣ̞	0	0.55	0	0	0	0	0.37	0	0	0.18	0	0.18	0	0	0.18	0	0	0	0	0	0	0.55	2.03
j	0	0	0	0	0	0	0.18	0.92	0	0	0	0	2.03	0	0	0	0	0	0	0	0	0	3.13
ɲ	0	0	0	0	0	0	0.37	0.18	2.58	0	0	0	0.18	0	0.37	0	0	0	0	0	0	0.18	3.87
ʝ	0	0	0	0	0	0	0	0	0	0	0	0	0.55	0	0	0	0	0	0	0	0	0.37	0.92
k	0	0	0	0	0	0.18	0.18	0	0	3.13	0	0	0	0	0	0.55	0.18	0	0.18	0.18	2.39	0.18	7.18
l	0	0.18	0	0	0	0	0	0	0	0	0	0	0	0	0.18	0.18	0.37	0	0.18	0.37	0	0	1.47
λ	0	0	0	0	0	0	0	0.37	0	0	0	0	0	0	0	0	0.18	0	0	0	0	0	0.55
m	0	0	0	0	0	0	0	0	0	0	0	0	0	1.66	0.37	0	0.18	0	0	0	0	0	2.21
n	0	0	0	0	0	0	0	0	0	0	0	0	0	1.10	0	0	1.10	0	0.55	0	0	0	2.76
ŋ̞	0	0	0	0	0	0	0	0	0	0	0	0	0	0.18	0	0	0	0	0	0	0	0	0.18
p	0.18	0	0	0	0	0	0	0	0	0.18	0	0.18	0	0	0	0	0.18	0	0	0	0	0	0.74
ɾ	0	0.37	0	0	0	0	0	0	0	0	0	0.37	0	0	0.37	0.37	0	0.74	0.55	0.18	0	0	2.95
r	0	0	0	0	0	0	0	0	0	0	0	0	0.37	0	0	0	1.66	1.10	0	0.18	0.18	0	3.50
s	0	0	0	0	0	0	0	0	0	0.74	0	0.37	0	0	0.37	0	0.37	0.18	1.47	0	4.42	0.37	8.29
t	0.18	0	0.18	0.18	0	0.18	0	0	0	0	0	0.37	0	0	0	0	0	0	0	0	0	0	1.10
θ	0	0	0.37	0.55	0.37	0	0.18	0	0	1.29	0	0	0	0	0	0.18	0.55	0.18	3.31	0.18	4.60	0.37	12.15
tʃ	0	0	0	0.18	0	0.18	0.18	0	0	0	0	0	0	0	0	0	0	0	0	0	0.18	0.18	0.92
x	0	0	0	0	0	1.84	3.50	0	0	0.18	0.18	0	0	0	0	0	0.18	0	0.18	0	0.37	7.92	14.36
Total (%)	8.66	18.42	0.92	1.66	0.37	2.58	4.97	1.47	2.58	6.45	0.18	1.84	3.13	3.31	2.03	1.47	5.89	2.21	6.45	1.84	13.08	10.50	

Table 7: Percentage of consonant substitutions.

<j> followed by <a>, <o> or <u> and <g> followed by <e>, <i>: both realized as a velar fricative [x].

<z> followed by <a>, <o> or <u> and <c> followed by <e> or <i>: both realized as an interdental fricative [θ].

<c> followed by <a>, <o> or <u> and <qu> followed by <i> or <e>: both are realized as a velar plosive [k].

<r> in word-initial position and after nasals or lateral consonants or <s> and <rr> between vowels: both are realized as an alveolar trill [r].

This is the reason of the high percentage of errors in target consonants [β̞] (18.23%), <x> (14.36%), <θ> (12.15%) and [k] (7.18%) and also in the consonants resulting from a substitution error: [β̞] (18.42%), [θ] (13.08%) [x] (10.50%) and [k] (6.45%) (Table 7). The lack of biunivocal correspondence between phones and graphemes is also patent in the most frequent confusions in manner of articulation within the class of fricative consonants (24.68%) –to which [x] and [θ] belong–, within the group of approximant consonants (20.07%) –[β̞]– and within plosive consonants (14.55%) –[k]–. Taps and trills are also involved as target phones or as errors, although to a lesser extent. The same trend is observed when place of articulation is considered: the largest number of confusions occur within the class of bilabials (26.70%) –which includes [β̞]– and inside the group of velars (19.15%) –which includes [x] and [k]. The interdental consonant [θ] appears as the result of substitution errors in 13.08% of cases and as target phones in confusions in 11.97% of cases.

Confusions between [s] and [θ] (4.42%) and between [θ] and [s] (3.31%) observed in Table 7might be in part explained by the geolectal phenomenon known as *seseo*, which consists in the systematic substitution of [θ] (interdental fricative) by [s] (alveolar fricative) so that [θ] is absent from the phonetic inventory of the speakers of the geographic areas in which this phenomenon occurs. The analysis of features of manner and place also point out in this direction if the confusions in the class of fricatives and in alveolar and interdental consonants are considered.

The presence of a 3.13% of cases in which [λ] appears as the result of a confusion error and the confusions between [j] and [λ] (2.03%) shown in Table 7 might be partially accounted for by the presence of *yeísmo*, i.e., a neutralization of the contrast between [j] (palatal approximant) and [λ] (palatal lateral) in favor of [j] which is common in most geographical varieties of Spanish. When substitutions in manner of articulation are considered, 2.58% of cases of confusions between laterals and approximants are found; part of the substitutions within the class of palatals (6.63%) may be also accounted for by the presence of *yeísmo*.

The 2.58% of confusions in [ɲ] (palatal nasal) that appear in Table 7 may be explained by the decision taken for the phonetic transcription of the corpus concerning a potential transfer from Catalan spelling rules. Since [ɲ] is spelled as <ñ> in Spanish and as <ny> in Catalan, it was considered that both <ñ> and <ny> were intended to represent the palatal nasal consonant.

Almost half of the substitutions found in consonants occur between phones that share their three features (48.43%), while confusions between consonants sharing one (19.52%) or two (26.15%) features are less commonly encountered. It is worth noting that confusions between consonants that do no have any phonetic feature in common take place in 5.52% of cases.

Finally, half of the consonant confusions in the corpus affect simultaneously voicing, manner and place features, a fact to be explained by the spelling irregularities mentioned earlier. When two features are involved in confusions, manner and place are simultaneously affected in 16.99% of cases and voicing and place in 9.77% of cases. If the confusion involves only one feature, it can be either place of articulation (9.96%) or voicing (9.57%).

In summary, the analysis of consonant substitutions reveals that the spelling mistakes in cases of lack of one-to-one correspondence between phones and graphemes are an important source of confusions within the same class of consonants and are phonetically motivated.

Substitutions affecting combinations of consonants represent a 0.60% (N = 7) of the total number of errors in the corpus. More than half of the errors within this category –4 cases– correspond to the target sound [ks], spelled as <x> in Spanish. The rest of the errors are found in heterosyllabic clusters formed by a plosive (or their approximant realizations) plus a liquid (i.e., a lateral or a rhotic consonant). No further phonetic analysis has been carried out due to the small size of the sample.

7. Conclusions and Future Work

Our Spanish list of dyslexic errors is still small but large enough to find some insights about dyslexic errors and to settle the annotation criteria. In fact, we believe that this collection is valuable if it allows the creation of more tools targeted to people with dyslexia. With respect to Pedler's confusion sets mentioned in Section 3, we believe our resource is of similar size and possibly more diverse as it includes a larger sample of the population. In future work we plan to enlarge our collection with more texts written by people with dyslexia and also using the Web as corpus. *DysList* resource is freely available in `www.luzrello.com/dyslist.html`

Acknowledgements

We thank Martí Mayo for computing some of the features. This work has been partially funded by the HIPERGRAPH project (TIN2009-14560-C03-01) from the Spanish Economy and Competitiveness Ministry.

8. References

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

Bustamante, F. R. and Díaz, E. (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.

Carrillo, M. S., Alegría, J., Miranda, P., and Pérez, S. (2011). Evaluación de la dislexia en la escuela primaria: Prevalencia en español (Evaluation of dyslexia in primary school: The prevalence in Spanish). *Escritos de Psicología (Psychology Writings)*, 4(2):35–44.

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

Connelly, V., Campbell, S., MacLean, M., and Barnes, J. (2006). Contribution of lower order skills to the written composition of college students with and without dyslexia. *Developmental neuropsychology*, 29(1):175–196.

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

Gelman, I. A. and Barletta, A. L. (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.

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

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

Holbrook, D. (1964). *English for the Rejected: Training Literacy in the Lower Streams of the Secondary School.* Cambridge University Press, New York, US.

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

International Phonetic Association. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Korhonen, T. (2008). Adaptive spell checker for dyslexic writers. In *Proceedings of the 11th international conference on Computers Helping People with Special Needs*, ICCHP '08, pages 733–741, Berlin, Heidelberg. Springer-Verlag.

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

Li, A. Q., Sbattella, L., and Tedesco, R. (2013). Polispell: an adaptive spellchecker and predictor for people with dyslexia. In *User Modeling, Adaptation, and Personalization*, pages 302–309. Springer.

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

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

Meng, H., Smith, S., Hager, K., Held, M., Liu, J., Olson, R., Pennington, B., DeFries, J., Gelernter, J., O'Reilly-Pol, T., Somlo, S., Skudlarski, P., Shaywitz, S., Shaywitz, B., Marchione, K., Wang, Y., Murugan, P., LoTurco, J., Grier, P., and Gruen, J. (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.

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

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

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

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

Piskorski, J., Sydow, M., and Weiss, D. (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 Press.

Rello, L. and Baeza-Yates, R. (2012a). Lexical quality as a proxy for web text understandability (poster). In *Proc. WWW '12*, pages 591–592, Lyon, France.

Rello, L. and Baeza-Yates, R. (2012b). The presence of English and Spanish dyslexia in the Web. *New Review of Hypermedia and Multimedia*, 8:131–158.

Rello, L. and Llisterri, J. (2012). There are phonetic patterns in vowel substitution errors in texts written by persons with dyslexia. In *21st Annual World Congress on Learning Disabilities (LDW 2012)*, pages 327–338, Oviedo, Spain, September.

Rello, L., Baeza-Yates, R., Saggion, H., and Pedler, J. (2012a). A first approach to the creation of a Spanish corpus of dyslexic texts. In *LREC Workshop Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, pages 22–27, Istanbul, Turkey, May.

Rello, L., Bayarri, C., and Gorriz, A. (2012b). What is wrong with this word? Dyseggxia: a game for children with dyslexia (demo). In *Proc. ASSETS'12*, pages 219–220, Boulder, USA, October. ACM Press.

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

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

Silva Rodríguez, A. and Aragón Borja, L. (2000). Análisis cualitativo de un instrumento para detectar errores de tipo disléxico (Qualitative analysis of an instrument to detect dyslexic errors, IDETID-LEA). *Psicothema*, 12(2):35–38.

Speech Therapy Association of Catalonia. (2011). *PRODISCAT: Protocol de detecció i actuació en la dislèxia. Àmbit Educativo (Protocol for detection and management of dyslexia. Educational scope.)*. Education Department of the Catalonian government.

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

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

Toro, J. and Cervera, M. (1984). *TALE: Test de Análisis de Lectoescritura*. Visor, Madrid, Spain.

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

Vellutino, F., Fletcher, J., Snowling, M., and Scanlon, D. (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.