Screening Dyslexia for English Using HCI Measures and Machine Learning

Luz Rello  
HCI Institute  
Carnegie Mellon University  
Pittsburgh, USA  
luzrello@cs.cmu.edu

Enrique Romero  
Department of Computer Science  
Universitat Politècnica de Catalunya  
Barcelona, Spain  
eromero@cs.upc.edu

Maria Rauschenberger  
Web Science and Social Computing Research Group  
Universitat Pompeu Fabra  
Barcelona, Spain  
maria.rauschenberger@upf.edu

Abdullah Ali  
Information School  
University of Washington  
Washington, USA  
xyleques@uw.edu

Kristin Williams  
HCI Institute  
Carnegie Mellon University  
Pittsburgh, USA  
krismawil@cs.cmu.edu

Jeffrey P. Bigham  
HCI & LTI Institutes  
Carnegie Mellon University  
Pittsburgh, USA  
jbigham@cs.cmu.edu

Nancy Cushen White  
Department of Pediatrics  
University of California San Francisco  
San Francisco, USA  
nancycushen.white@ucsf.edu

ABSTRACT

More than 10% of the population has dyslexia, and most are diagnosed only after they fail in school. This work seeks to change this through early detection via machine learning models that predict dyslexia by observing how people interact with a linguistic computer-based game. We designed items of the game taking into account (i) the empirical linguistic analysis of the errors that people with dyslexia make, and (ii) specific cognitive skills related to dyslexia: Language Skills, Working Memory, Executive Functions, and Perceptual Processes. Using measures derived from the game, we conducted an experiment with 267 children and adults in order to train a statistical model that predicts readers with and without dyslexia using measures derived from the game. The model was trained and evaluated in a 10-fold cross experiment, reaching 84.62% accuracy using the most informative features.

CCS CONCEPTS

• Computers and Society; • Social Issues; • Assistive Technologies for persons with disabilities;

KEYWORDS

Dyslexia, screening, early detection, diagnosis, linguistics, serious games, machine learning

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:

1 INTRODUCTION

More than 10% of the population has dyslexia [11, 26]. The DSM-V [1] defines dyslexia as a specific learning disorder with a neurological basis. According to the World Federation of Neurology it occurs in children who, despite conventional classroom experience, fail to attain the language skills of reading, writing, and spelling commensurate with their intellectual abilities [28]. In summary, dyslexia is frequent, universal and related to school failure. However, it remains under-diagnosed. For instance, in the UK, a country that effectively treats dyslexia as compared with other countries, only 5% of the individuals with dyslexia are diagnosed and given appropriate help, and [7]. It is estimated that over 85% of adult illiterates have dyslexia [7].

Yet, early detection is crucial for addressing dyslexia and effective remediation. Often, students are under-diagnosed because current procedures for diagnosis are expensive [16, 21] and require professional oversight [3, 8]. Our goal is for anyone to know as early as possible if they might have dyslexia in an inexpensive way.

To achieve this goal, we have created a computer game that records a wide variety of web-page interaction measures to screen dyslexia for English. We conducted a user study with 267 participants to collect data to train a machine learning model that is able to correctly determine if a person has (or does not have) dyslexia with 84.62% accuracy.
2 BACKGROUND AND RELATED WORK
The complexity of administering paper-based diagnostic tools, and the time they require, have led educators to turn towards computer-based screening methods to derive a quick assessment.

2.1 Commercial Software
Among the available commercial software to detect dyslexia in English there is Lexercise Screener [13] and Nessy [17]. We could not find studies behind these commercial applications, although they are widely used in practice. To our knowledge, they are not based on a machine learning model predictive of dyslexia, and we have not found publication of their accuracy.

2.2 Computer Based Games
There are a number of computer games designed to screen for dyslexia, but they do not use machine learning models.

Lyytinen et al. [15] created the computer game Literate, later called GraphoGame [14], to identify children at risk in Finland. The game was tested with 12 and 41 children between 6 and 7 years old - with statistically significant differences.

There are three other on-going projects for early risk detection of dyslexia that have not yet reported significant results yet: one approach for Italian tested with 24 pre-schoolers [9], and a language-independent approach MusVis evaluated with German, English and Spanish children [19, 24].

2.3 Machine Learning Approaches
Machine learning approaches to predict dyslexia are more recent. In 2015, the first method to screen dyslexia in Spanish was introduced; it used eye-tracking measures from 97 subjects (48 with dyslexia) [21]. Later in 2016, eye-tracking measures were also used to predict dyslexia for Swedish (185 subjects, 97 of them with high-risk of dyslexia)[2]. Both methods used Support Vector Machines. Another study detected dyslexia subtypes in the Hebrew language using data derived from existing medical records [12].

The only approach we are aware of to predict risk of dyslexia using features derived from computer-based measures is the game Dytective for Spanish. The screener, Dytective, was first evaluated with 343 people (95 with diagnosed dyslexia) and attained 83% accuracy in a held-out test set with 100 participants using Support Vector Machines [23]. Later, the model was improved by applying a neural network model (Long Short-Term Memory Networks (LSTMs) [10]) to a larger dataset—4,335 participants (763 with professional dyslexia diagnosis)—attaining 91.97% accuracy [22]. This model was integrated into a free online tool Dytective which has been used over 100,000 times.\(^1\) An earlier study piloted Dytective’s screening measures with 60 English speaking children and found the feature set was promising, but the study did not fully incorporate machine learning methods [25].

We advance these approaches by (i) extending these methods to the English language and (ii) include a wider number of items targeting cognitive indicators predictive of dyslexia.

\(^1\)https://dytectivetest.org/

3 USER STUDY
We conducted a within-subject study (267 participants) with all participants exposed to the same linguistic items integrated into an online game Dytective.

3.1 Procedure and Ethics Statement
Participants completed the experiment remotely, through a computer at home, school, or in a specialized center in the USA (mainly from the states of Pennsylvania, New York and Texas). All participants agreed to participate through an online consent form, and children provided assent along with their parent or legal guardian following protocols approved by our institutional review board (IRB). Parents/legal guardians were specifically warned that they could not help their children complete the study exercises. When schools and specialized centers oversaw participation, parental/legal guardian consent was obtained in advance, and the study was supervised by the school counselor or therapist.

The first part of the study consisted of a questionnaire collecting demographic data. This questionnaire was completed by the participant’s supervisor (school counselor or therapist) in cases when the participant was under 18 years of age. Then, following oral instructions, participants were given 20 minutes to complete the test exercises.

3.2 Participants
We recruited 267 participants from one specialized center, three schools, and from individuals with dyslexia who knew about our study through our public call online.

Subjects ranged in age from 7 to 60 years old. We classified these participants into three groups. Of the participants, 52 were diagnosed with dyslexia -Class D (dyslexia)- (28 female, 24 male, \(M = 11.16, SD = 6.31\)) and 206 without a diagnosis of dyslexia served as a control group -Class N (Not-Dyslexia) (94 female, 112 male, \(M = 11.89, SD = 5.11\)). There were 9 participants at risk of having dyslexia or suspected of having dyslexia -Class M (Maybe)- (4 female, 5 male, \(M = 17.66, SD = 16.17\)).\(^2\)

The first language of all participants was English, although 84 participants spoke another language (mostly Spanish in the Texas area). A total of 224 participants reported having trouble with language classes at school.

3.3 Dependent Measures
Participants’ performance was measured using the following dependent measures for each of the exercises: (i) Number of Clicks per item; (ii) Hits (i.e., the number of correct answers); (iii) Misses (i.e., the number of incorrect answers); (iv) Score (i.e., the sum of correct answers for each stage’s problem type); (v) Accuracy (i.e., the number of Hits divided by the number of Clicks); and (vi) Miss Rate (i.e., the number of Misses divided by the number of Clicks).

We later used these performance measures together with the demographic data as features of our prediction model’s dataset (see Section 4).

\(^2\) All were either adults or children under observation by professionals, the step before having an official diagnosis.
We integrated test items into a software game to serve as the primary material of our study. 3.4 Materials We integrated test items into a software game to serve as the primary material of our study.

3.4.1 Design and Implementation. Dyteective is a cross-platform web-based game built in HTML5, CSS, JavaScript and a PHP server and a MySQL database. It was designed with a high level of abstraction to make it easily portable for future native implementations. The interface design of the game implements the guidelines that, according to the latest findings in accessibility research, ensure the best on-screen text readability for this target group. Text is presented in black using a mono-spaced typeface Courier and a minimum font size of 14 points [20].

3.4.2 Playing Dyteective. At each phase, the player’s goal is to accumulate points by solving a linguistic problem type as many times as possible in a 25-second time window. For example, the player hears the target, non-word crench and then a board is shown on screen containing the target non-word as well as distractors that are particularly difficult for people with dyslexia to differentiate (See Figure 1 (a)). After each time window, the player continues on to the next item corresponding to a new linguistic problem type.

3.4.3 Content Design. The test items are composed of a set of attention and linguistic exercises addressing three or more of the following indicators belonging to different types of Language Skills, Working Memory, Executive Function, and Perceptual Processes. These indicators are related to dyslexia [5, 6, 27].

The exercises were designed according to linguistic knowledge and the expertise of dyslexia therapists (specific to the English language). In addition, to assist item selection (exercises) we used the following criteria:

(i) linguistic analyses of 833 confusion sets, created from the errors of people with dyslexia writing in English [18]; and
(ii) Performance measures from the linguistic exercises of an online game called Pirateetas. This game is part of a previous work targeting children with dyslexia to improve spelling performance [24]. We selected exercises that were more challenging for the players (those with higher error rates and need for more time to be solved) since those exercises were more likely to manifest dyslexia difficulties.

Table 1: Indicators used for the design of the test items.

<table>
<thead>
<tr>
<th>Language Skills</th>
<th>Working Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alphabetic Awareness</td>
<td>Visual (alphabetic)</td>
</tr>
<tr>
<td>Phonological Awareness</td>
<td>Auditory (phonology)</td>
</tr>
<tr>
<td>Syllabic Awareness</td>
<td>Sequential (auditory)</td>
</tr>
<tr>
<td>Lexical Awareness</td>
<td>Sequential (visual)</td>
</tr>
<tr>
<td>Morphological Awareness</td>
<td><strong>Executive Functions</strong></td>
</tr>
<tr>
<td>Syntactic Awareness</td>
<td>Activation and Attention</td>
</tr>
<tr>
<td>Semantic Awareness</td>
<td>Sustained Attention</td>
</tr>
<tr>
<td>Orthographic Awareness</td>
<td>Simultaneous Attention</td>
</tr>
<tr>
<td><strong>Perceptual Processes</strong></td>
<td><strong>Auditory Discrimination and Categorization</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Table 1: Indicators used for the design of the test items.</strong></td>
</tr>
<tr>
<td></td>
<td><strong>3.4 Materials</strong></td>
</tr>
</tbody>
</table>

3.4.4 Materials

4 DATASET

The dataset is composed of 226 features per participant (i.e., total of 60,342 data points). Each participant from the dataset was marked as D if the participant has dyslexia, N if not, and M (maybe) if the participant suspects that he or she has dyslexia but is not diagnosed. From the dataset we extracted the following features:

1. **Gender** of the participant. A binary feature with two values, female and male.
2. **Age** of the participant ranging from 7 to 60 years old.
3. **Second language.** A binary feature with two values, no and yes, when the participant had a second language in case of bilingualism.
4. **Language subject.** This is a binary feature with two values, no and yes, when the participant declares that she has trouble with language classes at school.

Features from 5 to 226 are performance measures; they correspond with the six dependent measures (Clicks, Hits, Misses, Score, Accuracy, and Missrate) per level played (37 levels).

These features target some of the skills presented in Table 1. Note that all the exercises involve attention, so all these features target the executive functions activation and attention, and sustained attention. In addition, some of them also target simultaneous attention when the participant pays attention to a number of sources of incoming information at the same time.

5-28 These features are performance measures related to alphabetic awareness and visual discrimination and categorization. For these tasks the participant hears the name of a letter, e.g., d, and identifies it from among the distractors (orthographic and phonetically similar letters, e.g. b, q, p) within a time frame, using a Whac-A-Mole-style game interaction.

29-52 These features relate to phonological awareness and auditory discrimination and categorization. The participant listens to the sound (phoneme) of a letter and identifies it from among distractors. For example, the participant hears the phoneme /n/ and then a board is shown containing the target <n> as well as distractors. We use distractors that are particularly difficult for people with dyslexia to differentiate (i.e., other phonemes that share phonetic features, such as nasal and sound consonants).

53-88 These features target syllabic awareness and auditory discrimination and categorization. The players hear the pronunciation of a syllable (e.g., /prin/) and identify its spelling from among orthographic distractors <pren>, <prein>, <prain>, <prean>, and <pryn>.

89-112 These features correspond to a set of exercises where participants identify a word’s spelling after hearing its pronunciation (e.g., /greet/ by discriminating among phonetically and orthographically similar words and/or non-words (e.g., <create>, <greater>, <great>, <creater>, <greater>, <crete>, <creat>). These features target lexical awareness, auditory working memory, and auditory discrimination and categorization.

113-136 These performance features correspond to exercises targeting visual discrimination and categorization, by requiring participants to find as many different letters as possible.

---

3 A confusion set is a small group of words that are likely to be confused with one another—such as weather and whether.

These features relate to 179-184
This set of features targets 161-172
These performance features target 191-196
These features were extracted from a set of exercises requiring participants to supply a missing letter [161-166] or delete an extra letter in a target word [167-172]. See Figure 1, exercises (d) and (e), respectively.
These performance features target 193-194
These features relate to 185-190
This set of features relates to 191-202
This set of features requires participants to separate words to make a meaningful sentence, e.g. Change sher-anupthehill to e.g. she ran up the hill.
This set of features targets 203-208
These performance features target 173-178
These features target sequential auditory working memory, and auditory discrimination and categorization. See Figure 1, exercise (a).
These performance features target lexical, phonological, and orthographic awareness; They are derived from exercises requiring participants to supply a missing letter [161-166] or delete an extra letter in a target word [167-172]. See Figure 1, exercises (d) and (e), respectively.
These features target morphological and semantic awareness. They are collected from exercises requiring participants to find a morphological error in a sentence when there is also a semantic error. For example, in the sentence, The affect of the wind was to cause the boat's sails to billow. (The word affect should be effect).
These features relate to syntactic awareness. Participants find an error in a sentence related to a grammatical or function word that changes, (e.g., of instead of on in “Smoking is prohibited of the entire aircraft”).
Features 191-196-require participants to rearrange letters to spell a real word (e.g., b e c u a s e) or to rearrange syllables to spell a real word (e.g., /na/ /na/ /ba)-syllabic, lexical, and orthographic awareness (Features 197-202).
Figure 1: Screenshots of the exercises requiring the player to click on the target non-word listed among the distractors; (a) select the different letter; (b) build a correct word by substituting a letter, (c) selecting a letter, (d) or deleting a letter (e).

Table 2: Classifier accuracy in the cross validation experiment, using the optimized feature set.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>84.62%</td>
</tr>
<tr>
<td>Precision – Class D (Dyslexia)</td>
<td>63.76%</td>
</tr>
<tr>
<td>Recall – Class D (Dyslexia)</td>
<td>80.24%</td>
</tr>
<tr>
<td>Precision – Class N (Not-Dyslexia)</td>
<td>93.88%</td>
</tr>
<tr>
<td>Recall – Class N (Not-Dyslexia)</td>
<td>85.83%</td>
</tr>
</tbody>
</table>

5 RESULTS AND DISCUSSION
To determine whether it is feasible to detect whether a user may have dyslexia, we set up a machine learning experiment. We carried out an experiment with a binary classifier of LIBSVM [4] in the Gaussian Support Vector Machine (SVM) setup. An SVM is a method for supervised machine learning that analyzes data and finds patterns for classification. As other Machine Learning algorithms, given a set of training examples, each marked as belonging to a category, an SVM training algorithm builds a model that assigns new examples into the categories. The particular bias of SVMs is that of constructing a hyperplane (either in the original space or in a transformed one) for the classification output. This hyperplane is constructed by combining the original input examples with the aim of maximizing the functional margin. Our SVM is trained on the dataset as the one described in Section 4.

We performed a 10-fold cross validation experiment by dividing the data into 10 different roughly equal subsets (10% of the data in each subset). Then, we trained a statistical model on the rest of the data (90%) and tested on the corresponding fold by iterating 10 times; at the end, all data was tested independently. We used 10-fold cross validation because it is normally recommended for smaller datasets when a single train-development test split might not be informative enough.
We randomized the data and used stratified sampling to ensure a similar distribution of data categories in all folds. Participants marked as M (Maybe) were assigned to the class D (Dyslexia). Outliers’ values in the number of Clicks and Misses were limited to a maximum fixed value. Subsequently, the data were scaled to zero mean and unit variance.

We analyzed the data for features whose distributions were different between dyslexic and non-dyslexic participants. To that end, a Kolmogorov-Smirnov test was performed. The number of Hits and Misses showed different distributions for a number of exercises.

Table 2 shows the accuracy of the SVM model (Gaussian kernel). This result suggests that the model is able to predict players with dyslexia quite accurately with a final result of 80.24% by using a subset of informative features. Note that the baseline (the percentage of subjects assigned to the class Dyslexia in the data set) is 22.85%.

The most informative features were a set of 10 features composed of Hits and Misses. Misses being the most informative ones at an individual level. These features are performance measures belonging to exercises that target Alphabetical Awareness, Phonological Awareness, Visual Categorization and Auditory Discrimination and Categorization. More concretely, these features come from exercises where the participant was required to map (or associate) a letter name or a letter sound with a grapheme (letter or letters). This is consistent with previous literature on dyslexia that focus on the deficit on the phonological component in dyslexia [26, 27].

6 CONCLUSIONS AND FUTURE WORK

We presented a method to screen for risk of dyslexia among English speakers that combines machine learning and web-based interaction data collected from a linguistic game. The method was evaluated with 267 participants and attained 84.62% accuracy on its prediction. These results build on earlier findings from the first version of Dytective [23], where only Spanish was considered.

These results should be taken as preliminary, since the model was trained on a small dataset. Further experiments with more participants under other less controlled conditions are needed. Our next step will be to conduct a large-scale study. With positive results, we will integrate the model in a tool to screen risk of dyslexia online. Since estimations of dyslexia are much higher than the actual diagnosed population, we believe this method has potential to make a significant impact.

ACKNOWLEDGMENTS

This paper was developed under a grant from the US Department of Education, NIDDR grant number H133A130057; and a grant from the National Science Foundation (#IIS-1618784).

We thank the Valley Speech Language and Learning Center in Brownsville in Texas, and the schools Winchester Thurston School (Pittsburgh, PA), and Ellis School (Pittsburgh, PA). We thank the volunteers who participated (Adam Brownold, Elsa Cárdenas-Hagan, Anne Fay, Susan Freudenberg) in supervising the participants. Thanks to Lola Álvarez, Susanne Burger, Debbie Meyer, and Regina Rash for their help with recruiting participants.

REFERENCES


