A New Error Annotation for Dyslexic texts in Arabic

Maha M Alamri
School of Computer Science
Bangor University
Bangor, UK
maha.alamri@bangor.ac.uk

William J Teahan
School of Computer Science
Bangor University
Bangor, UK
w.j.teahan@bangor.ac.uk

Abstract

This paper aims to develop a new classification of errors made in Arabic by those suffering from dyslexia to be used in the annotation of the Arabic dyslexia corpus (BDAC). The dyslexic error classification for Arabic texts (DECA) comprises a list of spelling errors extracted from previous studies and a collection of texts written by people with dyslexia that can provide a framework to help analyse specific errors committed by dyslexic writers. The classification comprises 37 types of errors, grouped into nine categories. The paper also discusses building a corpus of dyslexic Arabic texts that uses the error annotation scheme and provides an analysis of the errors that were found in the texts.

1 Introducation

Gallagher and Kirk (1989) divided learning disabilities into two types: developmental learning disabilities and academic learning disabilities. Developmental learning disabilities include attention, memory, perceptual, perceptual-motor, thinking and language disorders; while academic learning disabilities include reading, spelling, handwriting, arithmetic and writing expression disorders. This paper focuses on spelling disabilities, with a focus on the spelling difficulties encountered by people suffering from dyslexia. The word dyslexia originates from the Greek and signifies "difficulty with words" (Ghazaleh, 2011). Dyslexia International (2014) has reported that dyslexia affects around one in ten individuals.

Dyslexia has become a topic of debate in different fields, including education, psychology, neuropsychology, linguistics and other sciences. Some studies have attempted to analyse and explain textual errors committed by writers with this condition, though to date there is no standard error classification specifically for dyslexia errors.

Most of the studies carried out in this field did not categorise the errors but focused only on listing them. This study addresses this gap by developing a new dyslexia error classification system based on the results of a number of dyslexia error analysis studies as described in the next section.

This paper is organised as follows. Section 2 covers studies that discuss the errors caused by dyslexia. Section 3 describes the classifications used to annotate Arabic dyslexia errors. Section 4 contains an evaluation of these classifications. Section 5 discusses building the Arabic dyslexia corpus, followed by section 6 which explains the annotation process. Section 7 shows the analysis of dyslexic errors. Lastly, some suggestions for further work and conclusions are presented in Section 8.

2 Basis of dyslexic error classification for Arabic texts (DECA)

The DECA developed for this study relies on the findings of the studies mentioned below that discuss dyslexia errors from different aspects. For instance, Burhan et al. (2014), discuss the errors using a survey of teachers on which errors they believe are most common.

According to Ali (2011), spelling disabilities often cause letter reversals, also known as mirror writing and writing from left to right. As Arabic is written from right to left, writing from left to right can result in a correctly written sentence; mirror writing causes the sentence to be reversed. Ali (2011) also highlights other common errors including omission, addition, substitution and transposition. Dyslexic students also have difficulties

Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP), pages 72–78,
Valencia, Spain, April 3, 2017.

differentiating between letters with similar forms and different sounds.

Abu-Rabia and Taha (2004) examined the spelling mistakes made by speakers and writers of Arabic. They compared dyslexia with two groups of participants, namely, participants with, a young readers' group, matched with the dyslexic participants by reading level and an age-matched group. The study revealed seven types of errors: phonetic errors, semi-phonetic errors, dysphonetic errors, visual letter confusion, errors relating to irregular spelling rules, word omission and functional word omission. Other errors included students spelling an Arabic word according to how it is pronounced in the local spoken dialect of Arabic that they use in their day-to-day life, rather than using the correct Arabic spelling for it.

In order to examine the errors of female students with dyslexia Alamri and Teahan (2013) created a corpus of 1,067 words in a pilot project. During analysis, they identified a number of common spelling errors, including but not limited to: inability to specify the correct form of the Hamza; difficulty in short and long vowels; Tanween and exchanging ظ with ض , ض with ظ , ت with ة or ه and ة or ه with ت .

Burhan et al. (2014) also studied common errors made by students with learning disabilities; however, they used the viewpoints of teachers to identify the degree of common errors of 28 different kinds of errors.

Abunayyan (2003) created a document called "Error Analysis in Spelling - تحليل الأخطاء في مادة الإملاء", which is used in Saudi Arabia to analyse the spelling errors of dyslexic students in primary schools, it contains 23 different error types.

The following are three studies that give a brief overview of further studies, which examined dyslexic errors, corpora or lists of errors in other languages. These studies are relevant as they are examples of error annotations language resources that have been developed in other languages (although as stated nothing similar has been done for Arabic until now).

Pedler (2007) created a spelling correction programme that focuses on errors in words committed by individuals with dyslexia. This version comprises approximately 12,000 English words and 833 marked-up errors. The corpus used in this study comprised different resources, such as homework, online typing texts, texts created by dyslexic students studying for the IT NVQ and texts created by students on the dyslexia mailing list. Pedler (2007) created an English confused words list defined as "a small group of words that are likely to be confused with one another", such as 'form' and 'from'. The list included 833 sets of words which are regularly confused that were extracted from the corpus of texts written by people with dyslexia.

Rello (2014) compiled a Spanish corpus (Dyscorpus) comprising texts written by dyslexic children aged 6-15 years. The corpus comprised 83 texts: 54 taken from school essays and homework exercises and 29 from parents of dyslexic children, totalling 1,171 errors. Dyscorpus is annotated and provides a list of unique errors.

Rauschenberger et al. (2016) collected texts written in German from homework exercises, dictations and school essays. The corpus comprised 47 texts written by 8 to 17 year old students. The texts contained a list of 1,021 unique errors. The researchers created a new resource of German errors and annotated errors with a combination of linguistic characteristics.

3 Dyslexic error classification for Arabic texts (DECA)

There seems to be a consensus among researchers on some types of errors made by people suffering from dyslexia, such as 'omission'. However, some types of errors are only reported in single studies, for instance the 'functional words omission' error reported by Abu-Rabia and Taha (2004). These errors were excluded from this study because the prospect of their appearance is limited.

Most of the types in the classification deal with unique specificities of the Arabic language. The system of Arabic writing contains characteristics such as diacritics which does not exist in other languages. However, there are some types in the classification that occur in other languages, such as omission, substitution and addition. A classification of annotated errors was created for the Arabic corpus of this study which can help researchers of dyslexia in Arabic understand and identify error types more easily.

The DECA classification comprises a list of errors grouped into types and categories. The category is more general than the type: it specifies

whether the error occurs in the Hamza, in the Almadd, and so forth. Each error category is further subdivided into a variable error type. The nine error categories are "Hamza, Almadd, Confusion, Diacritics, Form, Common error, Differences, Writing method, Letters written but not pronounced (or Vice Versa)". A category called "Other" was also created to handle any error that does not yet have a "tag". The first version of the classification contained 35 error types. In each category, an error type called "Other" is added if the errors are not listed in the category. Alfaifi et al. (2013) suggests the use of two characters to represent the tag: the first specifying the category and the second specifying the error type; for example, in الهمزة على الألف (Alif Hamza Above), the tag would be <HA>with the (H) indicating the category الهمزات (Hamza), and the (A) indicating the error type (Above)على الألف .

To illustrate further, if the erroneous word is ثيمر and the correct word is ثمار; thus, the writer would write ي instead of the diacritical ثِ and deleted the letter ا. The erroneous word has one wrong letter added in one location and another correct letter missing in another location. Therefore, to indicate the two different types of errors, (_) can be used between the tags as follows: <DY_AA>.

4 Evaluating the DECA

Pustejovsky and Stubbs (2012) suggest that on the first round of annotations, it is best to select a sample of corpus to annotate in order to find out how well the annotation task works in practice. This will also help to evaluate the comprehensiveness, appropriateness and clarity of the classification and to determine if it serves the purpose of the error analysis.

Following Pustejovsky and Stubbs (2012) approach, 5000 words were chosen as a sample. The annotators used the classification Version 1 to annotate all errors completely manually, using the original handwritten text before transcribing it into an electronic form. They then provided a list of the types of errors encountered that matched with the classification and indicated if there were any new types not listed in the classification. The findings showed that all errors in the samples were annotated using the classification, except for two new types, which are "تكرار الحروف - Repeated letters" and " عدم القدرة على التفريق بين شكل الحرف إذا كان في بداية الكلمة أو وسطه أو نهايته - Form of the letter in the Beginning, Middle or End". Version 1 was edited to include these two errors. Therefore, Version 2 of the classification contained 9 categories and 37 errors types, as shown in Table 1.

Following this exercise, questionnaires were sent to two evaluators who had agreed to participate in this study. The evaluators were primary school teachers who teach children with learning disabilities. They were given the DECA Version 2 and were asked to read through the list of error categories and give feedback on whether they felt it comprised all the errors committed by dyslexic students and if the categories were appropriate. They were also asked to read through the sample text and tag it with the appropriate error tag.

Both evaluators found the correct tag for all sentences, except for one sentence containing the error word "Which -اللتي" where one chose the <FR>tag rather than <LT>. Both found the tags to be appropriately named. When asked how easily they found the right tag, their answers ranged from easy to difficult according to the sentence. Moreover, they found that the table presented all the types of dyslexic errors and that it was comprehensive.

5 Building the Arabic corpus (BDAC)

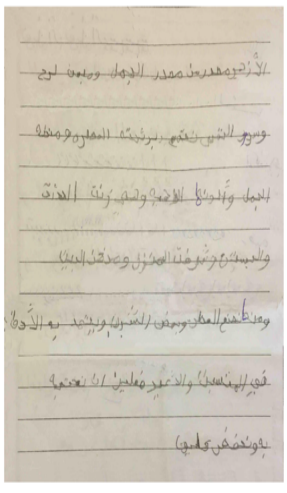
The size of the BDAC corpus is 27,136 words and 8000 errors in texts collected from Saudi Arabian primary schools, online forms and texts provided by parents. All participants were diagnosed with dyslexia by professionals. The texts written by dyslexics aged between 8 to 12 year olds, with some texts written by youths aged 13. The BDAC corpus contains texts written by both male and female students.

As some texts were handwritten, further work is needed for transcription into an electronic form. In addition, since some teachers or parents did not transcribe the correct text that the dyslexic wrote, further work is also required either by trying to find the correct text or by choosing the word in accordance with the written text as much as possible.

An example of a handwritten text written by 10 year-old girl with dyslexia shown in Figure 1.

TAG	Error Type - نوع الخطأ	Category - فئة
<HH>	Hamza on Line - الهمزة على السطر	الهمزات Hamza
<HA>	Alif Hamza Above - الهمزة على الألف	
<HB>	Alif Hamza Below - الهمزة تحت الألف	
<HY>	Ya Hamza Above - الهمزة على الياء	
<HW>	Waw Hamza Above - الهمزة على الواو	
<OH>	Other - لم تذكر في الهمزات	
<AA>	Alif Madd - مد الألف	المدود Almadd
<AW>	Waw Madd - مد الواو	
<AY>	Ya Madd - مد الياء	
<OM>	Other - لم تذكر في المدود	
<CT>	Confusion in Tah and Tah Marbuta/Hah - بين التاء المفتوحة والتاء المربوطة أو الهاء	الخلط Confusion
<CH>	Confusion in Hah and Tah Marbuta - بين الهاء والتاء المربوطة	
<CA>	Confusion in Alif and Alif Maksura - بين الألف الممدودة والألف المقصورة	
<CD>	Confusion in Dha and Tha - بين الظاء والضاد	
<CV>	Confusion in Similar Letters - الخلط بين حروف متشابهة	
<OC>	Other - لم تذكر في الخلط	
<DN>	N in Tanwin - نون مكان التنوين	الحركات Diacritics
<DW>	W in Damma - واو مكان الضمة	
<DY>	Y in Kasra - ياء مكان الكسرة	
<OD>	Other - لم تذكر في الحركات	
<FW>	Word Boundary Errors - وصل ماحقه الفصل من الحروف أو فصل ماحقه الوصل من الحروف	شكل الكلمة Form
<FR>	Repeated Letters - تكرار الحروف	
<FM>	Multi Errors - أخطاء متعددة	
<OF>	Other - لم تذكر في الشكل	
<MO>	Omission - حذف	الاخطاء الشائعة Common errors
<MA>	Addition - إضافة	
<MS>	Substitution - تبديل	
<MT>	Transposition - تحويل	
<DD>	Different Forms, Same Phonetics - عدم القدرة على تفريق بين حروف متشابهة لفظا المختلفة شكلا	الاختلافات Differences
<DF>	Form of the Letter in the Beginning, Middle or End - عدم القدرة على التفريق بين شكل الحرف إذا كان في بداية الكلمة أو وسطه أو نهايته	
<DI>	Local Language - الكتابة بناء على اللهجة المحلية	
<DS>	Writing a Word that is Similar to the Meaning - كتابة كلمة مشابهة للمعنى	
<OI>	Other - لم تذكر في الاختلافات	
<WM>	Mirror - مرآة	طريقة الكتابة Writing method
<WL>	Left to Right - الكتابة من اليسار لليمين	
<OW>	Other - لم تذكر في الكتابة	
<LS>	Sun Letters - لام الشمسية	حروف تكتب ولا تنطق أو العكس Letter written but not pronounced or vice versa
<LM>	Adding letter (L) to words start with letters (AL) - دخول اللام على مافيه (ال)	
<LA>	Alif Fariqa - ألف بعد واو الجماعة	
<LL>	(Lakn ...) - (لاكن - لاكنها ...)	
<LH>	(Hada ...) - (هاذا - هاذه - هاذان)	
<LT>	(Allty) - (اللتي)	
<LD>	(Alldhy) - (اللذي)	
<LK>	(Dahlk ...) - (ذالك – بذالك ...)	
<OL>	Other - لم تذكر في حروف	
<OT>	Other - لم تذكر في أي مجموعة	أخرى- Other

Table 1: Dyslexic error classification for Arabic texts (DECA).



الأزهار مصدر من مصادر الجمال ومبعث للراحة وسرور النفس نتمتع برائحتها العطرة ومنظرها الجميل وألوانها الزاهية وهي زينة الحدائق والبساتين وشرفات المنازل ومداخل البيوت ومنها نصنع العطور وبعض المربيات والشراب ويتهادى بها الأصدقاء في المناسبات والأعياد فعلينا ان نعتني بها ونحافظ عليها.

Flowers are a source of beauty sources and a source of comfort and self we are enjoying by the smell fragrant and their beautiful look and bright colours, also use to decorate gardens and orchards, balconies and entrances to homes, also, we can use it to produce perfume and some jams, drink and give it to friends on special occasions and holidays, so we have to take care of it.

Figure 1: Text written by 10-year-old girl with dyslexia.

In comparison with other languages, three studies carried out on different languages — English (12,000 words), Spanish (1,171 words) and German (1,021 words) (Pedler, 2007; Rello, 2014; Rauschenberger et al., 2016) — provide strong evidence that a small corpus of around 1,000 errors can yield useful results.

6 Annotating the BDAC corpus

As Granger (2003) points out, error annotation is a very tedious task that needs to be undertaken with care, but it has an immensely significant outcome as it makes it possible for the researcher to gain quick access to particular error statistics.

In order to illustrate the annotation process, Figure 2 shows a screenshot of a Java program that was created in order to speed up the annotation

process. A Java program was developed to convert (tokenise) the text into tokens. Each token is located in a separate line, and the erroneous words are manually annotated with each type of error based on the classification and the correct spelling of the erroneous word.

As shown in Figure 2, the text includes 43 tokens. In the example below, the first error is located in token 2. Thus, the annotator chose token 2, as it is an error word, by double-clicking on the error (token 2) in the text area labelled "Raw Text _ النص الاصلي", then chose the correct word from the text area labelled "Correct Text _ النص الصحيح" , again by double-clicking. Next, the appropriate tag was selected from the list. After that, "Apply - تنفيذ" is clicked, and it appears in the "Raw Text _ النص الاصلي" area as shown in Figure 2. The procedure is repeated with each error found in the text. In the case of a word that contains more than one type of error, as denoted by token 6, the annotator can add another tag via the "+" button, and choose another tag which is separated by (_). As a result, the annotation for token 6 is:

```
Tn="6" CorrectForm="أقتبس" Tag="HA_MA" ErrorForm="اسقتبس"
```

Figure 2: Screenshot of Java program to aid manual annotation process.

Each error token requires two annotations: one for the correct word and the second for the error type, as follows:

```
Tn="1" CorrectForm="الشمس" Tag="LS" ErrorForm="اشمس"
```

where:

- **Tn** = Token number (position of the word within the sentence).
- **CorrectForm** = The correct spelling of the word.
- **Tag** = Contains abbreviation of the error type.
- **ErrorForm** = The error word.

The BDAC corpus (27,136 words) has been fully annotated using DECA Version 2. The combined information was ultimately converted to an XML file as shown in Figure 3 below:

```
<Record_info>
   <Participant_info>
      <Age>10</Age>
      <Gender>Female - انثى</Gender>
   </Participant_info>
   <Text_info>
      <TextSource>HW - الواجب</TextSource>
      <Text n="1">
         <Category>word - كلمة</Category>
         <RawText>البن</RawText>
         <CorrectText>اللبن</CorrectText>
         <Error_analysis>
            <Token Tn="1" CorrectForm="اللبن" Tag="LM">البن</Token>
         </Error_analysis>
      </Text>
   </Text_info>
</Record_info>
```

Figure 3: A sample of the XML format used for the BDAC.

7 Analysis of Dyslexic Errors

Annotating the corpus has a significant advantage in terms of being able to search for particular error types or groups of errors in exactly the same way as individual words are searched (Nicholls, 2003). Once the annotation is carried out, corpus analysis becomes the simple procedure of extracting the tags or error and their corresponding target word. Some errors occur more than others in the corpus. Table 2 below shows the frequency of errors for the top five errors.

Error word	Number of Occurrences
On - علا	64
Which - اللتي	59
Which - اللذي	47
To - الى	35
That - ذالك	31

Table 2: Frequency of errors.

The correct form for the first error (علا) is (على). The error type is (CA), which falls under the "Confusion – الخلط" category. The second, third and fifth errors fall under the "Letter written but not pronounced or vice versa – حروف تكتب ولا تنطق أو العكس" category, for which the correct forms are (ذلك), (الذي), (التي), respectively. Finally, the fourth error falls under the "Hamza – الهمزات" category, where the correct form is (إلى) and the error type is (HB).

The highest number of errors for specific category was for the "Common errors – الأخطاء الشائعة" category with 2,717 error words; followed by 1,621 errors in the "Hamza – الهمزات" category and 1,553 errors in the "Confusion – الخلط" category. The lowest two types of errors fell within the "Differences – الاختلافات " and "Form – شكل الكلمة " categories.

The Alif Madd (مد الألف) error was the most frequent type of error making up 13.43% of total number of errors. This is in contrast with Burhan et al. (2014) finding that (كلمات مبدوءة بـ (أل) إذا سبقتها اللام المكسورة) are the most frequent type of errors made by Arabic dyslexic students from the teachers' viewpoint.

The most common errors in made by dyslexic persons are addition (13.4%), omission (10.98%), substitution (6.36%) and transposition (3.23%). This contrasts with Alamri and Teahan's (2013) study which found that the highest number of errors were errors of omission rather than addition.

Dyslexic people are popularly known to confuse Tah and Tah Marbuta/Hah (بين التاء المفتوحة والتاء المربوطة أو الهاء), with 6.55% of the errors falling under this type of error. This is consistent with Burhan et al. (2014) who found that this type of error is noticeably more apparent in the writing of people who suffer from dyslexia.

8 Conclusion and recommendations for further work

The DECA was introduced in response to the lack of a standard classification for dyslexia errors in Arabic. It was developed on the basis of prior error classification studies. Two people assessed the DECA classification for Arabic dyslexic errors and found it to be reliable and effective. The last version of the DECA includes 37 types of errors classified under nine categories.

The findings could be helpful for the field of pedagogy in general and for researchers of dyslexia in particular. This classification is valuable and can serve as a springboard to provide improved aid to this target group and also make the annotators' task less stressful.

Further work is required to improve the DECA in collaboration with special education needs and corpus linguistics specialists. Since the BDAC was collated from writings of residents of only one country (Saudi Arabia), one way to improve the classification is by collecting further texts from various countries. This may yield different types of errors, which could then be added to the classification developed in this study as a standard error classification which could be applied to other Arabic dyslexia corpora.

Acknowledgments

We deeply thank teachers, parents and all children for providing Arabic texts written by dyslexics.

References

Salim Abu-Rabia and Haitham Taha. 2004. Reading and spelling error analysis of native. *Reading and Writing*, 17(7-8):651–690.

Ibrahim S. Abunayyan. 2003. Error analysis in spelling (form 5 /M3).

Maha M. Alamri and William J. Teahan. 2013. Investigating dyslexic Arabic text. Master's thesis, School of Computer Science, Bangor.

Abdullah Alfaifi, Eric Atwell, and Ghazi Abuhakema. 2013. Error annotation of the arabic learner corpus. In *Language Processing and Knowledge in the Web*, pages 14–22. Springer.

Mohammed A. M. Ali. 2011. *Learning difficulties between skills and disorders*. Dar Safa Publishing - Distribution, Amman.

Hamadneh Burhan, Mohammad M. Al-Salahat, Maher T. Al-Shradgeh, and Wael A. Alali. 2014. Degree of common misspellings of students with learning disabilities. *The International Interdisciplinary Journal of Education (IIJOE)*, 3(6).

Dyslexia International. 2014. DI-Duke report, April.

James J. Gallagher and Samuel A. Kirk. 1989. *Educating exceptional children*. Boston: Houghton Mifflin Company.

Esfandiari B. Ghazaleh. 2011. A study of developmental dyslexia in middle school foreign language learners in iran. *Argumentum*, 7:159–169.

Sylviane Granger. 2003. Error-tagged learner corpora and call: A promising synergy. *CALICO journal*, pages 465–480.

Diane Nicholls. 2003. The cambridge learner corpus: Error coding and analysis for lexicography and elt. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, pages 572–581.

Jennifer Pedler. 2007. *Computer correction of real-word spelling errors in dyslexic text*. Ph.D. thesis, Birkbeck, University of London.

James Pustejovsky and Amber Stubbs. 2012. *Natural language annotation for machine learning*. O'Reilly Media, Inc.

Maria Rauschenberger, Luz Rello, Silke Fchsel, and Jrg Thomaschewski. 2016. A language resource of German errors written by children with dyslexia. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Luz Rello. 2014. *A Text Accessibility Model for People with Dyslexia*. Ph.D. thesis, Department of Information and Communication Technologies, University Pompeu Fabra.