

Bangor University

DOCTOR OF PHILOSOPHY

Analysing and Correcting Dyslexic Arabic Texts

Alamri, Maha

Award date:
2019

Awarding institution:
Bangor University

[Link to publication](#)

PREVIEW

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 24. Sep. 2020

ProQuest Number:28214790

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent on the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 28214790

Published by ProQuest LLC (2020). Copyright of the Dissertation is held by the Author.

All Rights Reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346



PRIFYSGOL
BANGOR
UNIVERSITY

School of Computer Science and Electronic Engineering
College of Environmental Sciences and Engineering

**Analysing and Correcting Dyslexic
Arabic Texts**

Maha Marzouq Alamri

Submitted in partial satisfaction of the requirements for the
Degree of Doctor of Philosophy
in Computer Science

Acknowledgements

Always and forever, all praise and thanks be to God (Allah) for everything and for guiding me towards the successful accomplishment of my PhD—a dream that has now become a reality.

I owe many people my sincere gratitude for helping me throughout the process. I should start with expressing my special gratitude to my supervisor, Dr. William J. Teahan, for every single act of support that he has offered for me—thank you for your substantial feedback and dedicated encouragement.

I am especially and profoundly grateful to the provenience of my love, my extended family. Many special thanks and love go to my parents, Marzouq and Faddah, for their infinite love, uncountable prayers and permanent wishes. My brothers and sisters, you know how excited I was throughout the past years in spite of the anticipated obstacles, and you were such huge supporters behind the scene, making me stronger and laugh louder whilst also inspiring me to keep going. I love and thank you all for everything you have done for me.

My special gratitude and appreciation also go directly to my immediate family. My husband, Turki, I know how proud you are feeling right now and

I know it is because my dream has now been achieved. I know there are no words that can describe the uncertainty we had experienced during the long journey of my study, however, you stood by me in everything and always allowed me to nurture my ambition and sustain my enthusiasm towards this moment. My deepest thanks should go to you, and to our daughter, Deema, the light of my life, whose presence with us is the true platform to our happiness.

I would also like to give special thanks to those who participated in this research and contributed to the study in spite of their exhausting academic schedules. Your contributions are key to the successful completion of this thesis, and I am sincerely grateful for your involvement. In addition, I wish to thank everyone who offered me invaluable advice, encouragement and support, including Dukhnah Alamri (lecturer) and Asma Aljathlan, Elham Alsaleh (teachers) and Maha Alammar.

I would like to thank Bangor University for providing an ideal environment for study and research. I am profoundly thankful to all staff and fellow researchers at the School of Computer Science and Electronic Engineering for their help and direction, especially Prof. Ludmila Kuncheva, Dr. Mohammed Mabrook, Dr. Noor Al-Kazaz and Dr. Nadim Ahmed.

Finally, I would like to acknowledge my country, the Kingdom of Saudi Arabia, and Al-Baha University for providing me the invaluable opportunity to study in the United Kingdom.

I dedicate this thesis

To my parents, for their love, prayers, endless support and encouragement;

To my sister Wadha, for standing by me when things turn bleak; and

To people with dyslexia, for inspiring me.

Abstract

Dyslexia is a disorder that involves difficulty with literacy skills and language related skills. It is related to the inability of a person to master the utilisation of written language and affects a significant number of people. This thesis describes the development of the Bangor Dyslexia Arabic Corpus (BDAC) in order to facilitate the analysis and automatic correction of dyslexic Arabic text. This thesis has also developed a new classification of errors made in Arabic by people with dyslexia which was used in the annotation of the BDAC. The dyslexic error classification scheme for Arabic texts (DECA) comprises a list of dyslexia spelling errors classified into 37 types, and grouped into nine categories.

This thesis also investigates a new type of classification – dyslexia text classification – that identifies whether or not a text has been written by a person with dyslexia. The text compression scheme known as prediction by partial matching (PPM) has been applied to the problem of distinguishing dyslexic text from non-dyslexic text. Experimental results show that the F_1 score for PPM-based classification was 0.99 and outperformed other classifiers such as Multinomial Naïve Bayes and Support Vector Machines.

A new system called Sahah is also proposed for the automatic detection and correction of dyslexia errors in Arabic text. The system uses a language model based on the PPM text compression scheme in addition to edit operations (omission, addition, substitution and transposition). The correct alternative for each error word is chosen on the basis of the compression codelength. Two experiments were carried out to evaluate the usefulness of the Sahah system. Firstly, its accuracy was evaluated using the BDAC

containing errors made by people with dyslexia. Secondly, the results of Sahah were compared with the results obtained when using word processing software and the Farasa tool. The results show that the Sahah system significantly outperforms Microsoft Word, Ayaspell and the Farasa tool with an F_1 score of 0.83 for detection and an F_1 score of 0.58 for correction.

PREVIEW

Contents

1	Introduction	1
1.1	Background and Motivation	1
1.2	Research Questions	3
1.3	Aim and Objectives	4
1.4	Contributions	5
1.5	Publications	6
1.6	Thesis Outline	8
2	Literature Review	11
2.1	Introduction	11
2.2	Dyslexia	12
2.2.1	Dyslexia Identification	13
2.3	Arabic Language	15
2.3.1	Arabic Script	15
2.3.2	Arabic Vowels	17
2.3.3	Arabic Morphology and Phonology	19
2.3.4	Arabic Encoding Methods	21
2.4	Spelling Errors	21
2.4.1	Spelling Errors by Writers with Dyslexia	23
2.4.2	Spelling Errors by Arabic Writers with Dyslexia	24

2.5	Corpus Linguistics	27
2.5.1	Types of Corpora	27
2.5.2	Dyslexia Corpora	29
2.6	Natural Language Processing	30
2.6.1	Evaluation Techniques	31
2.6.2	Text Classification	33
2.6.2.1	Multinomial Naïve Bayes	33
2.6.2.2	Support Vector Machines	34
2.6.2.3	Arabic Text Classification	35
2.6.3	Dyslexia Classification	36
2.6.4	Spelling Correction	37
2.6.4.1	The Noisy Channel Model	38
2.6.4.2	n-gram Language Models	39
2.6.4.3	Arabic Spelling Correction	41
2.6.4.4	Dyslexia Spelling Correction	43
2.7	Text Compression	44
2.7.1	Prediction by Partial Matching	45
2.7.1.1	Variants of Prediction by Partial Matching	46
2.7.1.2	Compression Codelengths using PPM-Based Models	50
2.7.1.3	Text Classification Using PPM	51
2.7.1.4	PPM Compression Method for Arabic	52
2.8	Conclusion	54
3	Bangor Dyslexic Arabic Corpus	55
3.1	Introduction	55
3.2	Identifying Dyslexia in Saudi Arabian Schools	56
3.3	Data Collection Procedure	57

3.4	Transcribing the Handwritten Data	58
3.5	The Bangor Dyslexia Arabic Corpus	62
3.5.1	Text Resources of the BDAC Corpus	63
3.5.2	Categorisation	64
3.6	Participant Information	64
3.7	BDAC Frequency Profiling	66
3.8	Conclusion	67
4	Error Annotation for Dyslexic texts in Arabic	70
4.1	Introduction	70
4.2	Basis of Dyslexic Error Classification Scheme for Arabic Texts	71
4.3	Dyslexic Error Classification Scheme for Arabic Texts	74
4.4	Evaluating the DECA	77
4.4.1	Annotation Sample Evaluation	77
4.4.2	Teachers' Feedback Evaluation	80
4.4.3	Inter-annotator Agreement Evaluation	80
4.5	The Annotation Process of the BDAC Corpus	81
4.6	Analysis of Dyslexic Errors in the BDAC	84
4.7	Conclusion	87
5	Distinguishing Dyslexic Text from Non-dyslexic Text	89
5.1	Introduction	89
5.1.1	Example of Dyslexia Classification	90
5.2	Text Corpora	91
5.3	Classification Experiments	92
5.3.1	Dyslexia Corpus and Learner Corpus Experiment . . .	93
5.3.2	Arabic Children Corpora Experiment	95
5.3.3	Experiment to Compare PPM with Other Classifiers .	100

5.4	Conclusion	102
6	Automatic Correction of Arabic Dyslexic Text	103
6.1	Introduction	103
6.2	Spelling Correction Functions	104
6.3	The Sahah System for the Automatic Spelling Correction of Dyslexic Arabic Text	106
6.3.1	Pre-processing Stage	107
6.3.2	Error Detection and Correction Stage	110
6.3.2.1	Sub-stage 2a: PPM Correction	111
6.3.2.2	Sub-stage 2b: Error Detection	115
6.3.2.3	Sub-stage 2c: Edit Operations	116
6.3.3	Post-processing Stage	117
6.4	Evaluation	118
6.4.1	Evaluation Methodology	118
6.4.2	Experimental Results	121
6.5	Conclusion	125
7	Conclusion	126
7.1	Introduction	126
7.2	Summary of the Thesis	126
7.3	Review of Research Questions	132
7.4	Review of Aim and Objectives	133
7.5	Future Work	135
	References	137
	Appendices	159

List of Figures

3.1	Screenshot of the text transcription form.	60
3.2	Example of a handwritten text written by a girl with dyslexia with the intended text transcribed by a teacher.	61
3.3	Example of a handwritten text written by a girl with dyslexia.	64
3.4	Number of documents from each category.	65
3.5	Age range of the participants.	65
3.6	Character frequency distribution.	68
4.1	Interface of the tool developed to facilitate the annotation process.	83
4.2	XML sample with stand-off annotation by tokens.	85
5.1	Dyslexia classification using PPMD.	90
5.2	Accuracy of dyslexia (BDAC) versus non-dyslexia (ALC) text classification using different PPMD orders.	94
5.3	Pre-processing and post-processing for PPM using bi-graph replacement.	97
5.4	Accuracy of dyslexia (BDAC) versus non-dyslexia (BNDAC) text classification using different BS-PPM orders.	99
6.1	Workflow of the Sahah system.	107

6.2 Workflow of the error detection and correction stage of the Sahah system.	110
--	-----

PREVIEW

List of Tables

1.1	Publications that relate to this study.	7
2.1	Arabic joining groups and group letters.	16
2.2	Table of Arabic diacritics with their ALA-LC Romanization representations and samples with the character ‘ت’.	18
2.3	Confusion matrix of two classes.	32
2.4	The PPMD model after the string ‘ <i>dyslexicornotdyslexic</i> ’ has been processed.	48
2.5	Models for predicting character streams (Teahan, 1998). The symbol \hookrightarrow in the table represents an escape	51
2.6	The PPMD model after the string ‘المعلم’ has been processed. .	53
3.1	Examples of words for which it was not possible to transcribe.	62
3.2	Word frequency distribution.	66
4.1	Error types recorded in Arabic studies.	72
4.2	Version 1 of Dyslexic Error Classification Scheme for Arabic Texts (DECA).	76
4.3	Version 2 of Dyslexic Error Classification Scheme for Arabic Texts (DECA).	79
5.1	Text corpora used in the experiments.	91

5.2	Some examples of text found in each of the corpora used in the experiments.	92
5.3	Classification results of dyslexia (BDAC) versus non-dyslexia (ALC) text using different PPMD orders.	94
5.4	Classification results of dyslexia (BDAC) versus non-dyslexia (BNDAC) text using different PPMD orders.	96
5.5	Bi-graphs frequency statistics for the BDAC corpus and BNDAC corpus.	98
5.6	Classification results of dyslexia (BDAC) versus non-dyslexia (BNDAC) text using different BS-PPM orders.	99
5.7	Different order of BS-PPM over dyslexia and non-dyslexia corpora.	100
5.8	Summary of F_1 experimental results for dyslexia classification using cross-validation.	101
6.1	The most common prefixes and suffixes based on the dyslexia corpus analysis.	108
6.2	Cases of removing the redundant characters.	109
6.3	Example of an error that corrects with a different ordering of sub-stages.	111
6.4	Improving the model of sub-stage 2a.	113
6.5	Improving the model of sub-stage 2c.	113
6.6	Set of transformations rules from the DECA used for sub-stage (2a) in the Sahah system.	114
6.7	The codelength of possible alternatives spellings by using the confusions in Table 6.6 for the erroneous word “احمد”.	115
6.8	Codelengths for different candidate trigrams for a sample correction.	117

6.9	Detection and correction results after the pre-processing stage and sub-stage 2a of the Sahah system.	121
6.10	Detection and correction result after all stages and sub-stages of the Sahah system were applied.	121
6.11	Detection comparison using the Bangor Dyslexia Arabic Cor- pus (BDAC) corpus.	122
6.12	Correction comparing the Sahah system with the Farasa tool.	123

PREVIEW

Chapter 1

Introduction

1.1 Background and Motivation

The word dyslexia originates from the Greek language and signifies difficulty with words (Ghazaleh, 2011), and specifically issues with reading, spelling, and word recognition (Grigorenko, 2001). The earliest consideration of dyslexia was presented by W. Pringle Morgan, in November 1896 (Morgan, 1896). The article described the case of a 14-year-old boy who was apparently at an adequate level of intelligence and logical reasoning for his age, yet struggled considerably in terms of reading and writing skills. This article is one of the first reports concerning congenital word blindness. Therefore, Morgan is often considered as being the pioneer in the field of dyslexia (Guardiola, 2001).

There seems to also be a significant amount of people who have dyslexia, as the International Dyslexia Association (2012) reported that dyslexia affects 15-20% of any given population. It should also be noted that there is no relationship between dyslexia and a person's level of intelligence. Dyslexia

is popularly identified with numerous famous figures, such as Richard Branson.

Dyslexia concerns difficulty with acquiring literacy skills, and this difficulty can be present throughout person's life, and might influence their education over the long-term. Furthermore, the disorder is not culture- or language-specific, and is therefore not exclusive to specific cultures, and is observable in all age groups and in all languages (Reid, 2010), such as Arabic.

The Arabic language is one of the most widely used in many parts of the world. A study conducted by Holes (2004) suggested that two fundamental reasons exist for the wide usage of this language, the first that Arabic is the language of the 'Holy Quran', and the second that other languages, such as Urdu and Farsi, employ Arabic letters. Thus, Arabic was selected as the focus of this thesis in addition to it being the researcher's native language. Moreover, the researcher wished to help Arabic people with dyslexia, and to add value for this target group.

Nevertheless, despite the widespread use of the language, academic research concerning dyslexia in Arabic is scarce, because dyslexia is not widely recognised in the Arab region (Aboudan et al., 2011). This is evidenced by the fact that the first dyslexia association in the Arab world was established in Kuwait in 1999, many years after the equivalent association was established in the West, where the oldest such association, the International Dyslexia Association was established in 1949.

There are a number of different approaches that can be employed to help and support people with dyslexia, such as tools to assist in its diagnosis and assessment, and applications such as word prediction software, text classification, and spelling correction. These tools have been developed through

different methods such as natural language processing and corpus linguistics.

The use of text corpora has expanded in recent years as it plays a significant role in different aspects such as computational linguistics, and Natural Language Processing (NLP) research. Although the use of text corpora has enjoyed a relatively high level of interest in research, availability of dyslexia corpora is scarce (Pedler, 2007), and only a few studies have considered the potential benefits of using dyslexia corpora.

Furthermore, there is an obvious lack of Arabic dyslexia corpora, which highlights the importance of improving and enlarging the extant resource that was previously developed by the researcher of this thesis (Alamri, 2013). Such a corpus can be used as a starting point for developing a more extensive understanding of dyslexic errors in Arabic, and how they are written and moreover, towards investigating and developing Arabic dyslexia applications. Furthermore, it can serve as a platform for future researchers to develop further studies in the area, or to employ in the creation of applications for dyslexics.

These points formed the main inspiration for conducting this research study.

1.2 Research Questions

The research questions explored for this study are as follows:

1. What is an effective spelling error classification scheme for annotating and analysing Arabic dyslexic corpora?
2. How well dose a compression-based language modelling method, such

as the Prediction by Partial Matching (PPM) text compression method, compare to two well performed algorithms such as Multinomial Naïve Bayes (MNB) and Support Vector Machines (SVM) for classifying a text that has been written by a person with dyslexia?

3. Can PPM, in conjunction with other methods, be effectively applied to correcting a text that has been written by a person with dyslexia?

1.3 Aim and Objectives

The aim of this study is to investigate the effectiveness of a new approach to classifying and correcting Arabic dyslexic text, specifically, using the PPM compression method. This study seeks to evaluate how well this approach performs in applications using Arabic dyslexic corpus.

Therefore, this study's objectives in investigating the research questions are as follows:

- Review the extant literature regarding dyslexia, Arabic language, dyslexia spelling errors, corpus linguistics, text classification, spelling correction, and text compression (see Chapter 2);
- Improve the existing Arabic corpus of texts written by people with dyslexia (the Bangor Dyslexia Arabic Corpus (BDAC)) (see Chapter 3);
- Create a new dyslexic error classification scheme for Arabic dyslexic texts (DECA) (see Chapter 4);
- Develop and evaluate a method to classify whether or not a text has been written by a person with dyslexia, using the PPM compression

scheme, and compare the performance of the PPM with other classification methods, such as the Multinomial Naïve Bayes (MNB) and Support Vector Machines (SVM), when they are employed for the purpose of classifying dyslexic text (see Chapter 5);

- Design and evaluate an automatic spelling correction system for correcting spelling errors in Arabic texts, produced by people with dyslexia, by comparing them with other spelling correction tools (see Chapter 6).

1.4 Contributions

Since there is currently a lack of an Arabic dyslexia corpus, this study has the potential to make valuable contributions to the field. The specific contributions are as follows:

- The first and foremost contribution of this research study is the enlargement of the Arabic dyslexia corpus (the BDAC) to comprise 28,203 words written by both male and female with dyslexia aged between 8 to 13 year olds. Based on the literature review, the BDAC is the first dyslexia corpus for Arabic.
- The second contribution is the development of dyslexic error classification scheme for Arabic texts (DECA) that can provide a framework to help analysing and annotating specific errors committed by writers with dyslexia. Also, this has been used to provide an annotated dyslexic corpus (the BDAC) and then analysis of Arabic dyslexic errors, based on the corpus.
- The third contribution is the creation of Bangor Non-Dyslexia Arabic

Corpus (BNDAC), consisting of 9,099 words written by non-dyslexic male and female between the ages of 8 and 13.

- The fourth contribution is the investigation of an effective, new method for classifying dyslexic text, based on the PPM compression method.
- The final contribution is the development and testing of a new system called Sahah to automatically correct Arabic dyslexic text by using PPM text compression scheme and an edit operation approach using compression code length.

1.5 Publications

The researcher has already published one conference paper based on this study. In addition, a further two journal papers have been submitted for publication. Table 1.1 shows specific papers which relate to this study.

The first paper, entitled “A New Error Annotation for Dyslexic Texts in Arabic”, is included in Chapter 4. The paper describes a new classification scheme of errors made in Arabic by people with dyslexia to be used in the annotation of the Arabic dyslexia corpus (BDAC). The dyslexic error classification scheme 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 writers with dyslexia. 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 corpus. The paper was presented at the Third Arabic Natural Language Processing Workshop

Table 1.1: Publications that relate to this study.

1	Title	A New Error Annotation for Dyslexic Texts in Arabic
	Authors	Maha M. Alamri, and William J. Teahan
	In	The Third Arabic Natural Language Processing Workshop (WANLP)
	Publisher	Association for Computational Linguistics (ACL)
	Year	2017
	Status	Published
2	Title	Distinguishing Dyslexic Text from Non-dyslexic Text
	Authors	Maha M. Alamri, and William J. Teahan
	In	Transactions on Computers (TC) Journal
	Publisher	IEEE
	Year	2019
	Status	Submitted
3	Title	Automatic Correction of Arabic Dyslexic Text
	Authors	Maha M. Alamri, and William J. Teahan
	In	Computers Journal
	Publisher	Multidisciplinary Digital Publishing Institute (MDPI)
	Year	2019
	Status	Published

(WANLP) co-located with EACL 2017, held in Valencia, Spain.

The second paper, entitled “Distinguishing Dyslexic Text from Non-dyslexic Text”, which Chapter 5 is based upon, investigates a classification problem, specifically dyslexia text classification, which involves identifying whether or not a text has been written by a person with dyslexia. For this purpose, we apply the PPM text compression scheme for the binary classification problem of distinguishing dyslexic text from non-dyslexic text. Various experiments were conducted to evaluate the method using three corpora. Experimental results show that the accuracy for PPM-based classification significantly outperformed standard feature-based classifiers such as Multi-