Table of Contents

الاهداء I

Acknowledgment II

Abstract III

ملخص V

Introduction 1

1.1 Scope and Motivation 1

1.2 Problem Statement and Thesis Goals 9

1.3 Summary of Contributions 9

1.4 Summary of Structural Work 10

Literature Review 12

2.1 Corpus Collection and Annotation 12

2.1.1 Development of a pilot Levantine Arabic Treebank (LATB) of Jordanian Arabic ... 12

2.1.2 Arabic Dialects Parser 14

2.1.3 Cross Lingual Arabic Blog Alerts (COLABA) 18

2.2 Dialectal Orthography 25

2.3 Dialectal Morphological Annotation 31

2.3.1 MADAMIRA 32

2.4 Remarks on the related works and literature 38

Corpus Collection 40

3.1 Corpus 40

3.1.1 Corpus collection methodology 40

3.1.2 Corpus Statistics 41

3.2 Processing and Storing collected text 50

3.2.1 Storing methodology 50

3.2.2 Storing Schema Statistics 55
Corpus Annotation 57
4.1 Annotation Methodology 57
4.2 Annotation process 64
4.2.1 Using MADAMIRA 64
4.4.2 Manual Annotation 67
4.2.3 Correct errors and fill in gaps using DIWAN 68
4.3 Discussion 70
4.4 Evaluation 72
Conclusion and Future Work 75
References 77
Appendices 83
Appendix #1 84