Reducing Out-of-Vocabulary in Morphology to Improve the Accuracy in Arabic Dialects Speech Recognition

by

Khalid Abdulrahman Almeman

A thesis submitted to
The University of Birmingham
for the degree of
Doctor of Philosophy

School of Computer Science
The University of Birmingham
March 2015

Abstract

This thesis has two aims: developing resources for Arabic dialects and improving the speech recognition of Arabic dialects. Two important components are considered: Pronunciation Dictionary (PD) and Language Model (LM). Six parts are involved, which relate to finding and evaluating dialects resources and improving the performance of systems for the speech recognition of dialects.

Three resources are built and evaluated: one tool and two corpora. The methodology that was used for building the multi-dialect morphology analyser involves the proposal and evaluation of linguistic and statistic bases. We obtained an overall accuracy of 94%. The dialect text corpora have four sub-dialects, with more than 50 million tokens. The multi-dialect speech corpora have 32 speech hours, which were collected from 52 participants. The resultant speech corpora have more than 67,000 speech files.

The main objective is improvement in the PDs and LMs of Arabic dialects. The use of incremental methodology made it possible to check orthography and phonology rules incrementally. We were able to distinguish the rules that positively affected the PDs. The Word Error Rate (WER) improved by an accuracy of 5.3% in MSA and 5% in Levantine.

Three levels of morphemes were used to improve the LMs of dialects: stem, prefix+stem and stem+suffix. We checked the three forms using two different types of LMs. Eighteen experiments are carried out on MSA, Gulf dialect and Egyptian dialect, all of which yielded positive results, showing that WERs were reduced by 0.5% to 6.8%.

Acknowledgements

Many thanks to my supervisor, Dr. Mark Lee for his support, guidance and advice during the PhD period.

Special thanks to my parents, Madhawi Alfanikh and Abdulrahman Almeman for their unfailing love and their encouragement. I am grateful to my wife, Amane Alsaheel for her encouragement, support and quiet patience during my MSc and PhD.

I wish to thank thesis group members, Prof. John Barnden and Dr. Iain Styles for their guidance and advice.

Contents

1 Introduction **1**
- 1.1 Problem statement . . . 4
- 1.2 Research questions . . . 5
- 1.3 A brief description of the work detailed in the thesis . . . 6
- 1.4 Summary of the key thesis contributions . . . 7
  - 1.4.1 Presenting a methodology for building a multi dialect morphology analyser . . . 7
  - 1.4.2 Demonstrating a methodology for collecting multi dialect Arabic text corpora automatically . . . 8
  - 1.4.3 Developing a methodology for building a multi dialect speech corpus 8
  - 1.4.4 Building a multi dialect speech recognition system and comparing it to separated dialects tasks . . . 9
  - 1.4.5 Improving of Arabic dialects PDs . . . 9
  - 1.4.6 Improving of Arabic dialects LMs . . . 9
- 1.5 Thesis outline . . . 10
- 1.6 Resulting publications . . . 12
- 1.7 Resulting dialects resources . . . 14

2 Modern Standard Arabic and Arabic Dialects **15**
- 2.1 Introduction . . . 15
- 2.2 MSA vs. dialects in usage . . . 16
- 2.3 The multiplicity of dialects . . . 16
- 2.4 Dialectic variation . . . 21
- 2.5 Morphology of the dialects . . . 22
- 2.6 Phonology of the dialects . . . 24
- 2.7 Some challenges for Arabic dialects . . . 26
- 2.8 Conclusions . . . 27

3 Related Work **29**
- 3.1 Introduction . . . 29
- 3.2 Automatic multi-dialect analysis of Arabic . . . 29
- 3.3 Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words . . . 31

3.4 Multi dialect speech parallel corpora . . . 34
3.5 A comparison of Arabic speech recognition for multi-dialect vs. specific dialects . . . 37
3.6 An incremental methodology for improving pronunciation dictionaries for Arabic . . . 40
3.7 Morpheme-based language models for improving the speech recognition of Arabic dialects . . . 41
3.8 Conclusions . . . 42

4 Automatic Multi-Dialect Analysis of Arabic 44
4.1 Introduction . . . 44
4.2 The motivations for multi-dialect morphology analysis . . . 45
4.3 Methodology . . . 46
4.4 Implementation . . . 49
4.4.1 The building of Arabic multi dialect morphology analyser webpage 52
4.5 Evaluation . . . 54
4.6 Conclusions . . . 58

5 Automatic Building of Arabic Multi Dialect Text Corpora by Bootstrapping Dialect Words 60
5.1 Introduction . . . 60
5.2 The motivations for building a multi dialect written text corpora . . . 61
5.3 Methodology . . . 62
5.4 Results . . . 67
5.5 Evaluation . . . 68
5.5.1 Comparing results . . . 68
5.5.2 Analysis . . . 70
5.5.3 Error evaluation . . . 76
5.6 Conclusions . . . 76

6 Multi Dialect Arabic Speech Parallel Corpora 78
6.1 Introduction . . . 78
6.2 The need for Arabic multi dialect speech corpora . . . 79
6.3 Methodology . . . 80
6.4 Implementation . . . 82
6.4.1 Write MSA text and diacritise it . . . 83
6.4.2 Translate into dialects and diacritise them . . . 83
6.4.3 Recording . . . 84
6.4.4 Audio segmenting . . . 85
6.5 File organisation . . . 85
6.6 Results . . . 86
6.6.1 Parallel texts results . . . 86
6.6.2 Parallel speech results . . . 86

6.7 Evaluation . . . 88
6.7.1 Text evaluation . . . 88
6.7.2 Speech evaluation . . . 90
6.8 Conclusions . . . 93

7 A Comparison of Arabic Speech Recognition for Multi-Dialect vs. Specific Dialects 95
7.1 Introduction . . . 95
7.2 Data . . . 97
7.3 Recognition system . . . 97
7.4 Results . . . 98
7.5 Discussion . . . 101
7.6 Conclusions . . . 103

8 An Incremental Methodology for Improving Pronunciation Dictionaries for Arabic Speech Recognition 105
8.1 Introduction . . . 105
8.2 Why do we need to improve Arabic pronunciation dictionaries? . . . 106
8.3 Data . . . 107
8.4 Methodology . . . 107
8.4.1 The pronunciation dictionary rules . . . 109
8.4.2 The incremental methodology for improving Arabic pronunciation dictionary . . . 110
8.5 Recognition system and baseline result . . . 111
8.6 Results . . . 113
8.7 Evaluation . . . 115
8.8 Conclusions . . . 116

9 Morpheme-Based Language Models for Improving the Speech Recognition of Arabic Dialects 118
9.1 Introduction . . . 118
9.2 Data . . . 120
9.3 Methodology . . . 122
9.4 Recognition system and baseline result . . . 123
9.5 Automatic Speech Recognition (ASR) experiments results . . . 124
9.6 Evaluation . . . 125
9.7 Conclusions . . . 127

10 Conclusions and Future Work 129
10.1 Introduction . . . 129
10.2 The methodologies for building Arabic dialects resources . . . 130
10.3 Improving speech recognition for Arabic dialects . . . 133
10.4 How this research can be extended to multi-dialect approaches to other languages . . . 135

10.5 Future work 137
10.5.1 Extending the multi dialect analyser 137
10.5.2 Classifying text corpora 138
10.5.3 Producing different version of speech corpora 138
10.5.4 Extending incremental methodology by applying it to other dialects 138
10.5.5 Returning to the original full word from prefix+stem or stem+stem after improving LMs 138
10.6 Contributions 139

List of References **140**

List of Tables

2.1 Arabic letters 17
2.2 The 13 combinations of short vowels for ب /b/ 'Baa letter' 18
2.3 Percentage of shared unigrams, bigrams and trigrams in the Egyptian corpus (ECA) and the MSA corpus, and for the conversational British English corpus (BE) and American English corpus (AmE) (Kirchhoff and Vergyri, 2005) 19
2.4 Some changes in phones between Arabic dialects compared with MSA . . . 22

3.1 Some existing MSA corpora 33
3.2 Some existing MSA speech corpora 36
3.3 Gulf speech corpora 37
3.4 Levantine speech corpora 38
3.5 Egyptian speech corpora 39
3.6 Microphone source speech corpora for MSA, Gulf, Egyptian and Levantine 39

4.1 Example of the output after the first layer 51
4.2 Example of segmented words 52
4.3 Example of analysed words after the last layer 54
4.4 Results before starting the experiments 55
4.5 Results after MSA analyser has adopted 55
4.6 The final results 56

5.1 Examples of categorised words and phrases 64
5.2 Total number of words 65
5.3 The estimation of how many pages we need per dialect 65
5.4 Results 67
5.5 Total results 67

5.6 Sentences counts and average of length . . . 68
5.7 Comparing unknown words . . . 69
5.8 Size comparisons . . . 69
5.9 Frequency of frequencies of token types . . . 70
5.10 10 greatest unigrams tokens frequencies (including function words) . . . 71
5.11 10 greatest unigrams tokens frequencies (non function words) . . . 72
5.12 Bigrams, trigrams and five-grams counts . . . 74
5.13 Commonest bigram for all dialects corpora . . . 75
5.14 Commonest trigram for all four corpora . . . 75

6.1 New phones representation in dialects . . . 81
6.2 Corpora distribution for sections . . . 83
6.3 Example of some sentences . . . 84
6.4 Recording attributes . . . 84
6.5 Tokens count . . . 86
6.6 Parallel texts sentences count . . . 86
6.7 Speaker count . . . 86
6.8 Speaker age . . . 87
6.9 Files count . . . 87
6.10 Utterances count . . . 87
6.11 Phones count . . . 88
6.12 Sharing sentences between four parallel corpora . . . 89
6.13 The word overlap for MSA with Gulf, Egyptian and Levantine . . . 89
6.14 Lexicon count . . . 90
6.15 Speech contrast evaluation . . . 92

7.1 One-to-one auto mapping for creating a baseline PD . . . 98
7.2 The WER variation with the tied states and the number of densities in multi-dialect system using all four corpora . . . 99
7.3 The best WERs for the four dialects when evaluated using multi-dialect data 100
7.4 The best WERs for the four dialects when evaluated against each dialect's own data . . . 100
7.5 The table shows Levantine dialect results when evaluated using MSA acoustic model with three different LMs . . . 101
7.6 A Student's t-test result . . . 102

8.1 An one-to-one automapping for creating baseline PD . . . 108
8.2 PD phonology and morphology rules . . . 111
8.3 MSA Result . . . 113
8.4 Levantine Result . . . 114

9.1 Description of the corpora used for creating closed and open LMs . . . 121
9.2 Reduction percentage in closed LM size- unique tokens . . . 121
9.3 Reduction percentage in open LM size- unique tokens . . . 121

9.4 Comparison of LM sizes 122
9.5 Baseline results 123
9.6 MSA recognition results 124
9.7 Gulf dialect recognition results 125
9.8 Egyptian dialect recognition results 125

List of Figures

1.1 An architecture of a simplified ASR system for decoding a sentence, from Jurafsky and Martin (2009) . . . 2
1.2 Standard n-gram backoff path for a 4-gram language model over words (a) and backoff graph for 4-gram over factors (b), from Kirchhoff et al. (2006) 4
1.3 Thesis parts . . . 6

2.1 Vocabulary growth for full words in Egyptian Colloquial Arabic (ECA), from Kirchhoff et al. (2003) . . . 23

4.1 How words be analysed by using the algorithm . . . 53

5.1 Zipf's law of Egyptian, slope = - 0.9761. log of rank on the X-axis versus frequency on the Y-axis . . . 72
5.2 Zipf's law of Gulf, slope = -0.9759. log of rank on the X-axis versus frequency on the Y-axis . . . 73
5.3 Zipf's law of Levantine, slope= -0.972. log of rank on the X axis versus frequency on the Y-axis . . . 73
5.4 Zipf's law of North African, slope = -0.9793. log of rank on the X-axis versus frequency on the Y-axis . . . 74

6.1 An example of how the files are organised . . . 85
6.2 wave example . . . 91

8.1 The Incremental Methodology Algorithm for improving Arabic PD in ASR 112

Acronyms

ASR Automatic Speech Recognition

CCA Corpus of Contemporary Arabic

EER Equal Error Rate

FLM Factored Language Model

GMM Gaussian Mixture Model

HMM Hidden Markov Model

KACST King Abdulaziz City for Science and Technology

LM Language Model

MADA Morphological Analysis and Disambiguation of Arabic

MSA Modern Standard Arabic

NLP Natural Language Processing

OOV Out-Of-Vocabulary

PD Pronunciation Dictionary

PDF Probability Density Function

POS Part-Of-Speech

SAAVB Saudi Accented Arabic Database

STER Stem Error Rate

TTS Text To Speech

WCAG Web Content Accessibility Guidelines

WER Word Error Rate

Chapter 1

Introduction

The goal of Automatic Speech Recognition (ASR) research is to build systems that map from acoustic signals to a string of words computationally (Jurafsky and Martin, 2009). Acoustic models, language models and pronunciation dictionaries -Lexicons- are the most important components of ASR systems; they work together to form a recogniser. Figure 1.1 shows a simplified ASR system, describing decoding a sentence, which begins with a wave and ends in a string of words, passing through the three main components listed above. The decoder combines the probabilities of a language model -n-gram-, the Hidden Markov Model (HMM)[1] of each word in the lexicon[2] and acoustic likelihood. The decoder in this model uses a Bayes' rule, which is represented as follows:

$$W^* = \arg\max_{W} \overbrace{P(O|W)}^{Acoustic\ model} \overbrace{P(W)}^{Language\ model}$$

Word Error Rate (WER) is the standard evaluation metric used for ASR systems (Jurafsky and Martin, 2009). To calculate the WER we must specify the number of

[1] Rabiner and Juang (1986) define an HMM as "a doubly stochastic process with an underlying stochastic process that is not directly observable (it is "hidden"), but can only be observed through another set of stochastic processes that produce the sequence of observed symbols".

[2] An HMM lexicon in Figure 1.1 is "a list of word pronunciations, each pronunciation represented by a string of phones. Each word can then be thought of as an HMM, where the phones (or sometimes subphones) are states in the HMM, and the Gaussian likelihood estimators supply the HMM output likelihood function for each state" (Jurafsky and Martin, 2009).

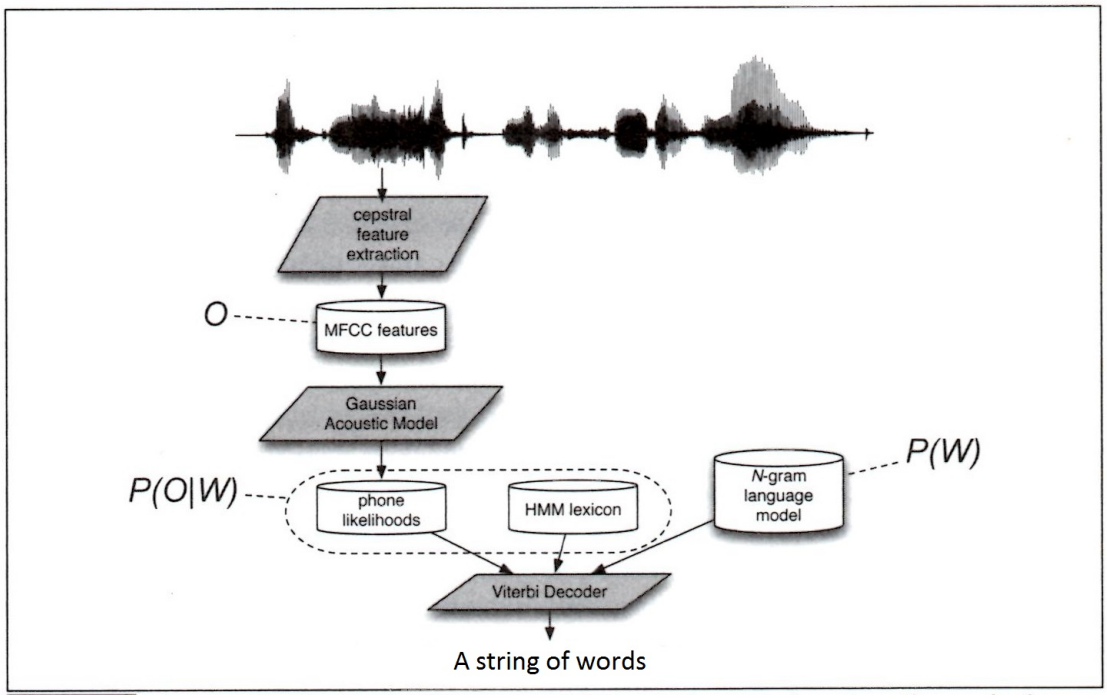

Figure 1.1: An architecture of a simplified ASR system for decoding a sentence, from Jurafsky and Martin (2009)

insertions, substitutions and deletions as follows:

$$\text{WER}\% = 100 \times \frac{\text{Insertions} + \text{Substitutions} + \text{Deletions}}{\text{Total Words in Transcript}}$$

Arabic has three major challenges that affecting ASR in two of its main components the Language Model (LM) and the Pronunciation Dictionary (PD)[3]: (1) the richness of the morphology, e.g. hundreds words can be generated from every single root; (2) the huge variations between Modern Standard Arabic (MSA) and dialects and between dialects themselves; and (3) the limited available resources that can assist in NLP and ASR tasks, such as different types of corpora, analytical tools, Part-Of-Speech (POS) taggers, dialects diacritisers, etc.

Two important factors that affect ASR directly relate to the language's rich morphol-

[3]These challenges will be discussed in greater detail in Chapter 2.

ogy: the LM and the PD. These two important components primarily relate to NLP, and this explains the relationship in this thesis between NLP and ASR. The LM and the PD considerably affect the accuracy of the speech recogniser. In a PD, every single word is represented phonetically. In the case of morphological languages that are not rich, such as English, it is not hard to add all or the majority of full word forms to every stem; for example there are four different full word forms of the stem "wait"; i.e. wait, waits, waiting and waited, all of these forms can be found in the English Voxforge lexicon and the CMU pronunciation dictionary. However, this is very difficult with Arabic, in which for the same stem too many different words can be generated. For this reason, it is very difficult to produce an Arabic PD manually. Therefore, an alternative approach is proposed. This involves producing an automatic PD for Arabic MSA and dialects. The other concern with Arabic PDs is how best to produce a phonetic representation of every possible word automatically; this concern is most associated with dialects in which there are no clear linking rules to apply from orthography to phonology.

MSA mainly follows a free word order (Staal, 1967; Farghaly and Shaalan, 2009). Examples of sentence structure are as follows: subject-verb-object, verb-subject-object or object-verb-subject. Some studies, such as Kirchhoff et al. (2006), have focused on this issue when building a LM, particularly with regard to how the decoder deals with the different orders of words, Figure 1.2 compares (a) a standard n-gram model with (b) Factored Language Model (FLM), which deals specifically with word order. This thesis focuses on the word-stem level, using the standard n-gram backoff path, i.e. (a) in Figure 1.2, without dealing with word order. The main reason for not dealing with word order is that SVO order is more commonly in use with Arabic dialects, (Chiang and Rambow, 2006; Farghaly and Shaalan, 2009) unlike MSA.

In view of the above, the main work of this thesis will focus on improving LM and PD from the perspective of inflectional approaches to reduce WER and improve accuracy in

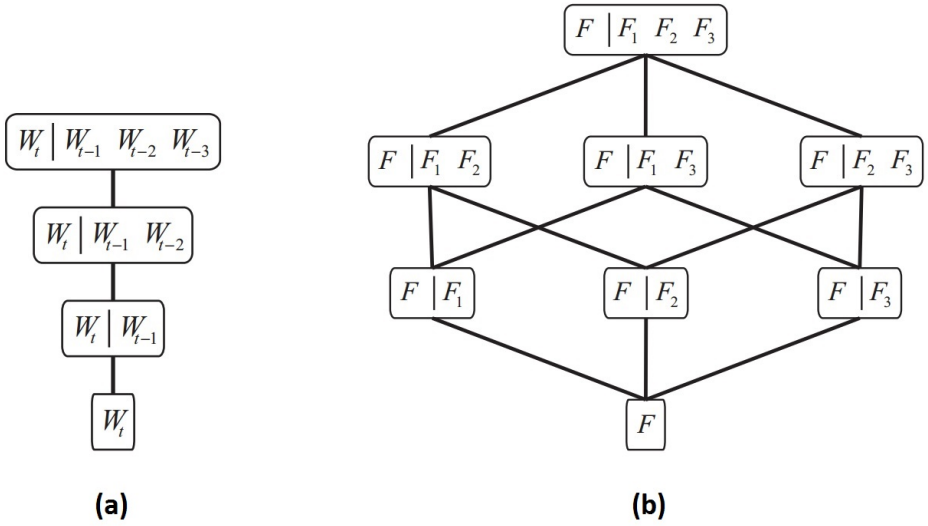

Figure 1.2: Standard n-gram backoff path for a 4-gram language model over words (a) and backoff graph for 4-gram over factors (b), from Kirchhoff et al. (2006)

Arabic dialects speech recognition tasks. However, before commencing work on PDs and LMs, there is a need to collect and build the requisite resources, which are very important in such tasks.

1.1 Problem statement

The problem raised and addressed by this thesis is twofold. The first part of the problem focuses on the lack of dialect resources and tools, which adversely affect Arabic dialects for performing ASR and NLP tasks. The absence of a multi dialect morphology analyser, which is a very important tool for dialect tasks, has led previous researchers to use either an annotated lexicon, which is unavailable in most dialects, or to extract stems and affixes manually, which is time-consuming for large tasks. The second issue is the general absence of large corpora for the dialects, though there are small corpora in the cases of some dialects. Dialect text corpora are very helpful when creating LM for dialects ASR. The final issue affecting dialect resources is the absence of parallel speech corpora. This means it is not possible to compare between Arabic dialects or extract comparable results

between Arabic dialects.

The other part of the problem, which will be explained in detail in this thesis, addresses Arabic dialect morphology in ASR tasks. In the case of Arabic dialects, it is very difficult to produce all full word forms manually, as hundreds of words can be generated from every single root. Moreover, to date, there has been no previous study clarifying how best to link orthography to phonology in Arabic dialects to create an automatic PD. The research that has been done either uses a manual lexicon, which is unavailable for most dialects, or uses a few unspecified rules to link orthography to phonology. Very few full word forms appear in Arabic dialect LMs, leading many Out-Of-Vocabulary (OOV)[4] words to arise. This then makes the WER increase for the recogniser.

1.2 Research questions

The research questions posed in this thesis are as follows:

- Question 1: How can we improve Arabic dialect speech recognition accuracy by representing the rich morphology of Arabic dialects in two main components; PDs and LMs?

- Question 2: How do parallel multi dialect speech corpora affect the accuracy of Arabic multi dialect speech recognition tasks? In addition, which type of data is best for dialects speech recognition systems; pooled data or data separated using a dialect classifier?

- Question 3: How can an Arabic multi dialect text corpora be collected automatically?

- Question 4: How can a multi dialect morphology analyser using linguistic and statistical methods be built?

[4]Unseen words.

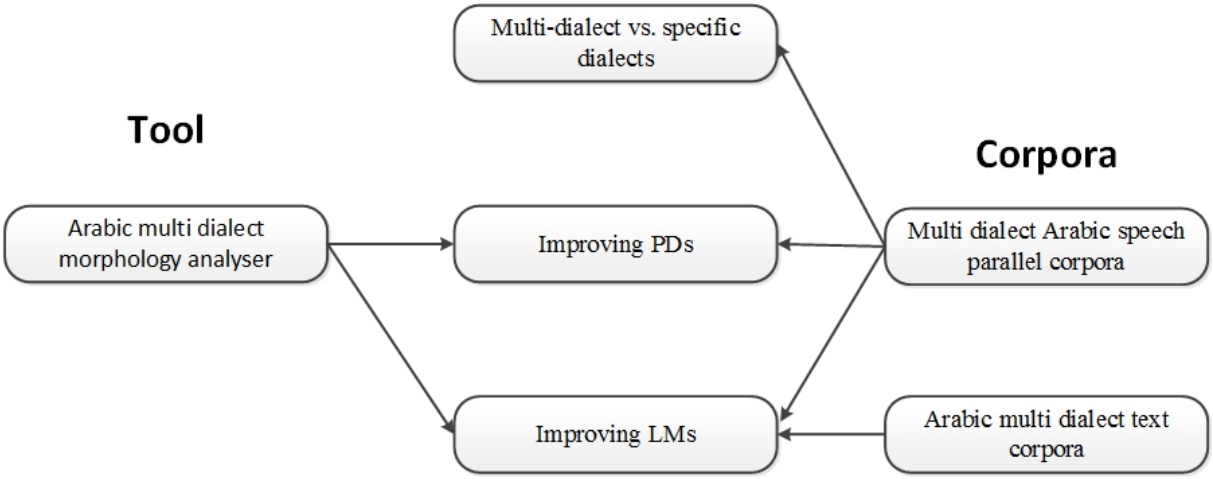

Figure 1.3: Thesis parts

1.3 A brief description of the work detailed in the thesis

This thesis focuses on two aims; developing appropriate resources for Arabic dialects and improving Arabic dialects speech recognition systems for two important factors, i.e. PD and LM. Figure 1.3 shows the six elements integrated into the thesis. Three parts out of six relate to resources, and three relate to improving applications of speech recognition dialects.

Three important components were involved when compiling Arabic dialects resources, as Figure 1.3 shows, i.e. (1) building multi dialect morphology analyser, (2) collecting a multi dialect Arabic text corpora, and (3) building a multi dialect Arabic speech corpus.

Once we finished building and collecting the required resources, we commenced building the multi dialect speech recognition task using a resulted speech corpus. In this task, we aimed to compare the results of multi dialect tasks with separated tasks, and sought to compare multi dialect speech recognition tasks with separated dialects tasks that use a dialect classifier. We referred to speech corpora to build and test this, as Figure 1.3 shows. In all the experiments that have been carried out in this thesis we used trigram

language models.

The improvement of the PDs and LMs for Arabic dialects was a main objective when conducting the work for this thesis. An incremental methodology was used to improve PDs for Arabic MSA and Levantine. This methodology made it possible to check existing orthography and phonology rules incrementally; such that any rule yielding improvement is added, and others will be ignored. The orthography to phonology rules are for MSA. Therefore, by using this methodology we were able to distinguish which of these rules affected PD dialect positively and which negatively. Two components were used for improving PDs, i.e. multi dialect speech corpora and multi dialect morphology analyser, as seen in Figure 1.3.

Many of the previous studies have aimed to improve LM in one of two ways; using either morphemes (word level) or word order (sentence level). The scope of this thesis was derived at the word level; therefore, we used morpheme to improve LMs for dialects. For this purpose we suggested three levels of morphemes; stem alone, prefix+stem and stem+suffix. We also checked the three forms using two different types of LMs; closed domain LM, and open domain LM. All three resultant resources i.e. speech and text corpora and multi dialect morphology analyser were used in this portion of the thesis, as Figure 1.3 shows.

1.4 Summary of the key thesis contributions

Below we list the contributions to knowledge resulting from this thesis:

1.4.1 Presenting a methodology for building a multi dialect morphology analyser

We addressed the problem of the analysis of a multi-dialect Arabic morphology. The methodology applied involved the synthesis of two methods. The linguistic method used

an adapted MSA morphology analyser to first handle the dialect prefixes and suffixes and then analyse the remaining word fragments. This method improved the accuracy of the dialect words from 32% to 69%. The second method involved segmenting the words and using 'the web as corpus' to estimate the frequency of different segment combinations, which were then used to guess the correct stem form and extract affixes where there were some. The overall synthesis was shown to have a 94% accuracy in a corpus of Arabic dialects (Chapter 4).

1.4.2 Demonstrating a methodology for collecting multi dialect Arabic text corpora automatically

The work described involved building multi dialect Arabic texts corpora automatically using the web as a corpus. From the results of the experiments, we obtained about 50 million tokens from the different Arabic dialects. These dialects included the four main dialects, i.e. Gulf, Levantine, Egyptian and North African, resulting in 14.5 million, 10.4 million, 13 million and 10.1 million tokens being obtained respectively. The total number of distinctive types across all the corpora was 2 million types (Chapter 5).

1.4.3 Developing a methodology for building a multi dialect speech corpus

The design and recording of a multi-dialect Arabic speech parallel corpus, which encompasses three main dialects; Gulf, Egyptian and Levantine as well as MSA, was undertaken for a specific linguistic domain i.e. travel and tourism. Travel and tourism domain has a clear vocabulary, which can be pronounced by any native speaker easily. Parallel texts were written for the four main dialects, involving 1291 sentences for MSA and 1069 sentences for other dialects. The recordings were conducted involving 52 participants. We obtained about 32 speech hours, which were then segmented into a total number of 67,132

speech files (Chapter 6).

1.4.4 Building a multi dialect speech recognition system and comparing it to separated dialects tasks

A comparison of Arabic ASR systems for specific dialects, versus a system trained using pooled data was done. The comparison covered three different dialects in addition to MSA. The best Word Error Rate (WER) that could be obtained for the multi-dialect recognition system was 13.7% and the average for the best WERs from the four dialects was extracted and found to be 10.2%. Thus, the difference was found to be -3.5%. Three different aspects were involved in this comparison; recognition accuracy, time required and accuracy of the dialect classifier (Chapter 7).

1.4.5 Improving of Arabic dialects PDs

A novel method was used to improve the accuracy of Arabic dialect PDs. In the incremental methodology, we showed how incremental cycles could be applied to Arabic speech recognition to reduce WER. By using phonology and morphology rules, an incremental methodology was used to improve the PDs for MSA and the Levantine dialect. The absolute accuracy of MSA PD is improved by 5.3% and 5% for the Levantine dialect. Eight phonology rules out of 11 improved the MSA PD and three rules improved Levantine PD. We obtained two PDs for MSA and the Levantine dialect to show how each orthography to phonology rule improved the accuracy of the speech recogniser (Chapter 8).

1.4.6 Improving of Arabic dialects LMs

Experiments were done to improve the LMs for three parallel dialects. In each dialect, two different LMs were produced: a closed domain LM and an open domain LM. The methodology of the second part of the multi dialect morphology analyser, involved retrieval of web frequencies for different parts of a word; this methodology was modified

and then used to extract the three suggested forms of the word; stem alone, prefix+stem and stem+suffix. Six results were then extracted per dialect, giving a total of eighteen results. The error rates in these results have been reduced between 0.5% to 6.8% (Chapter 9).

1.5 Thesis outline

Chapter 2 introduces the general features of the Arabic language. It also looks at the Arab world and dialectic variation. The chapter focuses on some of the challenges associated with Arabic dialects in relation to speech processing. Three challenges will be discussed: (1) a lack of standardisation in Arabic dialects, (2) the large gap between MSA and the dialects and between the dialects themselves, and (3) the shortage of available resources for Arabic dialects.

Chapter 3 provides an overview of relevant work. Related work has been categorised into two main sections; the resources and the improving of dialects speech recognisers. In the resources section, we will discuss three important resources; (1) the morphology analyser tool, (2) the texts corpora, and (3) the speech corpora for Arabic dialects. The other section of the related work will discuss building on and improving Arabic dialects speech recognition systems in two main categories, i.e. PDs and LMs.

Chapters 4, 5 and 6 discuss the building and the collection of the required Arabic dialect resources. Chapter 4 describes a multi-dialect morphology analyser, which utilises both a linguistic basis and a statistical basis upon which to analyse Arabic dialects, loanwords and MSA words that were not analysed by the MSA morphology analyser. The linguistic foundation utilises the MSA morphology analyser and adapts it to accept different dialects' affixes. Using the four forms created by the segmenter the notion of using the web as a corpus was proposed. The web resource makes it possible to use the frequencies retrieved for each word segment to distinguish the stem and then extract the

affixes where applicable.

Chapter 5 and 6 explain the methodologies used for building the text and speech corpora in Arabic dialects. Chapter 5 presents the Arabic multi dialect text corpora, which was created by exploiting the web as a corpus. Four steps involved in creating the corpora will be explained; the collection and grouping of around 1500 dialect words from different Arabic websites. Then, a survey was conducted with a group of people from different Arab countries to ensure that they would use only the words on the designated list. Next, the researcher collected the links using Bing API and downloaded web pages, likely to have the same dialects. The final stage was the cleaning and normalisation of the downloaded web pages. Our aim was that the resulting corpus should include four main dialects; Levantine, Egyptian, Gulf and North African.

In Chapter 6, we introduce the methodology we followed when building the parallel dialects speech corpora for Arabic. It shows the steps performed when writing the MSA text and then diacritising it. The next step is for dialect native speakers to translate the MSA diacritised text into the local dialects. In the recording step, we show how we obtained high quality recordings. The resultant corpora have four parallel speech corpora; MSA, Gulf, Levantine and Egyptian. There is also an MSA numbers speech corpus, produced by native Arabic speakers from different dialect backgrounds. The data resulting from both the speech, in Chapter 6, and text, in Chapter 5, is available for public use.

Chapter 7 describes the building of a multi dialect system before starting work on PDs and LMs to check the accuracy of the speech data collected, and to compare the multi dialect speech recognition task results, versus the separated dialects results. Therefore, Chapter 7 presents an Arabic multi-dialect speech recognition system created to recognise MSA and three different dialects.

Chapters 8 and 9 introduce methodologies for improving dialects speech recognition

systems in two main components; PDs and LMs. Chapter 8 introduces an incremental method for improving Arabic PDs, where orthography to phonology in Arabic, in general follows regular rules. However, these rules still needed to be checked to ascertain efficiency. Therefore, an incremental methodology for applying phonological rules is introduced in this chapter for MSA and Levantine dialects. This allows the application of each new rule to the baseline system, initiating a new cycle to check the effect of the additional rule. If there was an improvement, we added this rule to our PD improvement rules for each dialect.

Chapter 9 continues working to improve the dialect recogniser by working on the different dialect stems in the language models. We show the results of the experiments conducted on three parallel dialects; i.e. MSA, Gulf and Egyptian. Two different LMs were applied for each dialect; a closed domain LM and an open domain LM. Chapter 9 also explained how the methodology of the second part of the multi dialect morphology analyser was used to extract the three suggested forms of the word; stem alone, prefix + stem and stem + suffix.

Chapter 10 introduces the conclusions to this thesis, and presents suggestions about future work. Chapter 10 summarises the contributions made in this thesis.

1.6 Resulting publications

Seven papers resulted from this thesis, i.e. two articles and five conference papers as follows:

Articles:

1. Khalid Almeman and Mark Lee. Automatic multi-dialect analysis of Arabic. *Linguistica Communicatio: International journal of Arabic language engineering & General Linguistics*, 5:95–108, 2013 (Almeman and Lee, 2013d).

2. Khalid Almeman and Mark Lee. Building a Multi-Dialect Morphological Analyser

for Arabic. *The Journal of Computer Science and Engineering, in Arabic (IJCSEA)*, 5(1):74–92, 2013 (Almeman and Lee, 2013e).

Conference Papers:

1. Khalid Almeman and Mark Lee, "A Comparison of Arabic Speech Recognition for Multi-Dialect vs. Specific Dialects", In *Proceedings of the Seventh International Conference on Speech Technology and Human-Computer Dialogue (SpeD 2013)*, Cluj-Napoca, Romania, 16-19 October 2013, 2013 (Almeman and Lee, 2013a).

2. Khalid Almeman and Mark Lee, "An Incremental Methodology for Improving Pronunciation Dictionaries for Arabic Speech Recognition", In *Proceedings of the Seventh International Conference on Speech Technology and Human-Computer Dialogue (SpeD 2013)*, Cluj-Napoca, Romania, 16-19 October 2013, 2013 (Almeman and Lee, 2013b).

3. Khalid Almeman and Mark Lee, "Automatic Building of Arabic Multi Dialect Text Corpora by Bootstrapping Dialect Words", In *Proceedings of the First International Conference on Communications, Signal Processing, and their Applications (ICCSPA'13)*, Sharjah, UAE, 12-14 Feb. 2013, IEEE, 2013 (Almeman and Lee, 2013c).

4. Khalid Almeman and Mark Lee and Ali Almiman, "Multi Dialect Arabic Speech Parallel Corpora", In *Proceedings of the First International Conference on Communications, Signal Processing, and their Applications (ICCSPA'13)*, Sharjah, UAE, 12-14 Feb. 2013, IEEE, 2013 (Almeman et al., 2013).

5. Khalid Almeman and Mark Lee, "Towards Developing a Multi-Dialect Morphological Analyser for Arabic", In *Proceedings of the Fourth International Conference*

on Arabic Language Processing (CITALA'12), Rabat, Morocco, 2-3 May 2012, PP. 19-25, 2012 (Almeman and Lee, 2012).

1.7 Resulting dialects resources

Three resources resulted from this thesis. These resources have been made available to other researchers. They are one tool and two corpora:

1. The Arabic multi-dialect morphology analyser; The multi dialect morphology analyser follows the algorithm that has been suggested in Chapter 4. The multi dialect morphology analyser can be used online through the webpage as a web service, http://www.arabicmorphologyanalyser.com, or using its source code.

2. The multi dialect Arabic speech parallel corpora project: this project has 32 speech hours recorded by 52 participants for MSA, Gulf, Egyptian and Levantine. It also has an MSA number corpus recorded using different dialect backgrounds. The resultant corpora totals more than 67000 wave files. Fully diacritised texts and PDs have also been prepared for this project.

3. The Arabic multi dialect text corpus project; this corpus was extracted automatically from the web. The multi dialect text corpus comprises four sub-corpora, categorised into the main dialects, i.e. Gulf, Egyptian, Levantine and North Africa. In total, the multi dialect text corpora has 50 million tokens.

Chapter 2

Modern Standard Arabic and Arabic Dialects

2.1 Introduction

The Arabic language is a member of the Semitic language family. It is the fourth most commonly spoken language across the world after Chinese, Spanish and English (CIA, 2013), with an estimated more than 422 million speakers in 2012 (Bokova, 2012). Some references such as Ethnologue (2013)[1] and CIA (2013) have divided the number between native speakers and those who speak Arabic as a second language. Ethnologue (2013) estimates the number of native speakers as 206 million and second-language speakers at 246 million, totalling 452 million speakers between the two categories. The Arabic language is an official language in 24 countries (CIA, 2013).

Written from right to left, the Arabic language has 28 letters. Arabic letters have up to four forms according to their place in the word; i.e. isolate, beginning, middle or ending. Table 2.1 lists the Arabic letters' names and how they have different forms according to their positions within the word; it also shows how each letter is pronounced when spoken

[1]Ethnologue is a web-based publication which contains statistics on more than 7000 languages (Ethnologue, 2013).

separately.

Diacritisation in Arabic shows the short vowel. There are up to 13 different combination forms for the Arabic letter. An example of these multiple forms for the different short vowels can be seen in Table 2.2, which shows the 13 combinations of short vowels for one Arabic letter ب /b/ 'Baa letter'. Overall, more than 350 different possible diacritised letters exist for the 28 Arabic letters.

With the exception of some religious and learning books, diacritisation (as depicted above) is not written in Arabic texts (Boudelaa and Marslen-Wilson, 2010). The absence of diacritisation leads to lexical and morphological ambiguity (Diab et al., 2007) and is one of the most critical challenges for processing Arabic texts (Elshafei et al., 2006). This is for Modern Standard Arabic (MSA) which, in most cases, has regular rules for writing; this challenge is more formidable when dealing with dialects which are irregular in many cases.

2.2 MSA vs. dialects in usage

MSA is the formal language of communication understood by most Arabic people. It is used in education and on different media (Clive, 2004), to communicate between people of different dialects, in courtrooms, and in other formal situations. Conversely, dialects are used in daily conversation and telephone communication and, in recent times, have begun to be used on both television and radio and also as written forms on the Internet. Unless there is a requirement for MSA, the local dialect is more commonly used.

2.3 The multiplicity of dialects

> A variety is an any body of human speech patterns which is sufficiently homogeneous to be analysed by available techniques of synchronic description and which has a sufficiently large repertory of elements (Ferguson, 1971).

Isolated	Name	IPA	Beginning	Middle	Final
أ[a]	Alif	/a/	أ	ـأ	ـأ
ب	Baa	/b/	بـ	ـبـ	ـب
ت	Taa	/t/	تـ	ـتـ	ـت
ث	Thaa	/θ/	ثـ	ـثـ	ـث
ج	Jeem	/ʤ/	جـ	ـجـ	ـج
ح	Haa	/ħ/	حـ	ـحـ	ـح
خ	Khaa	/x/	خـ	ـخـ	ـخ
د	Dal	/d/	د	ـد	ـد
ذ	Dhal	/ð/	ذ	ـذ	ـذ
ر	Raa	/r/	ر	ـر	ـر
ز	Zaa	/z/	ز	ـز	ـز
س	Seen	/s/	سـ	ـسـ	ـس
ش	Sheen	/ʃ/	شـ	ـشـ	ـش
ص	Saad	/sˤ/	صـ	ـصـ	ـص
ض	Daad	/dˤ/	ضـ	ـضـ	ـض
ط	Taa	/tˤ/	طـ	ـطـ	ـط
ظ	Dhaa	/ðˤ/	ظـ	ـظـ	ـظ
ع	Aain	/ʕ/	عـ	ـعـ	ـع
غ	Ghain	/ɣ/	غـ	ـغـ	ـغ
ف	Faa	/f/	فـ	ـفـ	ـف
ق	Qaf	/q/	قـ	ـقـ	ـق
ك	Kaf	/k/	كـ	ـكـ	ـك
ل	Lam	/l/	لـ	ـلـ	ـل
م	Meem	/m/	مـ	ـمـ	ـم
ن	Noon	/n/	نـ	ـنـ	ـن
ه	Haa	/h/	هـ	ـهـ	ـه
و	Waw	/w/	و	ـو	ـو
ي	Yaa	/j/	يـ	ـيـ	ـي

[a]The Alif Letter has eight different forms for writing i.e. ا أ آ إ ى ئ ء ؤ. Some forms are similar to the Waw letter, for example, ؤ /o/, and some forms are similar to the Yaa letter.

Table 2.1: Arabic letters

FatHah	Kasrah	Dhammah	Sukwn
بَ	بِ	بُ	بْ
بً	بٍ	بٌ	
بَّ	بِّ	بُّ	
بًّ	بٍّ	بٌّ	

Table 2.2: The 13 combinations of short vowels for ب /b/ 'Baa letter'

All languages have one or more dialects. Carter et al. (2011) defined a dialect as "a form of the language that is spoken in a particular part of the country or by a particular group of people". Any linguistic practices that are characteristic of specific socioeconomic groups, gender and age groups constitute dialects (Finegan, 2008). The variability between dialects includes differences in words, grammar, morphology, syntax, phonetics, etc.

The gap between dialects and the standard language or between dialects themselves differs from one language to other. For example a large gap can be found between Arabic dialects comparative to English dialects; an interesting experiment conducted by Kirchhoff and Vergyri (2005) computed overlap in vocabulary between dialects as a percentage of shared unigrams, bigrams and trigrams for Egyptian dialect and MSA, and compared these numbers with equivalent statistics for American English and British English, Table 2.3. In Arabic, the inventory of words (unigrams) only overlaps by 10% and there is a minimal overlap of bigrams and trigrams. One can compare this with English, in which the percentage of shared unigrams was found to be 44% and approximately 20% and 5% for bigrams and trigrams, respectively. The overlap in unigrams is indicative of the gap size in words and morphology, and the overlap in bigrams and trigrams indicates gap size in syntax and grammar.

As the gap between dialects grows, dealing with more than one dialect in ASR tasks becomes harder, and a lower accuracy rate is likely. In the case of a large gap, multi di-

	ECA-MSA	BE-AmE
Shared unigrams (%)	10.3	44.5
Shared bigrams (%)	1	19.2
Shared trigrams (%)	<1	5.3

Table 2.3: Percentage of shared unigrams, bigrams and trigrams in the Egyptian corpus (ECA) and the MSA corpus, and for the conversational British English corpus (BE) and American English corpus (AmE) (Kirchhoff and Vergyri, 2005)

alect tasks can be treated as different languages. The matter of the gap between dialects becomes more challenging in languages that are rich morphologically, as the rich morphology presents a new angle in addition to the differences in syntax, grammar, words, etc.

In some languages, there is agreement between concerning what is standard language, but in others, there is no agreement or one standard. For example, there is no agreement about which is the standard language in English; Finegan (2008) states, "no single variety of English can be called the standard. Here it should also be remembered that there are different national standards for British, American, Australian, and Canadian English (among others). Furthermore, at least with respect to pronunciation, there may be several standard varieties of a national variety. The simple fact is that many varieties of standard English exist.". Arabic is a different case, as MSA is the Arabic standard. MSA is a formal language used to communicate between speakers of different dialects; this makes it a second dialect for Arab speakers.

The differences in patterns in some classes of POS sometimes follow regular rules from standard to dialect or between dialects. One example of this is verbs in Arabic dialects; according to Haak (1996), in many cases the stem patterns of Arabic verbs in dialects are identical to those of MSA in many cases. However, this regular pattern in Arabic is not seen with loanwords- noun or verbs-, adjective, adverb, names, etc. Regular patterns between dialects are very helpful in NLP; for example, the accuracy of our multi dialect morphology analyser, in Chapter 4, has been increased from 32% to 68%, when adding

new affixes from dialects to enable the MSA analyser to analyse verbs used in dialects.

Arabic dialects multiplicity

Most current Arabic dialects were generated from the interaction between different old dialects of Arabic and other languages that existed in neighbouring regions (Habash, 2010). For example, both the Berber and French languages have influenced the North African dialect (*ibid.*).

The majority of words in dialect originated from either MSA, such as the expression أُولِت لُه /Aulit luh/ (Levantine)[2] has the origin قُلتُ لَهُ /Qultu lahu/ 'I said to him' in MSA; or they are loanwords e.g. ساندويشه /sAndawyʃah/ (Gulf) 'sandwich'. In both cases, there have been significant transformations between the original words and their current expression.

The changes in words that originate from MSA are expressed on three different levels: (1) some changes are expressed by converting some letters to other letters, in other words, by changing consonants or long vowels[3]; (2) some have been changed with just the diacritisation; i.e. short vowels such as بَيت /bayt/ which is spoken as بِيت /biyt/ (Gulf) 'house'. (3) the rest have been changed by leaving out some of the *Tajweed*[4] rules, which might have an effect on either the word or the sentence level.

The significant transformation that occurred from MSA to dialects also took place

[2]In this thesis, Arabic MSA and dialect words are represented in some or all of four variants, according to context: Arabic word /HSB transliteration scheme (Habash et al., 2007)/ (dialect) 'English translation'.

[3]In Arabic there are three long vowels i.e. ا /A/, و /W/ and ي /Y/.

[4]Which means in Arabic 'to read in a correct way'.

between the words in different dialects. There are large gap between the dialects in different aspects (Versteegh, 2001). There are large differences in prefixes, suffixes, stems, loan words and the usage of words. Chapter 5 describes the results of the survey that we conducted, which showed that hundreds of words are used uniquely in one dialect but are not used at all in other dialects.

The relationship between MSA and the dialects in general is similar to that in any other language and its dialects. However, in the case of Arabic, there are two linguistic aspects: (1) the large differences between MSA and the dialects; and (2) the MSA is not native to an Arabic speaker (Habash, 2010).

2.4 Dialectic variation

Ethnologue (2013) lists 30 different Arabic dialects. Each Arabic country has its own particular main dialect, and some countries have more than one main dialect. The main dialects of each country can be divided into a group of sub-dialects; e.g. the Saudi dialect includes Najdi (Central) dialect, Hejazi (Western) dialect, Southern dialect, etc.

The Gulf region is located in the eastern part of the Arab world; the Levantine is in the north-east; North Africa is to the west; Sudan is to the south, and Egypt resides in the centre. Due to its central position, Egypt shares many of its dialect words with either those spoken in the Gulf region to the east, the Levantine region to the north-east, and North Africa to the west.

MSA	θ	ð	q	j
has converted to:				
Egyptian	s (or) t	z (or) d	A	g
Levantine	s (or) t	z	A	j
Gulf	θ	ð	g	j (or) g
North Africa	θ	ð	g	j

Table 2.4: Some changes in phones between Arabic dialects compared with MSA

2.5 Morphology of the dialects

For the morphology of the Semitic languages, one of the key distinguishing features is the root and pattern (Watson, 2007). One example from the Arabic language is the root كتب /KTB/, from which we can generate كَتَبَ /KaTaBa/ 'he wrote', كَاتِب /KATiB/ 'writer', كِتَاب /KiTAB/ 'a book', a multitude of different stem forms (Hudson, 1986).

Arabic is one of the morphologically rich languages (Olive et al., 2011; Habash, 2010; Soudi et al., 2007; Al-Sughaiyer and Al-Kharashi, 2004; Versteegh, 2001). Beside the different forms of stem, Arabic word has many different types of affixes; for example وحيلعبوها /waHayalabwhA/ 'and they will play it' is represented as five words in English, including واو العطف coordination,حاء المستقبل future particle, ها الضمير pronoun and واو الجماعة the Waw of plural. Moreover, Arabic affixation also includes همزة الاستفهام interrogative, حروف الجر prepositions, لام التوكيد emphasis, لام التعريف definite article, and others (Habash, 2010). The combination of the roots and patterns with affixes generates an excessively high number of words and leads to high Out-Of-Vocabulary (OOV) rates (Kirchhoff and Vergyri, 2005). If all the new affixes and roots that are used in the dialects were added, this would render this estimation even higher.

For speech recognition tasks, the multitude of the full word forms of the Arabic language affects the Word Error Rate (WER): for example, a 64K word lexicon typically

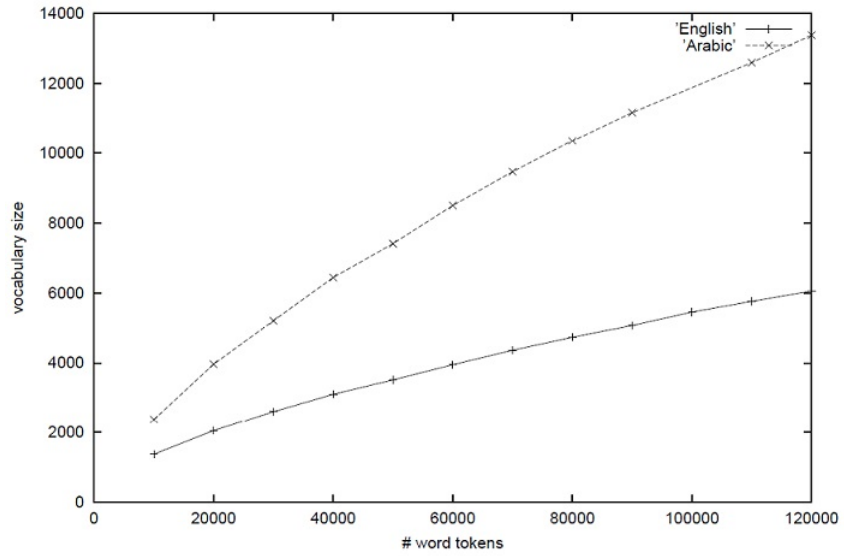

Figure 2.1: Vocabulary growth for full words in Egyptian Colloquial Arabic (ECA), from Kirchhoff et al. (2003)

leads to a 0.5% OOV rate for English, but results in approximately 5% OOV for Arabic (Afify et al., 2006). Figure 2.1 illustrates how the Egyptian Colloquial Arabic (ECA) vocabulary growth rate of full word forms exceeds that of English by a significant amount (Kirchhoff et al., 2003).

One solution for the multitude of the full word forms of the Arabic Automatic Speech Recognition (ASR) tasks is to deal with words which are missing prefixes/suffixes as correct words; for example, in the word وسيفهموها /wasayafhamuhA/ 'and they will understand it', if the واو العطف conjunction 'and' is omitted by the recogniser, this does not affect the meaning, so if the recognised word is سيفهموها /sayafhamuhA/ 'they will understand it', this would be deemed a valid interpretation.

When using a stem rather than a full word form, the affixes will be removed, this will reduce the PD size and then reduce OOV for the speech recognition task. Experimental results in Chapter 9 show the PD size is reduced by up to 34% when using stems.

2.6 Phonology of the dialects

One of the distinct features of the Semitic languages is a limited vocalic system and a rich consonantal system (Watson, 2007). For Arabic there are three basic vowels ا /A/, ي /y/, و /w/, for both short and long forms. Arabic has regular rules, in most cases, for linking orthography to phonology. Alghamdi et al. (2004) lists the rules for MSA, and how they are applied computationally. However, due to the differences between MSA and the dialects, it is unclear whether these rules can also be applied to dialects.

For the dialect words that originate from MSA, they have undergone changes in the affixation and stems of the words. Many Arabic dialects convert the phones of prefixes and suffixes, for example in the Egyptian dialect the prefix س /s/ meaning 'will' is converted to ح /H/; in North Africa, the suffix ـوا /wA/ 'they' is replaced with ـوش /wš/. Some changes also occur in stems; for example, حرف القاف /q/ is changed to ء /'/ in Levantine; in the Gulf region people use /ts/ or /tʃ/ instead of /k/ in some words. Table 2.4 shows some of these changes and compares MSA with some of the main dialects.

The developments in phonetics are reflected in the writing system. As a result of new phonetics that have been developed and absorbed into Arabic, additional symbols have been introduced for certain phonemes which previously did not exist in the MSA

alphabet; an example of one of these phonemes is ڤ /v/, which is primarily used in loanwords (Kirchhoff et al., 2003), the point here is that the usage of these representations of new phonetics is not standardised; a secondary point is that not all of the new phonemes are represented yet, such as /ts/ phonics which is used in Najdi dialect in the centre of Saudi Arabia.

In the experiments that will be shown in Chapter 8, we made use of the rules that directly affect the data that have been used in the work. Some of these rules will be listed as examples[5].

The Arabic letter ة /t/ or /h/ 'Ta-Marboutah letter' is used only at the end of a word. It can take one of two phones: when stopping it is pronounced هـ /h/, and when the speaker connects the word containing Ta-Marboutah letter with the following word, it is pronounced ت /T/. The ة Ta-Marboutah letter rule is similar to the short vowel rule for most letters at the end of the words where they take one of two phones; one with stopping i.e. Sukun 'no vowel', and one for continuation which is one of the thirteen different forms.

The /L/ phoneme in the definite article ال /AL/ can be divided into use with the 14 sun letters[6], and with the remaining moon letters[7]. In the sun letters the /L/ phoneme will be ignored (it is silent). However, with the moon letters it is pronounced. As an example, the word الشمس /Alšams/ 'the sun' is pronounced /aššams/ by deleting the /L/ phoneme; however, القمر /Alqamar/ 'the moon' is pronounced as it is written.

The last phonology rule we are going to show here is the pronunciation of Alef-Alwasl which is pronounced either ء /'/ when starting with it, or ignored when it is linked to the previous word. For example, in the word البيت /'lbayt/ 'the home' we say /'/ at the beginning as it is written because we started with it. However, when we say بالبيت

[5]See Alghamdi et al. (2004) for the full explanation of all MSA orthography to phonology rules.

[6]They are ت ث د ذ ر ز س ش ص ض ط ظ ل ن (Ryding, 2005).

[7]They are أ ب ج ح خ ع غ ف ق ك م ه و ي (*ibid.*).

/bilbayt/ ‘in the home’ we removed the phoneme /’/ from the word البيت /’lbayt/ ‘the home’ because of its position.

2.7 Some challenges for Arabic dialects

Beside the lack of diacritisation in almost all Arabic texts and the multitude of the full word forms, in terms of Natural Language Processing (NLP) and speech recognition there are three important issues that are very important to consider when discussing Arabic dialects; (1) dialects have a lack of standardisation, (2) the gap between MSA and the dialects, and (3) the shortage of available resources, especially with respect to the various dialects.

While MSA is a uniform system that creates a writing standard among different Arab countries, in Arabic dialects there is almost no standardisation of orthography (Farghaly and Shaalan, 2009), for example, among the many different forms that can represent the loanword ‘sandwich’ there are; ساندوتش /sAndwitʃ/, ساندويشة /sAndwiyʃat/, ساندويشه /sAndwiyʃh/, ساندوش /sAndwiʃ/, سندوش /sandwiʃ/, سندوتش /sandwitʃ/, سندويتش /sandwiytʃ/, and so forth. The lack of standarisation makes it very hard to generate generalisation of the rules, making some NLP tasks (such as the analysis of a dialect’s morphology) harder.

The large gap between MSA and different dialects has affected pronunciation and phone inventories as well as morphology, word order and vocabulary (Kirchhoff and Vergyri, 2005). We extracted shared words and sentences in the multi dialect speech corpus between the texts for MSA and three other dialects. Despite using parallel texts, we found

almost no sharing between MSA and other dialects with the exception of the names of cities, which were identical in different languages. We found that for MSA and the three other dialects, the percentage of shared sentences was no more than 1% when city names were excluded.

Finally, the resources are the fuel in NLP and speech tasks. In Arabic dialects there is a shortage of available tools, speech and written corpora for Arabic NLP and speech processing tasks. Most Arabic texts are written in MSA. The lack of spoken and written resources in Arabic dialects presents a serious obstacle for Arabic researchers in the area of NLP and speech processing.

2.8 Conclusions

This chapter has looked at the main features of MSA and dialects, and introduced the following points:

1. An introduction about Arabic language and general features; its dissemination, number of speakers, a quick look at Arabic font and diacritisation and the usage of MSA and dialects.

2. We have looked at the Arab world and the dialectic variation. There are about 30 main dialects and many sub-dialects across the Arab world.

3. We have introduced a brief description of Arabic phonological and morphological features and some differences between MSA and dialects.

4. Arabic language morphology raises two issues. The first relates to the root and pattern, which produces a multitude of different stem forms. The second is the affixes, which produce a multitude of full word forms. The synthesis between roots and patterns and affixes produces an extremely high number of words, which then leads to high OOV rates for speech recognition tasks.

5. In dialects morphology we have to deal with similar previous issues in respect to changes between MSA and dialects. Moreover: (1) new stems in Arabic dialects, (2) new affixes in Arabic dialects, and the most important thing (3) rules for irregular dialects.

6. Arabic generally follows regular rules based on letter to sound. However, these rules still need to be checked in order to ascertain the most efficient. In addition, there is a need to test the rules for Arabic dialects and determine if they give positive or negative results.

7. This chapter has outlined some of the challenges in Arabic dialects as they relate to NLP and speech processing. We have discussed three issues; (1) a lack of standardisation in Arabic dialects, (2) the large gap between MSA and the dialects and between the dialects themselves, and (3) the shortage of available resources for Arabic dialects.

Chapter 3

Related Work

3.1 Introduction

In this chapter we provide an overview of relevant research. We categorise the related work chapter by the area of our research. The relevant work has been divided into two main sections; (1) the morphology analyser tool and the texts and speech resources of Arabic dialects will be discussed in Sections 3.2, 3.3 and 3.4, and (2) the building and improving of Arabic dialects speech recognition methods will be discussed in Sections 3.5, 3.6 and 3.7.

3.2 Automatic multi-dialect analysis of Arabic

There are two main methods that can be used for designing an Arabic morphology analyser: statistical and linguistics based methods (Heintz, 2010). Statistical methods can be either supervised or unsupervised. For example, Darwish (2002) reports using a supervised technique to extract Arabic roots with 92.7% accuracy tested on 9606 words. Lee et al. (2003) used an unsupervised technique to build an Arabic segmenter; this reportedly achieved 97% accuracy on a corpus of 28,449 words. Although statistical methods yield high results for Modern Standard Arabic (MSA), two important issues effect the statistical basis for building multi dialect morphology analysers. The first issue is the lack of

Arabic corpora (Hammo et al., 2002), especially a diacritised pure dialects corpora, which makes means the use of statistical methods for building a multi dialect analyser may yield unreliable results. The second issue is the large gap between the dialects (Versteegh, 2001).

The linguistic method makes use of linguistic based information to define the rules of the language. Beesley (1998) made use of linguistic knowledge to produce the first morphological analyser for Arabic, following root and pattern using finite-state theory. Buckwalter (2002) built an Arabic morphological analyser using combinations of stems, prefixes, suffixes and rules to create all possible forms of an Arabic word. Although both Beesley (1998) and Buckwalter (2002) have a high coverage for MSA words, none of them can analyse dialects words. Boudlal et al. (2011) created the Al-Khalil analyser, which worked in a similar manner to the Buckwalter analyser. However, it also provides an analysis of the root of the word rather than just the stem. Al-Khalil (*ibid.*) accepts few contemporary words. Arabic dialects in many cases follow irregular rules, which makes the building of a dialects analyser using pure linguistics a very complex task linguistically.

Some researches have been done on dialects, such as that by Riesa and Yarowsky (2006) who introduced a supervised algorithm for morpheme segmentation in the Iraqi dialect; their work yielded an improvement of 50% in a machine translation task. Salloum and Habash (2011) added Egyptian and Levantine affixes to the Buckwalter analyser to build an Analyser for Dialectal Arabic Morphology (ADAM). ADAM was also designed to improve machine translation. Afify et al. (2006) conducted work in a similar area, building a dialect segmenter by adding Iraqi affixes to the Buckwalter analyser to improve Iraqi dialect speech recognition. Afify et al. (2006); Riesa and Yarowsky (2006); Salloum and Habash (2011) addressed one dialect's affixes and did not address multi dialectal affixes, their works also do not tackle irregular rules and neologistic words.

MAGEAD is an Arabic dialect morphology analyser that has provided an analysis of a root + pattern representation (Habash and Rambow, 2006). MAGEAD was developed

and tested on MSA and the Levantine dialect (*ibid.*). Therefore, when more dialects are included it becomes necessary to define additional rules. In addition, MAGEAD is limited to verbs in the dialect (*ibid.*). CALIMA is a dialect morphology analyser for the Egyptian dialect, which is linguistic based (Habash et al., 2012). In CALIMA the authors made use of the Egyptian Colloquial Arabic Lexicon (ECAL) (Kilany et al., 1997). When building dialects morphology analysers using a linguistic basis, there is a need to reference resources which are unavailable in most Arabic dialects. The final issue here is that there are more than 30 main dialects (Ethnologue, 2013) with a significant difference between each (Versteegh, 2001), which makes it difficult to build a separate morphology analyser for each dialect.

3.3 Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words

Resnik (1999) made use of a web corpus to build two parallel corpora (English-French) automatically. Ghani et al. (2001) suggested a way of creating corpora for minority languages by exploiting a web corpus; these researchers were the first to publish work describing the use of search engine queries to build a corpus from the Internet. Kilgarriff and Grefenstette (2003) listed many researches who have built text corpora by exploiting the web.

Many researchers use seed words to retrieve URLs and then build their written corpora, one example is the work of Parameswarappa et al. (2012). In a survey conducted with a group of Arabic dialects native speakers, we extracted about 1000 distinct words categorised into the four main dialects. These words can be used as seed words in a way similar to that suggested by Parameswarappa et al. (2012), for bootstrapping the web and collecting dialects corpora; to the author's knowledge no work has been done building such a corpora for Arabic dialects.

Arabic written corpora

Most of the current Arabic written corpora are from newswire resources including the Al-Hayat corpus (Goweder and De Roeck, 2001), the Gigaword corpus (Parker et al., 2011) and the LDC Arabic Newswire Corpus (Cole et al., 2001). These corpora are mostly written in MSA syntax rather than in dialects. As pointed out previously, a large difference between MSA and dialects in most aspects of language makes the use of MSA tools and corpora not useful for dialect-related tasks.

Table 3.1 shows some existing MSA corpora. It shows the source, the size of the corpus, and whether the corpus is free or not. LDC and ELRA produce the majority of different languages corpora, including Arabic written corpora.

Corpus of Contemporary Arabic (CCA) (Al-Sulaiti and Atwell, 2006) is an available corpus for the current Arabic Natural Language Processing (NLP). It consists of a contemporary text that includes expressions associated with some dialects, especially the Egyptian dialect. The transcripts of dialects speech corpora such as Appen Pty Ltd (Gulf) (2006); Appen Pty Ltd (Iraqi) (2006); Gadalla et al. (1997); Maamouri et al. (2007) have pure dialect texts for different countries, such as the Gulf, Levantine and Egyptian at both the word and sentence levels. Mostly, the texts are transcribed from telephone conversations. However, they are not big enough for most NLP tasks. Moreover, none of them are free.

Size is another important issue for text corpora. If we examine the size of the three available corpora. The Corpus of Contemporary Arabic (CCA) (contemporary) (Al-Sulaiti and Atwell, 2006) has less than 1 million tokens; Watan (MSA) (Abbas et al., 2011) has about 73 million tokens; and Khaleej (MSA) (Abbas and Berkani, 2006) has about 3 million tokens. By comparing the sizes of these previous corpora with the Arabic Gigaword corpus (MSA) (Parker et al., 2011) with in excess of 1000 million tokens, a large

Name of Corpus	Source	Size	Free?	Reference
Arabic Newswire Corpus	LDC	80 million tokens	No	(Cole et al., 2001)
CLARA	Charles University, Prague	50 million tokens	No	(Zemanek, 2001)
Al-Hayat Corpus	ELRA	18.6 million tokens	No	(Goweder and De Roeck, 2001)
Arabic Gigaword	LDC	Around 1077 million tokens	No	(Parker et al., 2011)
Watan	http://sourceforge.net /projects/arabiccorpus /files/	73 million tokens	Yes	(Abbas et al., 2011)
Khaleej	http://sourceforge.net /projects/arabiccorpus /files/	3 million tokens	Yes	(Abbas and Berkani, 2006)
A Modern Standard Arabic Corpus	Building a modern standard Arabic corpus	21 million tokens	No	(Abdelali et al., 2005)
CCA	University of Leeds, UK	Less than 1 million tokens	Yes	(Al-Sulaiti and Atwell, 2006)
An-Nahar corpus	ELRA	24 million tokens	No	(An-Nahar, 2000)

Table 3.1: Some existing MSA corpora

difference is apparent. Corpora of smaller sizes might affect the quality results of NLP tasks, failing to provide a sufficient size for training.

3.4 Multi dialect speech parallel corpora

Although many parallel text corpora have been built, very few parallel speech corpora have been produced. One of the reasons for this is that text corpus can be produced automatically or collected from different sources without big effort, whereas the building of speech corpus is a time-consuming process where one or more of the following three elements must be done manually; the recording, the text[1] and the segmentation.

Several research studies have been conducted in multilingual tasks in order to improve speech processing or NLP tasks. For example, Anumanchipalli et al. (2012) produced about two hours of parallel speech corpora for English and Portuguese and about 25 minutes parallel for English and German. Their work has been done to improve speech-to-speech translation of English-Portuguese and English-German. Pérez et al. (2012) did similar work, where they produced parallel text and speech corpus for Spanish and Basque languages. Erjavec (2004) produced a parallel multilingual speech corpus. The work started from English text which was then translated into other four languages, and the recording stage took place after that (*ibid.*).

In addition, some researchers built parallel corpora for one language but had the same text, such as Kain et al. (2011) which suggested producing two parallel speech corpora for English. One corpus is for spontaneous recording and the other one is for clear style recording, so the same text works for both corpora. The purpose of their work is to study the difference between spontaneous and clear recording acoustically.

To the author's knowledge, no one has built a parallel dialect speech corpus. This type of corpus will help researchers to study the dialects and focus on the differences between

[1]Including diacritisation for Arabic.

the dialects, especially for those languages that show a marked difference between the source of origin and the dialects such as Arabic.

Arabic speech corpora

There are many good speech resources for MSA (Elmahdy et al., 2012), however there is a shortage of speech resources which are available for Arabic dialects (*ibid.*). This shortage affects the speech processing tasks and the work on NLP for Arabic dialects.

Three main sources for the Arabic speech corpora are: microphones such as in the West Point corpus (LaRocca and Chouairi, 2002), telephone calls such as in Gulf Arabic Conversational Telephone Speech (Appen Pty Ltd (Gulf), 2006) and CALLHOME Egyptian Arabic Speech corpus (Canavan et al., 1997), or radio transmitters and receivers such as in the NEMLAR Broadcast News Speech Corpus (Maamouri et al., 2006).

LDC and ELRA distribute the majority of different language resources, including Arabic speech corpora, however most of their resources are not available for free to researchers. Most of the speech corpora from LDC and ELRA for dialects are telephone sources, where the telephone is the easiest way to gather spontaneous data in local dialects.

Table 3.2 shows some existing MSA speech corpora for different sources. Most of the TV and radio programs are introduced in MSA, which means the building of an MSA corpus using the receiver source is not difficult. However, most Arab speakers use their own dialects when calling by telephone or mobile, which makes the collecting of telephone conversations of MSA a more difficult job.

There are three speech corpora for the Gulf dialect, and as shown in Table 3.3 none of these is a microphone source. However, as can be seen in Tables 3.4 and 3.5 there is one microphone source corpus for Levantine and one for the Egyptian dialect while the remainder are all sourced from telephone calls. The last observed point in Tables 3.3, 3.4 and 3.5 is that there are no receiver speech corpora for Gulf, Egyptian and Levantine

Source	Name of Corpus	Produced by	Hours	Number of speakers
Microphone	West Point Arabic Speech	LDC	11.42	110
Microphone	Egyptian Arabic Speecon database [a]	ELRA	unknown	550
Receiver	TDT4 Multilingual Broadcast News Speech Corpus	LDC	unknown	unknown
Receiver	NEMLAR Broadcast News Speech Corpus	ELRA	40	unknown
Receiver	NetDC Arabic BNSC	ELRA	21	unknown
Receiver	Arabic Broadcast News Speech	LDC	8 recordings/each 30 or 60 minutes	unknown
Receiver	GALE Phase 2 Arabic Broadcast Conversation Speech Part 1 and 2	LDC	123	unknown
Telephone calling[b]	NIST projects (multi lingual)[c]	LDC	variety	variety
Telephone calling	Orientel United Arab Emirates MSA	ELRA	unknown	500

[a]Modern Standard Arabic as spoken in Egypt.
[b]They are Telephone calling for multilingual and microphone for English.
[c]Most of the NIST data are in English, but some may be collected in Arabic or in other languages.

Table 3.2: Some existing MSA speech corpora

dialects.

Source	Name of Corpus	Produced by	Hours	Number of speakers
Telephone calling	Gulf Arabic Conversational Telephone Speech	LDC	93	975
Telephone calling	Orientel United Arab Emirates MCA (Modern Colloquial Arabic)	ELRA	Each speaker utters specific items	880
Telephone calling	Saudi Accented Arabic Voice Bank	King Abdulaziz City for Science and Technology	96	1033

Table 3.3: Gulf speech corpora

Table 3.6 combines details of microphone sources corpora, and there are four total microphone source speech corpora for MSA and dialects, two corpora for MSA, one for Egyptian and one Levantine. There is no Gulf microphone source speech corpora to date.

It is not an easy task to gather a large amount of data for microphone sources where there is a need to record in special conditions, unlike telephone calling or receivers which seem easier to collect than microphones. The average source length of the three corpora of known length is about 25 hours per corpus, as Table 3.6 shows.

3.5 A comparison of Arabic speech recognition for multi-dialect vs. specific dialects

Although much work has been undertaken in regards to multi-dialect and multilingual tasks in many other languages, such as English (Chengalvarayan, 2001), Chinese (Liu and Fung, 2006), Spanish (Caballero et al., 2009) and German (Beringer et al., 1998), few work has been done on Arabic multi-dialect.

Source	Name of Corpus	Produced by	Hours	Number of speakers
Microphone	BBN/AUB DARPA Babylon Levantine Arabic Speech	LDC	45	164
Telephone calling	Levantine Arabic Conversational Telephone Speech	LDC	about 90	982
Telephone calling	Arabic CTS Levantine Fisher Training Data Set 3	LDC	50	unknown
Telephone calling	Levantine Arabic QT Training Data Set 4	LDC	133.6	1802
Telephone calling	Levantine Arabic QT Training Data Set 5	LDC	250	3318
Telephone calling	Fisher Levantine Arabic Conversational Telephone Speech	LDC	45	unknown
Telephone calling	OrienTel Jordan MCA (Modern Colloquial Arabic) database	ELRA	758	757

Table 3.4: Levantine speech corpora

Some research studies have been done to improve cross-dialectal recognition for the Arabic language, such as Kirchhoff and Vergyri (2005) and Biadsy et al. (2012). In Kirchhoff and Vergyri (2005), the authors reduced the Word Error Rate (WER) from 42.7% to 41.4% for MSA and Egyptian. Kirchhoff and Vergyri (2005) and Elmahdy et al. (2012) made use of MSA data to enrich other Arabic dialects. In Biadsy et al. (2012) the authors collected data for five different dialects then built cross dialects systems, finding the best WER to be 20.4% for Jordanian and Lebanese dialects; both of which are Levantine dialects.

The shortage of speech resources for Arabic dialects and the multitude of the full word forms of the MSA and dialects have affected the accuracy of the research studies on the Arabic dialects for Automatic Speech Recognition (ASR) tasks. Our multi-dialect Arabic parallel speech corpus can be considered to be the first freely available multi-

Source	Name of Corpus	Produced by	Hours	Number of speakers
Microphone	A-SpeechDB	ELRA	20	205
Telephone calling	CALLHOME Egyptian Arabic Speech + supplement	LDC	120 conversations & 20 for supplement. Each call up to 30 minutes	unknown
Telephone calling	1997 HUB5 Arabic Evaluation	LDC	20 conversations	unknown
Telephone calling	CALLFRIEND Egyptian Arabic	LDC	60 conversations, Each call between 5-30 minutes	unknown
Telephone calling	OrienTel Egypt MCA (Modern Colloquial Arabic) database	ELRA	unknown	750

Table 3.5: Egyptian speech corpora

Dialect	Name of Corpus	Hours
MSA	West Point Arabic Speech	11.42
MSA	Egyptian Arabic Speecon database	unknown
Egyptian	A-SpeechDB	20
Levantine	BBN/AUB DARPA Babylon Levantine Arabic Speech	45

Table 3.6: Microphone source speech corpora for MSA, Gulf, Egyptian and Levantine

dialect parallel speech corpus for the MSA and dialects. Such a corpus can be used to build multi-dialect speech recognition applications, and compare between dialects in phonetics, grammar and general context.

Some research studies to date have improved Arabic dialect classification, such as Biadsy (2011); Alorifi (2008); Torres-Carrasquillo et al. (2004). However, concern about the dialect classifier exists because for languages such as Arabic that have a large gap between dialects, the dialect classifier may give very low results when misidentifying dialects. Moreover, in dialect or language classifier research time is an important factor as, according to Arslan and Hansen (1996), classification accuracy becomes higher when test

utterance length increases, and this might affect real-time dialogue systems for different dialects.

3.6 An incremental methodology for improving pronunciation dictionaries for Arabic

Most research studies which have been conducted on Arabic Pronunciation Dictionaries (PDs) built new limited PD to cover the vocabulary in the text then used the Buckwalter morphology analyser (Buckwalter, 2002) to extract the different forms of each word in this PD. Maamouri et al. (2003), Maamouri et al. (2006), Soltau et al. (2007), Afify et al. (2006) mapped an one-to-one to develop PDs for their works, without employing the phonology rules, except very few unnamed rules (Biadsy et al., 2009).

Vergyri et al. (2008) used MADA (Habash et al., 2009) for selecting the best choice of each diacritised word i.e. the correct morphology. After that, they used an one-to-one mapping with a few pronunciation unspecified rules. The most cutting edge research for improving PD has been carried out by (Biadsy et al., 2009), where they added some specific phonological and morphological rules after using MADA for the best choice of word. They got a significant improvement of 4.1% in accuracy in ASR, however there are two important issues concerning Biadsy et al. (2009)'s work: (1) it is not known how much each rule has improved the total accuracy; and (2) it might be that some phonological rules, as proven in Chapter 8, have negative results, even if they are real phonological rules. In that case we need to check how much each rule positively or negatively affects the total accuracy.

Both Vergyri et al. (2008) and Biadsy et al. (2009) used their work to improve the MSA PD, yet there is still a need to apply phonology and morphology rules to check their effect on dialects and test the extent to which the gap between MSA and dialects might be affecting the PDs.

3.7 Morpheme-based language models for improving the speech recognition of Arabic dialects

In morphologically rich languages such as Arabic the large size of the lexicon is a concern of NLP and speech processing tasks. Any new word which has not been recognised will be counted as Out-Of-Vocabulary (OOV) word. This makes OOV in the Arabic language very high compared to English and some European languages that are less rich in their morphology. One of the most successful ways to solve the issue of OOV is to break the word into sub-word units, which will mean that the unique words are fewer (Heintz, 2010), and they can also have a list of affixes and then it will yield better results in ASR systems. For example, in Billa et al. (1997); Kirchhoff et al. (2006); El-Desoky et al. (2009), improvements in WER were made by using sub-word in Language Models (LMs). There will be a large difference when comparing between the full word forms versus the stem-based form for Arabic. Xiang et al. (2006) show that by using segmentation they could obtain results for a 64K lexicon that are similar to a lexicon which has a 300K vocabulary.

Billa et al. (1997) removed the definite article /il/ 'the' for the Egyptian dialect, which then reduced the vocabulary size by 7% and gave an improvement of 1% for speech recognition tasks. They used the CALLHOME Egyptian speech corpus (Canavan et al., 1997). In more in-depth work with the same speech corpus, Kirchhoff et al. (2006) extracted 22 different affixes for the Egyptian dialect, which then gave an improvement of 0.1%. Afify et al. (2006) defined prefixes and suffixes for Iraqi dialects and used them to perform a segmenter over the training corpus, and with this they obtained a WER improvement in their work of approximately 13%.

Not all research studies that have been carried out on Arabic dialects yielded an improvement. For example, Creutz et al. (2007) carried out their research on multilingual

and Egyptian Arabic using Morfessor (Creutz and Lagus, 2005) and found that there were improvements in all languages there except the Egyptian dialect. According to the authors, the reason for this is that the rate of vocabulary growth for Egyptian is not as big as MSA, however this result different from the previous results that yielded improvements for Egyptian and Iraqi dialects (Billa et al., 1997; Kirchhoff et al., 2006; Afify et al., 2006).

The research studies that have been done by extracting affixes achieved good improvements in their WER such as Afify et al. (2006); Sarikaya et al. (2007) on the Iraqi dialect which got an improvement of 13% and 7.45% respectively, and Xiang et al. (2006); Nguyen et al. (2009) obtained a WER improvement on MSA of 9.84% and 3.72% respectively. However, no research has examined which case from the following produced better improvement; (1) stem alone, (2) prefix+stem, or (3) stem+suffix.

Another issue that needs to investigate is that all of the research studies that have been done on Arabic dialects morphology that have been discussed previously in this section dealt with one dialect rather than multi dialect, and therefore the question that arises here is whether there are any issues if we try to apply the same data for multi dialect tasks. This would be beneficial and would allow us to compare the results and figure out the differences in dialect features.

3.8 Conclusions

This chapter has reviewed the relevant work, and introduced the following points:

The first area is dialects resources, i.e. tools and corpora. Therefore, we have discussed the methodologies of the research studies conducted on dialects morphology analyser tools and speech and text corpora for Arabic dialects.

The second area is improving Arabic dialects ASR by reducing the WER. We have covered three different aspects: (1) the comparison between multi dialect tasks vs. separated dialects tasks for Arabic dialects, (2) the improving of PDs for Arabic dialects, and

finally, (3) the work on sub-words to improve LMs for Arabic dialects ASR tasks.

Chapter 4

Automatic Multi-Dialect Analysis of Arabic

4.1 Introduction

In this chapter, we address the problem of the analysis of multi-dialect Arabic morphology based on the synthesis of two main methods. The first method is linguistic. A combination of dialect affixes have been added to the MSA morphology analyser to address those words that have changed as a result of their affixes.

A segmenter combined with a web engine search has been used for the second method. The word is segmented into four main forms: full word, virtual (prefix + stem), virtual (stem + suffix) and virtual stem to enable the system to check the frequency for each form. The web search consists of using the web as a corpus to extract the frequency of all word segments. This method can be used to distinguish between the full form of the word and the stem, and then the stem, prefix and suffix extracted.

The chapter is organised as follows: Section 4.2 gives the motivations of the building multi dialect morphology analyser; Section 4.3 presents the methodology that has been utilised; Section 4.4 lists the process of implementation; Section 4.5 provides an evaluation

of the work; Finally conclusions are presented in Section 4.6.

4.2 The motivations for multi-dialect morphology analysis

The large discrepancies between Modern Standard Arabic (MSA) and the dialects make it very difficult to directly use a MSA specific morphology analyser in NLP tasks for analysing dialect words. The result we obtained for the dialect corpus evaluated against MSA morphology analyser suggested just 32% accuracy. Dialect words have been amended in their roots, stems, prefixes and suffixes. There are also new loanwords that will not have been recognised by MSA morphology analyser.

The low accuracy which has been most affected by the differences between MSA and the dialects studied, motivates the possibility of building a more accurate dialect morphology analyser. Building a more precise multi-dialect morphology analyser requires the identification of both linguistic and statistical methods to determine which are applicable and most useful for succeeding in the work.

The majority of Arabic dialects occur in spoken rather than written forms. However, there is a need for written dialectal forms that can be applied to address spoken dialects in Automatic Speech Recognition (ASR) applications and Text To Speech (TTS) applications that are useful for handling dialects. The language model should be based on the same dialect which we would use in speech. Kirchhoff et al. (2003) tried to use MSA text in speech recognition for Egyptian dialect. However, their experiment did not yield any improvement. Thus, the need for written forms to describe Arabic dialects is crucial for NLP dialect tasks, especially those tasks that deal with speech.

Once we have obtained these written forms of dialects, the multi-dialect morphology analyser would then be used for the important task of extracting the stem or the root of the word so as to highlight the affixes and to show the features of the selected word/s.

4.3 Methodology

The methodology used in the multi-dialect morphology analyser is divided into three parts. The first part is related to adding dialect affixation to the MSA morphology analyser. According to Haak (1996), the stem patterns of Arabic dialects are identical to those of MSA in most cases. Therefore the first action is to ascertain how the MSA morphology analyser has been improved following the addition of the dialectical affixes.

The second part is to implement a special segmenter for segmenting words that have not been recognised in the adapted MSA morphology analyser. Four patterns for each word are be extracted from this segmenter: the full word, stem, virtual (prefix + stem) and virtual (stem + suffix).

The third part is to use a search engine to get the frequency of each pattern. The Arabic content on the Internet has in recent years increased dramatically (Tawileh and Alghamedi, 2011), and now exceeds 2 billion pages (Alarifi et al., 2012). Thus the content is deemed comprehensive enough to yield reliable results. This component will depend on the frequency of segmented words observed from the previous part, as retrieved from the search engine.

The hypothesis for the using search engine to obtain word fragments frequency is that most Arab dialects use similar stems in both loanwords and MSA altered words; however, the difference is mainly based on affixes or diacritisation. Even for those dialects that use the same affixes they use stems more often than full words. So, from this hypothesis, we can say the frequency of the stems allows a higher frequency of retrieval based on them, than when using full word forms. An example of this hypothesis is this stem يلعب /yalʕab/ 'he plays', should be greater frequency than the words حيلعبوا /HayalʕabwA/

‘they will play’ (Egyptian), هيلعبوا /hayilʕabwA/ ‘they will play’ (Libya) or ما يلعبوش /ma yalʕabwʃ/ ‘they do not play’ (Levantine) in different dialects (the frequencies given respectively are: 3,880,000; 4,100; 5,500; and 2,300). So to evaluate the hypothesis, all four forms have been tested to compare which provides the greatest frequency.

Some of the dialect verbs and other loanwords have an actual prefix but do not have a suffix. Thus they just have similar letters to one of the suffixes. For example the word الميكروفون /Almykrwfwn/ ‘the microphone’ has a prefix but does not have a suffix. It has ـون /wn/ which is similar to the pronoun ‘they’ in the second part of the word; but these letters are actually from the original word and not an actual suffix.

For the previous example, four shapes were produced from the second step: the full word: الميكروفون /Almykrwfwn/ ‘the microphone’; virtual stem: ميكروف /mykrwf/ ‘wrong word’; virtual (prefix + stem):الميكروف /Almykrwf/ ‘wrong word’ and virtual (stem + suffix): ميكروفون /mykrwfwn/ ‘microphone’ with the following frequencies respectively that have been retrieved from the third step: 223,000; 1,400; 3,800 and 367,000. For this example the form virtual (stem + suffix) is the correct analysis and it is reported as having the greatest frequency as shown.

Sometimes the correct analysis will be the virtual (prefix + stem); and sometimes the full word would be the correct analysis of the word such as the word بزوره /buzurah/ ‘children’ (Hejazi).

Four conditions in the sequence have been set for verification as outlined below:

1. The word is not in the MSA words list.

2. The change/s in stem and not just in the affixes.

3. The search in a set of only Arabic language.

4. The greatest word frequency from four forms is chosen, and its frequency has to be greater than or equal to 10000 as a threshold.

For the first condition the word is not in the MSA words list, if it were it would be an MSA word. MSA words should be analysed before entering any of the subsequent steps.

The second condition is the change/s in stem and not just in the affixes, those which are a dialectical words with a change to either prefix or suffix, so e.g. words حيلعبوا /Hayalabw/ 'they will play' (Levantine), حيناموا /HaynAmw/ 'they will sleep' (Egyptian) and بيقولوا /byqwlw/ 'they will say' (Gulf). All of these words are analysed based on the first part where all of these dialect words have been changed just in their affixes.

The third condition limits the search to Arabic language, to avoid any results from other languages that use all or some of the Arabic script; e.g. Sindhi, Pashto, Persian, Malayalam, Urdu and Ottoman also utilise Arabic script (Nakanishi, 1980).

The greatest word frequency from four forms is chosen in the fourth condition, and its frequency has to be greater than or equal to the threshold of 10000. Therefore, choosing the greatest word frequency is based on the main hypothesis that the use of the stem by Arab speakers is greater than the use of a word that has a prefix or suffix. The frequency has to be more than ten thousand times; Kilgarriff and Grefenstette (2003) states that lower frequencies from the web might be unreliable. So based on ten thousand results we

can be certain this is right word, not based on typing mistakes, and this might be dialect word, loanword, or even MSA word is used in somewhere in the Arab countries.

Algorithm 1 represents the method steps for both the linguistic and statistical stages in processing.

4.4 Implementation

Algorithm 1 works on three layers. The first layer is to implement the analyser for dialect affixes that have been inherited from the MSA words. The next layer is to implement a segmenter for the words that have not been recognised. This segmenter produces four forms; the full word, virtual stem, virtual (prefix + stem) and virtual (stem + suffix). The last layer is to implement a tool to distinguish segmented words by using external knowledge.

Both the linguistic and statistical bases are needed for this algorithm; whereby the linguistic base is needed to analyse dialectic words that have been adapted from the morphology analyser by using the MSA roots database and some affixes shared between MSA and dialects; the statistical base plays an important role in extracting the other stems from the web corpus depending on segmented words frequency.

A list of affixes for different Arabic dialects has been collected. Some of these affixes consist of more than one prefix and/or suffix i.e. a combination of prefixes and/or suffixes. Here are some examples of prefixes وهـ /waha/ 'and will' (Levantine and North Africa), حـ /Ha/ 'will' (Egyptian), وبـ /wib/ 'and will' (Gulf), وحـ /waHa/ 'and will' (Egyptian).

```
Data: words list
Result: analysed word [(prefix) (stem) (suffix)], OR: Does not need to analyse,
        Otherwise: unknown word
initialization;
Constant frequency AS INTEGER-TYPE = 10000;
length AS INTEGER-TYPE = 3;
while not at end of the list do
    read next word;
    if there is a solution IN adopted dialect morphology analyser then
        OUTPUT stem + (dialect) prefix + (dialect) suffix;
    else if stem.length ≥ length then
        Segment the word into four forms;
        switch the frequencies for each form of the word in web do
            case virtual stem.frequency is the greatest AND ≥ frequency
                OUTPUT stem + prefix + suffix;
            case virtual (prefix+word.frequency) is the greatest AND ≥ frequency
                OUTPUT (prefix + stem) + suffix;
            case virtual (word + suffix.frequency) is the greatest AND ≥ frequency
                OUTPUT (stem + suffix) + prefix;
            case word.frequency is the greatest AND ≥ frequency
                OUTPUT Arabic word + (analysis not required);
            otherwise
                OUTPUT is not an Arabic word;
            end
        endsw
    else
        search word frequency in web;
        if word.frequency ≥ frequency then
            OUTPUT Arabic word + (does not need to analyse);
        else
            OUTPUT is not an Arabic word;
        end
    end
end
```

Algorithm 1: Multi-dialect morphology analyser

Also other examples of suffixes are ـوش /wʃ/ ‘they’ (Levantine and North Africa), ـوا /wA/ ‘they’ (Levantine). There are also some MSA prefixes and suffixes that have been added to test if the analyser can extract MSA words that have not been recognised by the MSA morphology analyser or not.

To implement the first layer, the open source MSA morphology analyser Al-Khalil (Boudlal et al., 2011) is used and applied. This provides a good initial in-depth analysis. Therefore the dialectical affixes have been added to the analyser which has provided an encouraging result. The new affixes that have been added include the definition to introduce a full analysis of the required word. This definition shows the type of prefix and suffix; e.g. the prefix حـ /Ha/ points to the near future in Egyptian dialect. After adopting the database of Al-Khalil morphology analyser (*ibid.*), we acquired a result that was produced from the first layer; Table 4.1 shows an example of the output after the first layer.

نتائج التحليل							
Analysis Results							
Input	Voweled Word	Prefix	Stem	Type	Pattern	Root	Suffix
ابخص	ابْخَصْ	#	ابخص	فعل أمر	افْعَلْ	بخص	#
ابضاي	no result						
اتربع	no result						
احتريك	no result						
هيجلس	هيَجْلِس	ه: هاء المستقبل القريب + ي: حرف المضارعة	يجلس	فعل مضارع مبني للمعلوم	يَفْعِل	جلس	#

Table 4.1: Example of the output after the first layer

The second layer involves building a segmenter that produces all of the possible forms of the word. These forms as stated as previous, including; full word, virtual stem, virtual

(prefix + stem) and virtual (stem + suffix). Table 4.2 shows an example of six produced words by this segmenter.

Full word	virtual stem	virtual (prefix+stem)	virtual (stem+suffix)	virtual prefix	virtual suffix
حيلعبوا	يلعب	حيلعب	يلعبوا	ح	وا
حيمتنعوا	يمتنع	حيمتنع	يمتنعوا	ح	وا
تلفزيون	تلفزي	تلفزي	تلفزيون	#	ون
الحلج	حلج	الحلج	حلج	ال	#
برتقان	رتق	برتق	رتقان	ب	ان
الصابون	صاب	الصاب	صابون	ال	ون

Table 4.2: Example of segmented words

After the completion of segmenter, the last layer involves taking the remaining words to measure their frequency. The Bing search API (Microsoft, 2011) has been used to retrieve the number of results for each segmented form of the word. An example of the process associated with the final procedure has also been given; the Figure 4.1 shows how the words be analysed by using the algorithm.

After the final step of implementation, the third layer, the result based on the rest of the words is shown. Table 4.3 shows an example of the output after the last layer.

4.4.1 The building of Arabic multi dialect morphology analyser webpage

We have built a webpage for the Arabic multi-dialect morphology analyser [1] to provide a service for the researchers. This webpage can be used by any researcher to analyse Arabic dialects or MSA words. It follows the algorithm suggested in this chapter. The source code for this webpage is also available to researchers[2].

[1] www.arabicmorphologyanalyser.com.

[2] C# and ASP.Net have been used to build this webpage.

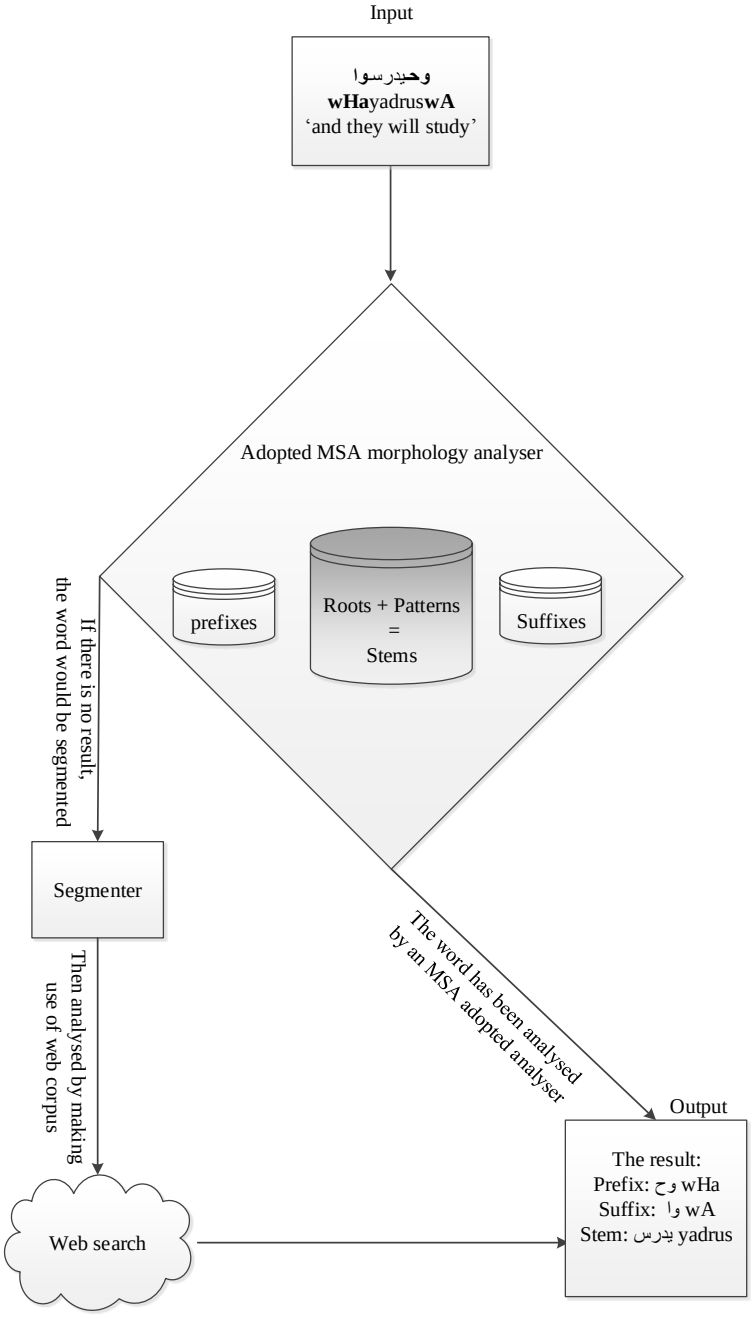

Figure 4.1: How words be analysed by using the algorithm

نتائج التحليل					
Analysis Results					
Input	Input-frequency	Prefix	Suffix	Stem	Stem-frequency
حيحتاجوا	1350	ح: حاء المستقبل القريب	وا: واو الجماعة	يحتاج	11400000
شدعوه	301000	MSA/dialect word does not need to analyse			
هيناموا	8190	ه: هاء المستقبل القريب	وا: واو الجماعة	ينام	4260000
بيتأرجح	2840	ب: باء المستقبل القريب	#	يتأرجح	210000
بيركع	13000	ب: باء المستقبل القريب	#	يركع	577000
أبضاي	38500	MSA/dialect word does not need to analyse			
أتاي	114000	MSA/dialect word does not need to analyse			
دنفسة	4230	no result			
حيسافروا	10900	ح: حاء المستقبل القريب	وا: واو الجماعة	يسافر	1960000
وازهله	4870	no result			

Table 4.3: Example of analysed words after the last layer

4.5 Evaluation

For the evaluation, 2229 dialectical words (names, verbs, adjectives and adverbs) were collected from six different Arab dialects: Egyptian, Levantine, Gulf, Sudani, North Africa and Iraqi. The advantage of the variety is to guarantee that this work covers most of the Arabic main dialects, and so gives reliable results.

The first test was to check these words on the MSA morphology analyser. As can be seen from Table 4.4, 68% of the total words were unknown; i.e. the MSA morphology analyser could analyse just 32% of the dialectical words.

An improvement occurred after the next step, which was the adaption of the MSA morphology analyser adding different dialectic affixation. The percentage of unknown words in the MSA analyser after modification was reduced from 68% to 31% of the total list of words, as shown in Table 4.5. This showed that there are a significant number of

The number of words	2229
Unknown words	1508
Unknown words (%)	68%
Recognized words	721
The total accuracy	32%

Table 4.4: Results before starting the experiments

dialectical words that have been changed purely by their affixation, without any change to their stems.

The number of words	1508
Recognized words	824
Recognized words (%)	55%
Unknown words	684
Unknown words (%)	45%
The total accuracy	69%

Table 4.5: Results after MSA analyser has adopted

The next step was to take the remainder of the words to search on the web after segmentation and applying the remaining conditions for the search. The recall, as is shown in Table 4.6, is 90% and the result of the search precision is 80%. The combined F-measure is 85% of the remaining words checked by the search engine. The error percentage based on total words was reduced from 31% in the previous step to 6%, which results in an accuracy for this analyser is 94%.

Table 4.6 shows that 20% of the words checked by web-as-corpus were wrongly analysed. 11% of these words did not meet the condition of a threshold frequency greater than 10,000 instances. For those words that appeared at a frequency less than the required,

The number of words	684
Recall (frequency >=10000)	90%
Precision (%)	80%
F-measure	85%
The total accuracy	94%

Table 4.6: The final results

the advantage of this analyser was that it updates; therefore, as soon as word becomes a usable it will be recognised.

To make sure the update benefit is effective we retested the experiment on the same list two months later and found that unknown words have reduced from 76 to 64 words.

The remaining 9% were incorrectly analysed. Those wrong analysed words either the use of the full word is more popular than the stem, and so got higher frequency such as الحلج /AlHalj/ 'the throat' (Kuwaiti), or they were given other meaning after segmentation and then converted to other words such as بيبي /byby/ 'baby' has converted to يبي /yaby/ 'he wants' (Gulf). To resolve these errors there is a need to add more linguistic rules. For example a rule has been added to deal with future tense verbs: حـ /Ha/ 'will' and هـ /ha/ 'will' to be like a dealing with a سـ /sa/ 'will'; i.e. they accept just four letters to follow them ا /A/, ن /na/, ي /ya/ and ت /ta/.

The method of using web search also works with some MSA words that were not analysed correctly in stage 1. For example, الخبراء /AlxubarA'/ 'the experts' آخرون /Axrwn/ 'others' and مسؤولون /masŵwlwn/ 'the accountants'. So even those morphology analysers that have large databases can be improved using this method.

The web search can also distinguish between those MSA or dialect Arabic words when they do not need to analyse and other words. There are three words in Table 4.3; أبضاي

/Abadˁ Ay/ ‘strong man’ (Levantine), أتاي /AtAy/ ‘tea’ (North Africa) and شدعوه /ʃdaʔwah/ ‘why’ (Gulf). These three words do not have neither prefix nor suffix. However, the analyser could extract and distinguish them.

One aspect of the use of web as corpus is that for Arabic the search engines did not support diacritics in the proper way; e.g. when using Bing search API (Microsoft, 2011) as a search engine it gives similar results for diacritised and non-diacritised words. These results are not accurate since most of the Arabic texts are non-diacritised. So were search engines to support diacritics in the future this would be of considerable benefit for such research.

The last advantage of this analyser is that in most cases it can be used to differentiate between those words that have an actual suffix, such as in the previous examples; آخرون /Axrwn/ ‘others’ and مسؤولون /masŵwlwn/ ‘the accountants’ and those that have just similar letters of suffix. For example, جيلاتين /jylAtyn/ ‘gelatine’ (loanword) when analysed by this analyser the segmenter will produce four forms; full word: جيلاتين /jylAtyn/ ‘gelatine’ (loanword), virtual (prefix + stem): جيلات /jylAt/ ‘wrong word’, virtual (stem + suffix): جيلاتين /jylAtyn/ ‘gelatine’ (loanword), virtual stem: جيلات /jylAt/ ‘wrong word’, with virtual suffix: ين /yn/, whereas here it will choose the right form, which is the full form. There are a significant number of words in the same situation such as تلفزيون /tilivizywn/ ‘TV’ (loanword) and الكرتون /Alkartwn/ ‘the cartoon’ (loanword). This shows one of the benefits in that it can distinguish neologistic dialectical

words.

4.6 Conclusions

This chapter describes a multi-dialect morphology analyser, which uses both a linguistic basis and a statistical basis to analyse words from Arabic dialects; it introduced the following points:

1. The suggested Arabic dialects morphology analyser has three parts; the linguistic base, the segmenter and the web search feature.

2. The linguistic base was prepared making use of the MSA morphology analyser and adapting it to accept affixes used in dialects. The overall accuracy rate improved from 32% to 69% after adoption.

3. The segmenter was created to give the four possible forms of each word. Full word form, virtual (prefix + stem), virtual (stem + suffix) and virtual stem.

4. Using the web as a corpus made it possible to search the frequencies retrieved for each segment of the word to distinguish between the full form of the word and the stem with affixes. This generated a result with 94% accuracy for 2229 different dialect words.

5. Some advantages to using the web as a corpus are; firstly, it allows identification and inclusion of MSA words that have not been recognised by the original MSA morphology analyser. Secondly, it facilitates distinction between the actual prefix and/or suffix and those letters that are similar to them but not actual prefixes/suffixes; and finally the method is up to date, allowing detection of any new popular dialect word.

6. One of the shortcomings of using the web as a corpus is that it does not yet support diacritics.

Chapter 5

Automatic Building of Arabic Multi Dialect Text Corpora by Bootstrapping Dialect Words

5.1 Introduction

As described before, in Arabic there is a shortage of available corpora for Natural Language Processing (NLP) tasks. This shortage becomes a huge when discussing dialect resources, especially dialects in written corpora, where most of the Arabic texts are written in Modern Standard Arabic (MSA). This point raises a big obstacle for Arabic researchers in the area of NLP.

The problem of shortage of resources in Arabic has several aspects. All researchers in Arabic NLP have to construct some or most of their own resources (Goweder and De Roeck, 2001). Each researcher spends a long time compiling sufficient resources. The standardisation issue in NLP tasks is another aspect to consider when discussing shortage of data.

For the purposes of this work we have presented the methodology for building auto-

matic Arabic dialects corpora by exploiting the web as a corpus. A key contribution of this work is dialect identification by creating words lists for each of the four main dialects. This involved conducting a survey with a group of Arabic speakers from different countries to assist in categorising the words. Once the lists were ready we bootstrapped the word by using a web search engine to download pages that were likely to have the same dialect as the word that we are searching for.

This chapter is organised as follows: Section 5.2 highlights the need for dialects written text corpora; Section 5.3 presents the methodology that has been used, including a description of the survey that was conducted; Section 5.4 describes the process of implementation; Section 5.5 presents an evaluation of the work; Conclusions are presented in Section 5.6.

5.2 The motivations for building a multi dialect written text corpora

The majority of the Arabic dialects occur as spoken rather than written forms. There is a need for dialectal written forms to deal with spoken dialects in many NLP applications and speech processing tasks such as in Automatic Speech Recognition (ASR) and Text To Speech (TTS) applications sensitive to the dialects. The language model should have same dialect which we would use in speech. Kirchhoff et al. (2003) have tried to use MSA text in speech recognition for the Egyptian dialect. However, their experiment did not yield any improvements.

The differences in words, syntax and phonetics between MSA and dialects and dialects themselves, as discussed in Chapter 2, make it very important to use dialect based corpora instead of using MSA speech or text based corpora.

One of the important issues is that to date no available multi-dialect text corpora for public use exist. Therefore, if one were to be developed for public use it would offer

considerable opportunity for researchers to work on Arabic NLP tasks. So, if (1) large-sized, (2) open-domain and (3) available multi-dialect corpora were built to address the shortage in Arabic resources, this would be very efficient, as it would permit researchers to compile more and more information in the area of Arabic NLP.

The Arabic content on the Internet has in recent years increased dramatically and now exceeds 2 billion pages (Alarifi et al., 2012). This content is comprehensive enough to exploit in order to build dialects corpora.

The huge number and variety of Arabic texts on the web mean a good selection of different text corpora is available. Although most of the Arabic texts on the web are written in MSA, there are still enough dialects on the web available in the blogs, comments, forums, Facebook, Twitter, Instagram, etc. if we can make use of these resources we can compile useful Arabic dialects text corpora.

5.3 Methodology

In this work, we gathered unique words from dialects and used them as seed words to retrieve URLs. We downloaded the required pages to build dialects corpora.

The methodology involves five main steps in the process of building multi dialect corpora: Collect multi dialect words and phrases, conduct a survey, estimate the counts, downloading and perform a cleaning and a normalisation.

Step 1: Collect multi dialect words and phrases

The first step is to collect dialect words and phrases that have been used in different Arab countries. All the words in each list should be placed in one of the four main dialects; Gulf, Egyptian, North African and Levantine.

The step of collecting words was done by exploring the web to extract dialects words. Some of these words were derived to get greater numbers of dialect words for example the

word احتريك /Aħtrek/ 'I am waiting for you' can be derived to احتريه /Aħtreh/ 'I am waiting for him' (Gulf) and to other forms. At the end of this step about 1500 dialect words had been collected from the web. These words were categorised into four main dialects.

Step 2: Survey dialect speakers

We ensured that the words in each dialect list were actually used in this dialect. However, we cannot be certain that these words do not appear in other dialects. Therefore, there was a need to conduct a survey with a group of Arab speakers from different countries to guarantee that they do not use any words from lists other than their own. The importance of the survey step was to distinguish between the words that are used in each main dialect.

To perform dialect identification, six people, 3 males and 3 females, were recruited participate in the survey. The job of each participant was to make sure that the entire words list, except for his/her word list were not used in his/her main dialect i.e. Gulf, Egyptian, North African and Levantine. Some of the words/phrases in the list are used in other dialects with minor changes, this is acceptable because we will search for an exact word/phrase. Most of the resulting lists contain separated words; some of them include Arabic expressions (such as dialects proverbs). Table 5.1 shows some examples of categorised words and phrases for the four main dialects.

By the end of the survey, the four lists of single-dialect words are ready for bootstrapping in the web corpus. Table 5.2 shows the number of words for each dialect and the total of words after the survey; which is about 1000 words and phrases.

As can be seen from Table 5.2 Egyptian was identified as having the lowest number of tokens. The reason for that is the location of Egypt in the centre of the Arab world. It shares words with many other Arab countries, unlike the Gulf which is located on the east of Arab world, and Levantine and North African which are respectively in the North

Gulf	Egyptian	North African	Levantine
باتسر /batsir/	اتلم /Atlam/	اتاي /Atay/	تؤبرني /tobirny/
ابتل /ibtil/	مسحول /masħwl/	البغرير /elbaɣrer/	ابوشريك /Aboshrek/
ابخص /Abxasˁ/	اشتغالة /eʃteɣalih/	طوبيس /tˁwbes/	ابيش /Abeʃ/
اتربع /AtrabbaʕA/	الباشكاتب /elbaʃkatib/	الدسة /eldassah/	اجبد /ejbid/
اتمرمط /Atmarmatˁ/	البتاع /elbetaʕA/	الرفيس /elrafes/	اختصر وقسم على عشرة /extasˁir wqssim ʕla ʕʃrah/
اثره /Aθroh/	الجون /Aljwn/	السمش /Alshamaʃ/	ادن من طين وادن من عجين /odn min tˁen wodn min ʕjen/
احتريك /Aħtrek/	الدفش يافت /Aldafʃ yafit/	الشاف / Alʃaf/	اطعج يابكرج /etˁʕj yabkrj/
ازبن /ezbin/	بايز /bayiz/	الشنين /Alʃnen/	الباب يلي بيجي منو الريح سذو واستريح /elbab ylly byje mnw elreħ sddw westireħ/
بشويش /biʃweʃ/	الكسكسة /Alksksah/	الصمطا /Alsˁmtˁa/	الحكي الك ياكنة سمعي ياجارة /Alħke elk yaknnah smʕy yajarah/

Table 5.1: Examples of categorised words and phrases

Dialect	Total of words
Gulf	430
North African	200
Levantine	274
Egyptian	139
Total	1043

Table 5.2: Total number of words

Dialect	Word count	Number of words per link (Average)	We need
Gulf	430	711 words	50 pages
North African	200	528 words	100 pages
Levantine	274	794 words	50 pages
Egyptian	139	979 words	100 pages
Average	—	753 words	—

Table 5.3: The estimation of how many pages we need per dialect

East and West of the Arab world.

Step 3: Estimating the counts

Before the downloading stage took place we needed to calculate the average for how many tokens will be produced per link. Table 5.3 shows the results of the estimation. In this step we intend to make sure that each dialect corpus will be more than 10 million tokens. To simplify the process we will choose either 50 or 100 pages for each dialect according to the estimation.

Based on the results in Table 5.3, it is possible to determine how many pages we need for each dialect to get comparable results. We set the page count to 50 pages for Gulf and Levantine, and 100 pages for North African and Egyptian; More than 100 links per dialect might represent the retrieval of duplicate results. This estimation should then give an average of greater than 10 million tokens for each dialect.

Step 4: Collecting the links and downloading the pages

After the estimation for the required pages of each dialect, we should move on to the next step, collecting links according to the seed words. Using Bing API (Microsoft, 2011) we searched for the seed words for every dialect, word by word. We retrieved the first 50-100 links for each search, then downloaded those pages, which likely have the same seed word dialect. For all four corpora, we downloaded more than 55000 valid web pages from different web resources, saved as HTML.

Although most of the Arabic web pages use cp1256 encoding, other coding is also used, including UTF-8 or ISO 8859-6. In this work, we use cp1256 encoding when saving HTML files.

Step 5: Cleaning and Normalisation

The cleaning of the resulted corpora is an important step as the biggest drawback in a web corpus is noise. The collecting of web pages from different resources such as forums, blogs, comments, etc. makes it very hard to make generalisations of the rules. So, each cleaning step includes removing unwanted symbols, tags, underscores, and more than one space, etc.

We also need to perform a normalisation for the resulted texts. The normalisation includes removing repeated words and phrases automatically in web pages especially the more popular words used in forums; such as الصفحة الرئيسية 'home page', عضو 'member', التسجيل 'register' etc. The normalisation also includes removing repeated sentences or paragraphs that appear in sequence automatically. However, if the sentences are repeated

Dialect	Size in million token	% of repeated sentences	Unique types
Gulf	14.5	41%	920K
North African	10.1	41%	720K
Levantine	10.4	33%	770K
Egyptian	13	42%	770K

Table 5.4: Results

Size in million token	50
Average of % of sentences repeated	39%
Unique types for all dialects	2 million
Number of words per link (Average for all)	753

Table 5.5: Total results

but not in sequence, we cannot remove the duplicate sentence as it might be come from a different context.

5.4 Results

Table 5.4 shows the results of the experiment after the cleaning and normalisation. Gulf, North African, Levantine and Egyptian corpora have 14.5 million, 10.1 million, 10.4 million and 13 million tokens respectively. Table 5.4 also shows the unique types in each dialect which are 920K, 720K, 770K and 770K respectively. Table 5.5 shows the total results of all four corpora sizes; which is around 50 million tokens. For these words there are more than 2 million unique types for all corpora. As the corpora have been built from the web, they have a high percentage of repeated sentences with an average of 39% as Table 5.5 shows.

There are more than 5 million sentences in all four corpora, as Table 5.6 shows. The average of sentences count is 1.3 million sentences per corpus and the average sentence length is about 10 words per sentence.

Dialect	Exactly size	Number of sentences	Average sentence length
Gulf	15,291,500	1,729,545	8.84
North African	10,570,195	1,044,495	10.12
Levantine	10,885,829	1,143,354	9.52
Egyptian	13,611,341	1,265,921	10.75
Total	50,358,865	5,183,315	9.8

Table 5.6: Sentences counts and average of length

5.5 Evaluation

To evaluate this work we compared it to the MSA Gigaword corpus (Parker et al., 2011) and a Corpus of Contemporary Arabic (CCA) (Al-Sulaiti and Atwell, 2006). Arabic Gigaword corpus (Parker et al., 2011) is the largest MSA corpus, covers most of the Arabic topics and is produced by LDC. CCA (Al-Sulaiti and Atwell, 2006) is an available and free contemporary corpus and has less than 1 million words. We covered different elements of this comparing (Out-Of-Vocabulary (OOV), size, syntax) before going into more in-depth analysis for the resulting corpora.

5.5.1 Comparing results

Out-Of-Vocabulary (OOV) comparison

We have extracted 50k unique types from the resulting corpora, Gigaword corpus, and CCA corpus. After analysing these unique types using the MSA morphology analyser Alkhalil (Boudlal et al., 2011), the percentage of unknown words was determined, as shown in Table 5.7. For Arabic Gigaword (Parker et al., 2011) and CCA (Al-Sulaiti and Atwell, 2006) the percentage of unknown words was less than 20%, whereas the percentage of unknown words for our dialect corpora was about twice this at 39%. As Alkhalil (Boudlal et al., 2011) is an MSA morphology analyser, it will not analyse many words from dialects, resulting in a high OOV for dialects corpora, as demonstrated by the results in the Table 5.7.

Corpus	The % of unknown words	Dialect(s)
Arabic Gigaword	18%	MSA
CCA	15%	Contemporary
Our four corpora	39%	Multi dialect

Table 5.7: Comparing unknown words

Corpus	Total tokens	Unique types
Gigaword	>1000 million	Unknown
CCA	600K	125K
Our four corpora	48 million	2 million
North African	10.1 million	720K
Levantine	10.4 million	770K
Egyptian	13 million	770K
Gulf	14.5 million	920K

Table 5.8: Size comparisons

Size comparison

Table 5.8 shows the result of comparing the resulting corpora with respect to their sizes with CCA (Al-Sulaiti and Atwell, 2006) and Gigaword (Parker et al., 2011). Arabic Gigaword corpus is designed for MSA. However, CCA has some dialects (contemporary) words while mostly retaining an MSA context. The Gigaword corpus is the largest MSA corpus. For the dialects, as can be seen from Table 5.8, a big difference in size is apparent between our four corpora, with an average of 12 million tokens, and CCA (Al-Sulaiti and Atwell, 2006) (which is less than 1 million tokens).

The size of CCA (Al-Sulaiti and Atwell, 2006) is also reflected on the count of unique types which is 125K, where the average unique types of our dialects corpora is 800K, with a total of 2 million for all. This big difference in unique types will affect the researches on NLP tasks, where the work on CCA (*ibid.*) gives a chance to deal with just 15% of the word variety of the average of the resulting corpora, and it has just 6% of the variety that in our four corpora.

Token frequency	Percentage of frequency
1	49%
2-9	39%
10-100	10%
>100	2%

Table 5.9: Frequency of frequencies of token types

Syntax comparison

It is very difficult to measure syntax differences between MSA and dialects corpora where there is no available multi-dialect parsing, dialects Part-Of-Speech (POS) tagger, or available parallel text corpora between MSA and dialects (Chiang et al., 2006) to use for parsing Arabic dialects.

When looking at the resulting dialects corpus, it is clear that its syntax differs from the MSA syntax in general context, sentence representation, word order, affixes and even the stems of words. In different aspects of writing, this issue is also noticeable among the resulting dialects corpora themselves.

5.5.2 Analysis

Table 5.9 shows the frequencies of token types in the four groups; those tokens that are listed once 'hapax legomena', those tokens that are listed from 2 to 9 times, those tokens that listed 10-100 times, and those tokens that appear more than 100 times. It is very common to have the majority be a percentage of tokens that have a frequency of less than ten times (Manning and Schütze, 1999), which is 88% in the resulted corpora. Also it is common that a very few tokens have a very high frequency (*ibid.*).

Table 5.10 shows the most-often occurring 10 unigrams with their frequencies for all four corpora. The unigrams that appear in Table 5.10 are originally MSA words. However, when these words are used in dialects there are some diacritisation changes; e.g. كل /kul/

Frequency	Token	Meaning(s) according to diacritisation
953440	من	From, that, who?
771032	في	In, mouth (very rarely used)
527566	و	And, conjunction
519358	على	On
400466	ما	That, not
349776	الله	Allah
264736	لا	No, not
228592	يا	O
182279	عن	From, instead of
163140	كل	All, eat, tired (very rarely used)

Table 5.10: 10 greatest unigrams tokens frequencies (including function words)

(MSA) is كل /kil/ (Gulf) 'eat'. The total of the 10 greatest unigrams is about 9% of the corpus size. As can be seen from Table 5.10, most of the words have more than one meaning and the different meanings are dependent on context.

Most of the tokens in Table 5.10 are function words. If we try to look at the most frequent tokens by excepting function words we will get the results that seen in Table 5.11.

To confirm that our corpus is comparable to other text corpora we have applied Zipf's law (Zipf, 2012) to all four corpora. Zipf's law suggests that there is a constant k such that:

$$f.r = k$$

where: f is the frequency and r is the rank

The results of Zipf's law for all corpora were almost exactly -1, which is predicted by Zipf's law. Figures 5.1, 5.2, 5.3 and 5.4 show the results.

Table 5.12 shows the bigrams, trigrams and five-gram counts that resulted. In our

Frequency	Token	Some meaning(s) according to diacritisation
349776	الله	Allah
100041	محمد	Mohammed
90579	والله	by Allah
71986	يوم	day
70455	علي	Ali
54443	واحد	one
48490	الموضوع	the subject
46277	اليوم	today
46005	السلام	the peace
45887	مثل	like, ideals (rarely used)

Table 5.11: 10 greatest unigrams tokens frequencies (non function words)

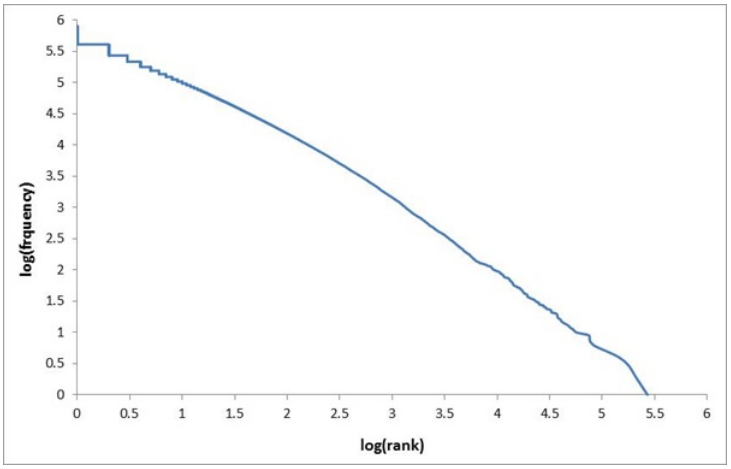

Figure 5.1: Zipf's law of Egyptian, slope = - 0.9761. log of rank on the X-axis versus frequency on the Y-axis

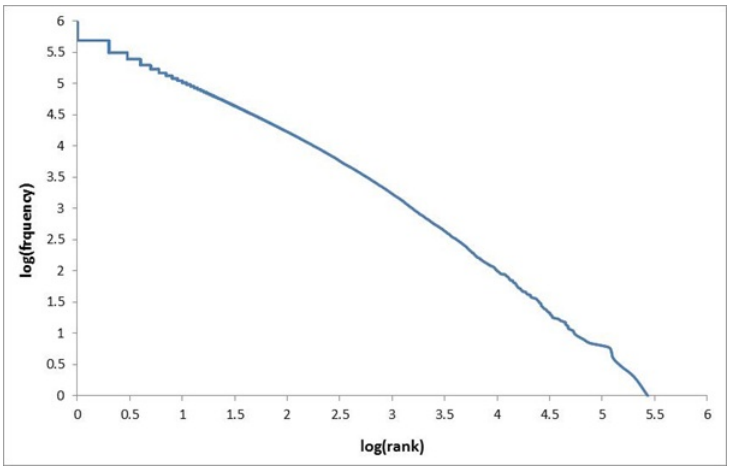

Figure 5.2: Zipf's law of Gulf, slope = -0.9759. log of rank on the X-axis versus frequency on the Y-axis

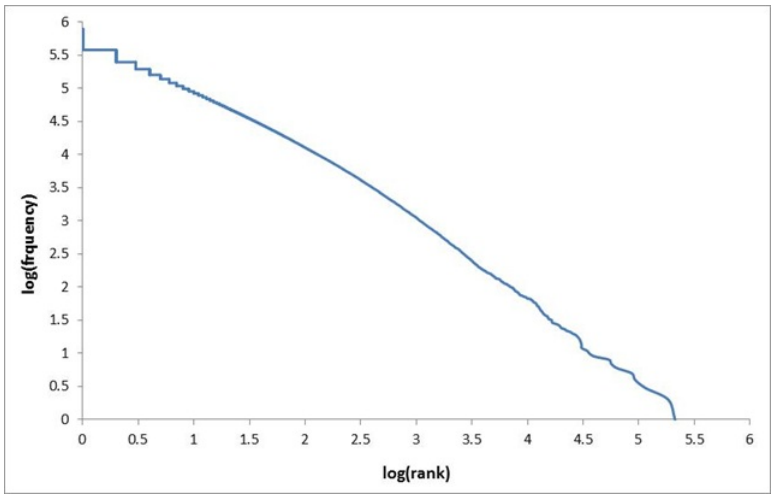

Figure 5.3: Zipf's law of Levantine, slope= -0.972. log of rank on the X axis versus frequency on the Y-axis

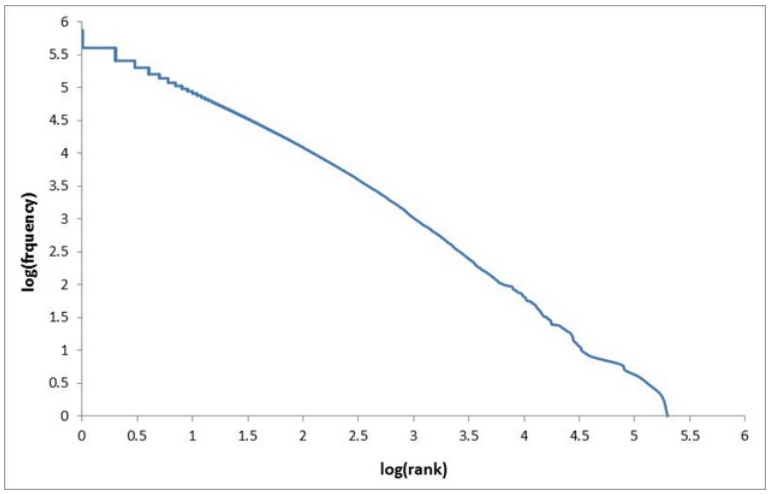

Figure 5.4: Zipf's law of North African, slope = -0.9793. log of rank on the X-axis versus frequency on the Y-axis

Dialect	Count of / million					
	bigram	unique bigram	trigram	unique trigram	five-gram	unique five-gram
Egyptian	11.8	4.8	10.6	6.3	8.4	5.6
Gulf	12.9	5.3	11.3	6.8	8.4	5.5
Levantine	9.3	4.3	8.2	5.3	6.2	4.3
North African	9.1	3.9	8.1	4.9	6.3	4.1
Our four corpora	43.1	15.1	38.2	21.5	29.3	18.8

Table 5.12: Bigrams, trigrams and five-grams counts

corpora, we have about 43 million bigrams, about 38 million trigrams and just less than 30 million five-grams. Remarkably, as the Table 5.12 shows, the bigrams have the greater count while having the lowest count with respect to unique bigrams (that is, about 15 million bigrams). The extraction of n-grams in written corpora is very useful, since they highlight the main features of the language/dialect.

The commonest bigram and trigram for all the dialects corpora are shown in Tables 5.13 and 5.14 respectively.

Frequency	Bi-gram	Meaning
25835	شاء الله	Allah willing
23056	السلام عليكم	the peace upon you
18251	الله عليه	Allah upon him (the meaning is incomplete)
17542	آ آ	A A
16851	و لا	and not
15563	محمد بن	Mohammed bin (the meaning is incomplete)
15400	صلى الله	Allah bless (the meaning is incomplete)
14804	ورحمة الله	and Allah's mercy
14400	عبد الله	Abd Allah (Abdallah)
14306	عرض الموضوع	display the topic

Table 5.13: Commonest bigram for all dialects corpora

Frequency	Trigram	Meaning
15591	آ آ آ	A A A
14953	صلى الله عليه	may Allah bless him
14197	انواع عرض الموضوع	thread display options
13701	ان شاء الله	if Allah wills
12867	عليكم ورحمة الله	upon you and mercy of Allah (the meaning is incomplete)
11921	ورحمة الله وبركاته	and Allah's mercy and blessings (the meaning is incomplete)
11693	الله عليه وسلم	Allah upon him and peace (the meaning is incomplete)
11036	السلام عليكم ورحمة	may the peace upon you (the meaning is incomplete)
6945	الله الرحمن الرحيم	Allah the most merciful, most compassionate
6864	بسم الله الرحمن	in the name of Allah the most merciful

Table 5.14: Commonest trigram for all four corpora

5.5.3 Error evaluation

There are a large number of derivations of each word for MSA. When the new prefixes and suffixes that appear in the dialects have been added, they will give a huge number of forms for each word, which might then produce a lot of spelling mistakes. Also many Arabic words differ in just النقاط 'the marks[1]', these words are more likely to have spelling mistakes than other words e.g. على 'upon' علي 'Ali (proper noun)', where the difference in just the marks of the last letter.

Another issue is that there are web pages using MSA as their main syntax. However, some of the words from the dialects could appear in these web pages. Because of this, we found that there are many MSA pages in the corpus, as these are very difficult to remove automatically. Generally, it is better for a corpus to have MSA syntax in addition to the main dialect; every dialect has MSA sentences within the dialect itself. Additionally, one last issue we found is that many web pages included mixed dialects, which might cause noise when managing more than one dialect concurrently.

5.6 Conclusions

This chapter has explained the methodology that have been used for building and evaluating an Arabic multi-dialect written corpora by the seeding of words from the dialects in the web. These corpora are for the main Arab dialects (Gulf, Levantine, Egyptian and North African). The chapter introduced the following points:

1. Five steps were involved in the Methodology (as will be described in points 2 through

[1]The marks are necessary for distinguishing different letters i.e. five dots, the short Kaf, three Hamzas and the Madda (Habash, 2010).

6 below).

2. Dialect words from different Arabic language websites have been collected and grouped together; about 1500 words were collected.

3. A survey was conducted with a group of people from different Arab countries to categorise the words into four main dialects; about 1000 words have been categorised.

4. An estimation of the counts is needed before the downloading stage to calculate the average for how many tokens will be produced per link.

5. Subsequently, web pages were downloaded that were likely to have the same dialects using a search engine API. More than 55000 web pages were downloaded.

6. Suitable cleaning and normalisation for the downloaded web pages was performed to remove unwanted HTML tags, symbols, and repeated sentences from resulting corpora.

7. We obtained a result of 14.5 million, 10.4 million, 13 million and 10.1 million tokens for the Gulf, Levantine, Egyptian and North African dialects, respectively; the total number of unique types in all corpora is more than two million types.

8. The results of the Zipf's law for all corpora were $\approx$ -1, which is predicted by Zipf's law.

9. We discussed the errors that were noticed in corpora, including spelling mistakes, mixed dialects and syntax mistakes.

Chapter 6

Multi Dialect Arabic Speech Parallel Corpora

6.1 Introduction

In a manner similar to the shortage of written corpora, there is a shortage of speech corpora in Arabic speech processing tasks. This shortage becomes a large obstacle for Arabic researchers in the area of speech processing when looking at available dialect resources.

In the survey conducted by Nikkhou and Choukri (2004), the participated companies listed the reasons why companies choose not to buy Arabic language resources. Some of these reasons are; (1) the data, if it is available, are too expensive, (2) the data do not meet the standard requirements, and (3) some of them do not buy the Arabic language resources because they do not have a high quality.

The building of a speech corpus is a time consuming process and without available corpora each researcher then needs to start building his own corpus, therefore it is beneficial to attempt to build a high quality speech corpus for Modern Standard Arabic (MSA) and dialects as this will give researchers a chance to engage in Arabic Natural Language

Processing (NLP) tasks.

For the purposes of this work, we have introduced the methodology for collecting multi-dialect Arabic speech parallel corpora. These include three main dialects; Gulf, Levantine, Egyptian as well as MSA. Four parallel text versions have been written to produce four parallel speech corpora. We began the process by writing MSA text. After this, we have translated the diacritised MSA text into the three other dialects. Then we have recorded the text by native speakers from different dialects. The final step was the segmentation of the speech files to identify the boundaries of sentences.

This chapter is organised as follows: Section 6.2 answers why do we need multi dialect speech corpora; Section 6.3 presents the methodology used to determine the procedure followed; Section 6.4 describes the process of implementation; Section 6.5 shows how the files are organised; Section 6.6 presents our results of the work; Section 6.7 highlights an evaluation of the work; and conclusions are presented in Section 6.8.

6.2 The need for Arabic multi dialect speech corpora

Many NLP applications such as Automatic Speech Recognition (ASR), Text To Speech (TTS), speaker identification tasks etc. require speech files to use in training. Any shortage in this data i.e. text or speech data, will result in deficiencies in the final product.

The importance of parallel speech corpora is well known for speech processing tasks and in the field of NLP. Parallel speech corpora are used as valuable resources for bi-lingual or bi-dialect NLP tasks. By using speech parallel corpora we can highlight and link the differences between dialects/languages in phonetics, grammar, etc. In addition, by making use of speech parallel corpora, multi-dialect/language tasks can be produced more easily than by using one dialect corpus especially for those languages that have a large gap between original language and current dialects such as Arabic. Parallel speech

corpora will also be helpful in speech-to-speech translation tasks.

The alterations in stems, phones, and word orders between MSA and the dialects and between the dialects, as argued in Chapter 2, make it so important to use dialect-based corpora instead of MSA-based corpora when building dialects applications. It is interesting to produce parallel corpora in both text and speech. By using parallel corpora, we can clearly distinguish the difference between MSA and dialects and dialects themselves. We can also make use of these differences by applying them in NLP tasks such as parsing dialects, or any speech tasks.

The lack of available Arabic speech corpora makes it very important for researchers to produce more and more resources for MSA and dialects. To date, no multi-dialect speech corpora is available for researchers and if one were to be developed it would offer considerable opportunities for researchers wishing to work on Arabic NLP and speech processing tasks.

6.3 Methodology

The methodology that has been used is similar to the methodology that has been used in the Multext-East project (Erjavec, 2004), where the project started with English text which was used to produce other language texts, and then the texts were recorded.

Four main steps have been included in the methodology. The first step involved writing MSA text, which is written with the correct rules of MSA. A text should contain words and sentences unrelated to any other dialect. Different sentence lengths should be included. Once the text are established, we then diacritise the MSA. In the diacritisation stage the short vowels will be shown. The letters themselves do not show the short vowel where Arabic uses just three vowel letters.

Diacritisation is not an easy step, as there is a need for sophisticated knowledge of the rules of MSA. These rules are not easy even for native speakers, as there are multiple

Original phone in MSA	Dialect	New phone
/f/ ف	in loanwords- all dialects	/v/
/q/ ق	in most of the dialects	/g/
/j/ ج	in Egyptian	/g/
/k/ ك	in Gulf	/t͡s/
/q/ ق	in Gulf	/d͡z/

Table 6.1: New phones representation in dialects

options available; i.e. 13 different possible marks for each letter, with some exceptions for vowel letters.

Once the diacritised MSA text is made, we moved on to translate it into the other dialects to produce parallel texts. We need to convert MSA text to another dialect, then diacritise it, and we repeat the process with the next dialects.

One of the two following cases will be chosen while converting MSA phones into dialects.

1. If the phone in the MSA word is converted to another phone, already used in MSA, we use the new rather than the old phone; e.g. when converting the word نقول /naqul/ (MSA) to انئول /inAul/ (Levantine) 'we say' using in this example /A/ which is used in the Levantine dialect rather than /q/.

2. If the phone in the MSA word has been converted to a new phone that is not used in MSA phones, we use a new phones as suggested in Table 6.1 to represent the new phones. According our work we found four new phones that have been used in dialects and not in MSA.

Step 3 relates to recording. In the recording stage we are going to record the text for each dialect according to special specifications that have been determined from which to produce a standard corpus. All the recordings have been done by native dialect speakers.

In step 4 the segmenting takes place, where it is necessary here to produce segmented files. Each file contains no more than one sentence as is presented in the text. Each segmented file should begin and end with silence.

We chose a specific domain in this work which is travel and tourism. Travel and tourism domain has a clear vocabulary, which can be pronounced by any native speaker easily. To split this domain into parts, we suggested eight general sections. Four sections directly related to travel and tourism: restaurant, hotel, transport and street and shopping; and the remaining four are necessary for tasks related to this area: days and times, currency, global cities and numbers.

The numbers section has an MSA numbers subsection which is for testing the differences between Arab speakers in how they pronounce MSA phones. So MSA numbers will be spoken by all participants, i.e. Levantine, Egyptian and Gulf people. The benefit of this operation was to build an MSA numbers corpus to describe different Arab background speakers. Each Arab speaker can speak MSA as their second dialect. However, as a conversion happened with the dialects by collecting MSA speech files we were able to understand and study the differentiation in phonetics between Arab speakers.

The research applied conditions suggested by TIMIT (Garofolo et al., 1993). Thus, from 70% to 80% of the information will be used for training and 20% to 30% of the corpus will be used for testing purposes (*ibid.*).

All dialects should be represented in both training and testing sets. Each speaker that appears in the training set should not appear in the test set.

6.4 Implementation

In the implementation stage, we started by writing MSA text that cover the main terminologies related to our specific domain, then diacritise it, then translate it, then diacritise the dialects texts, then record and segment speech files.

6.4.1 Write MSA text and diacritise it

We began by writing the MSA text with a total of 1291 sentences distributed across eight sections, and 18 subsections, as Table 6.2 shows. Out of 325 numbers, 222 were MSA numbers that did not require translation into other dialects. Therefore the three other dialect texts had 1069 per dialect with a total of about 4500 sentences for all. We have made use of the text that have been used in Saudi Accented Arabic Database (SAAVB) (Alghamdi et al., 2008) for writing days and times, currency, global cities and MSA numbers.

All of the words and sentences were to be evaluated according to MSA rules after diacritisation. The first column of Table 6.3 shows some examples of MSA diacritised sentences.

In our speech corpora, most parts have sentences. However, a few parts have single words and non-sentences, such as single numbers, greetings and farewells.

6.4.2 Translate into dialects and diacritise them

The MSA text should then be converted into the other dialects to be covered. The translating step is an important one when needing to produce parallel dialect texts. As stated previously, we left out the MSA number section as we wanted all the participants

Section	Count of subsections	Sentences
Restaurant	5	207
Hotel	4	137
Transport	2	90
Street and shopping	2	111
Days and times	1	207
Currency	1	42
Global cities	1	172
Numbers	2	325
Total	18	1291

Table 6.2: Corpora distribution for sections

MSA	Gulf	Levantine	Egyptian	Meaning
لا أَسْتَطِيْعُ سَمَاعَ صَوْتِك	مَا أَقْدَر أَسْمَعْك	مَا بَأْدِرْ أَسْمَعْ صُوتَك	أَنَا مُش سَامِعْ صُوتَك	I cannot hear you
مِنْ فَضْلِكَ أُرِيْدُ قَهْوَة	لَوْ سَمَحتْ، أَبِيْ قْهَوة	لَوْ سَمَحتْ بِدِّي أَهْوة	لو سَمَحْتِ عَايِزْ اَهْوَه	Please, could you bring coffee
أُرِيْدُ التَّحَدُّثَ مَعَ المسئُول	أَبْتَكَلِّمْ مَعْ المُدِيْر	بِدِّي أَحْكِي مَعْ المَسْؤُوَل	عَايِزْ اَتْكَلِّم مَعَ اَلْمَسْئُول	May I speak with in charge person
ثَمَانِيَةَ عَشَر	ثِمَانْ طَعَشْ	اِتْمَنْطَعَش	تَمَنْتَاشَر	Eighteen

Table 6.3: Example of some sentences

Sampling rate(Hz)	Sample format	Channels	Output
48,000	16-bit	1 channel (Mono)	Linear PCM

Table 6.4: Recording attributes

to speak using MSA rather than their local dialects. Table 6.3 in the second, third and fourth columns shows Gulf, Egyptian and Levantine examples respectively, all having been translated from MSA alongside their meaning, as given in the last column.

Having developed diacritised texts in four dialects, we then built statistical Language Model (LM) for every dialect, that is, four LMs. These LMs are used as a baseline LMs. As described in Chapters 8 and 9, additional LMs are produced for dealing with different parts of the word.

6.4.3 Recording

We used a Blue Yeti microphone for recoding. A cardioid pattern[1] was used for recording, and the recording attributes values were set as shown in Table 6.4. We used Audacity (Mazzoni and Dannenberg, 2006), to record our corpora.

The recordings were made in quiet rooms in case when background noise occurred, the sentence was re-recorded.

[1]Blue Yeti has four different patterns; cardioid, omnidirectional, bidirectional and stereo. When recording in cardioid pattern, sound directly in front of the microphone is picked up while the sound at the rear and sides of the microphone is not picked up (Yeti, 2011).

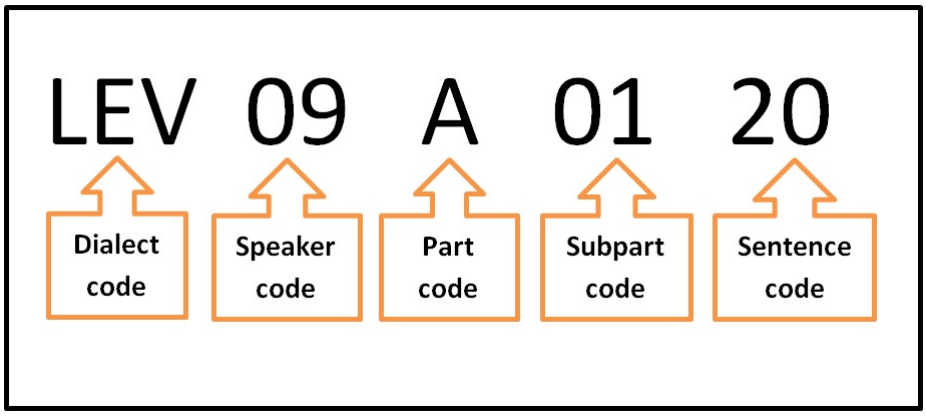

Figure 6.1: An example of how the files are organised

Every participant recorded full text in his/her dialect. The text lasts about 37 minutes in length in average, as the length would be different between different speakers.

6.4.4 Audio segmenting

Every participant recorded the entire text without segmentation. A transcriber tool (Barras et al., 2001) was used to segment each audio file into sentences and to then produce segmented speech files. About half a second from the start and end of each audio file there was silence for both parties. Each audio file has one sentence, which must be same as the diacritised text sentence.

6.5 File organisation

Figure 6.1 shows an example of the file structure. The main directory includes the dialects folders, i.e. MSA, Gulf, Levantine and Egyptian. Each dialect has a code pointing to its folder. Each speaker's folder includes files for all text for his/her dialect. 1291 segmented files per speaker folder. The part code points to the section, e.g. restaurant, hotel, shopping and street. There were several subparts inside, e.g. in the restaurant part, there are greetings, bill, services etc. The wave file was shown as a sentence number.

Dialect	MSA	Gulf	Levantine	Egyptian	For all
Tokens	2790	2146	2273	2356	9565

Table 6.5: Tokens count

6.6 Results

6.6.1 Parallel texts results

Table 6.5 shows token counts for the text for each dialect. The total size of the resulted texts is close to the 10K tokens. The average number of tokens per dialect is close to 2.5K, as Table 6.5 shows.

Table 6.6 shows the resulting parallel texts for sentences in the multi dialect corpora. 1291 sentences for MSA. 1069 sentences per dialect, where they give 222 for MSA numbers. The sentences for all dialects totalled about 4.5K sentences.

MSA	Gulf	Levantine	Egyptian	All
1291	1069	1069	1069	4498

Table 6.6: Parallel texts sentences count

6.6.2 Parallel speech results

Table 6.7 shows the number of speakers of each dialect. The total number of participate in our corpora was 52 speakers, with 12, 12, 8 and 20 speakers for MSA, Gulf, Levantine and Egyptian dialects respectively.

Age groups of the participating speakers are as shown in Table 6.8. The majority of

Speaker count	MSA	Gulf	Egyptian	Levantine	Total
Men	12	12	18	7	49
Women	0	0	2	1	3
Total	12	12	20	8	52
%Total(about)	23%	23%	38.5%	15.5%	

Table 6.7: Speaker count

Speaker age	MSA	Levantine	Gulf	Egyptian	Total
under 16	1	0	1	4	6
16-30	11	6	11	15	43
31-60	0	2	0	1	3
Total	12	8	12	20	52

Table 6.8: Speaker age

MSA	Gulf	Egyptian	Levantine	files total
15492 files	15492 files	25820 files	10328 files	67132 files

Table 6.9: Files count

speakers were between 16 and 30 years old. Only 6% of the participants were over 30 years, and about 12% were under 16 years.

Table 6.9 shows the files count after the segmenting stage. The total number of files is more than 67,000 speech files, including MSA and three other dialects i.e. Gulf, Egyptian and Levantine dialects.

Table 6.10 shows the utterance counts for our text sheets for the different dialects. The total utterances count for all dialects numbered about 160K utterances as Table 6.10 shows.

If we suppose each Arabic letter produces one phone. The total of phones that have been created in our corpora is about 800k phones, as can be seen in Table 6.11.

Dialect	Speaker count	Utterance count/speaker	Utterance count/all
MSA	12	2790	33,480
Gulf	12	2531	30,372
Egyptian	20	2702	54,040
Levantine	8	2643	21,144
total	52	10,666	160,180

Table 6.10: Utterances count

Dialect	speaker count	Phones count/speaker	Phones count/all
MSA	12	15,505	186,060
Gulf	12	14,290	171,480
Egyptian	20	15,088	301,760
Levantine	8	14,898	119,184
Total	52	59,781	778,484

Table 6.11: Phones count

6.7 Evaluation

6.7.1 Text evaluation

As can be seen in Table 6.12, which shows whole sentence sharing between four parallel corpora, there is a little sharing between MSA and the other dialects except for the names of cities. Most of the names given to cities are identical, even for different languages, so it is a normal result to got shared names between different dialects.

For MSA and all three other dialects the sharing is no more than 1% when excluding global cities, and is 14% if they are included. The greatest sharing as can be seen in Table 6.12 is between MSA and Egyptian, which reached 26% overall, and 11% when excluding global cities. Although this pair has the greatest commonality, there are some phonemes different in their pronunciations between MSA and Egyptian, such as /j/ which is spoken as /g/ in Egyptian dialect.

The lowest commonality in our four parallel corpora is between Levantine and the Gulf. When discounting global cities, just 15 sentences out of 897 include sharing between Gulf and Levantine, as Table 6.12 shows.

Table 6.13 shows a unigram depicting sharing between MSA and Gulf, and Egyptian and Levantine. The table confirms Kirchhoff and Vergyri (2005)'s result, which shows sharing between MSA and Egyptian was about 10.3%. Although our texts are parallel, overlapping between words did not exceed 20% in the best case.

	All	MSA * GULF	GULF * LEV[a]	LEV * EGY[b]	MSA * LEV	MSA * EGY	GULF * EGY
Restaurant	4	10	7	12	11	11	4
Hotel	5	16	8	17	15	26	13
Transport	0	0	0	3	2	18	0
Street and shopping	0	0	0	8	9	18	0
Days and times	0	12	0	3	2	13	0
Currency	0	0	0	13	11	22	0
Global cities	140	159	145	149	153	165	154
Numbers	0	0	0	10	6	7	0
Total	149	197	160	215	209	280	171
%ALL	14%	18%	15%	20%	20%	26%	16%
%ALL excepting global cities	1%	4%	1%	6%	5%	11%	2%

[a]LEV = Levantine.
[b]EGY = Egyptian.

Table 6.12: Sharing sentences between four parallel corpora

	MSA * Gulf	MSA * Egyptian	MSA * Levantine
Including cities names	13%	14%	19%
Excluding cities names	7%	8%	14%

Table 6.13: The word overlap for MSA with Gulf, Egyptian and Levantine

The sharing results, which have been shown in Tables 6.12 and 6.13, are supported by the variety between MSA and the dialects, and the dialects themselves. We can observe that the sharing percentage is very low in most cases between the dialects. Thus, for the same texts when we tried to translate them into other dialects, we got low results.

Table 6.14 shows the lexicon size for each dialect. The average lexicon size per dialect is about 1000 tokens, if we ignore the diacritisation, and 1300 with diacritisation. The average of the increasing percentage between undiacritised and diacritised text is about 28%.

In many cases, possibly most cases, the word in MSA differs from the dialect word in the word shape or in phones or both, as discussed in Chapter 2 in relation to varieties.

	Unique undiacritised tokens	Unique diacritised tokens	Percentage increase
MSA	1087	1481	36%
Gulf	1056	1315	25%
Levantine	1000	1154	15%
Egyptian	1032	1385	34%
For all	2145	3988	86%

Table 6.14: Lexicon count

These include, as stated, the difference between MSA and the dialects and between the dialects themselves.

As is conspicuous from Table 6.14, we acquired more than 2000 unique tokens for all undiacritised texts and close to 4000 unique tokens for diacritised texts, representing an increase of 86%. The difference in diacritisation gives different words meaning and also different phonemes. This suggests that one challenge for Arabic is its morphology, which in some way relates to the diacritisation.

6.7.2 Speech evaluation

For the recordings, we were able to obtain high-quality sound by using the following techniques: (1) recording in conditions that were as quiet as possible; (2) removing any recording that had noise in the background; (3) using a professional microphone for recording, i.e. Blue Yeti. Figure 6.2 shows a wave example of the recording of several words. We obtained high-quality sound in the recordings, which yielded results comparable to those that could be obtained in a soundproof room.

To evaluate the speech (foreground) and the noise (background) in the corpora, we checked the contrast between them. The contrast is measured to ensure that any noise or music in the background is very quiet. According to WCAG2.0 (2008)[2], the background

[2]Web Content Accessibility Guidelines (WCAG) 2.0 is a guideline for accessible audio files on the internet, which is recommended by W3C.

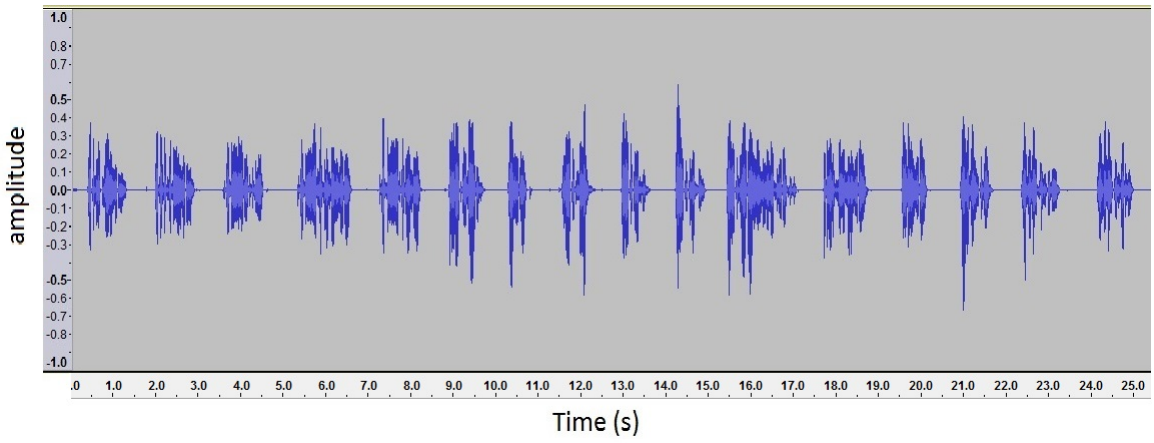

Figure 6.2: wave example

noise must be at least 20 rms[3] db lower than the content of the foreground speech. We randomly chose 30 speech files, and then we measured the contrast. The results of this measurement are shown in Table 6.15. The differences between the foreground and the background were over 20 rms db. The average difference over all was 35.7 rms db, which achieved WCAG conditions and indicated that the noise in the background of our corpora was at an acceptable level.

One issue when building the speech corpora was the difficulty in making the output from different speakers uniform. However, to solve this issue the output should be between 0.3 and -0.3 in sound amplitude as far as possible for the speaker; we accepted 0.5 to -0.5, but no greater variation than 1.0, -1.0 as this causes distortion, where it will be out of the range, Figure 6.2 shows an example of the output range.

To evaluate errors we completed a full review of all recordings to check the errors made with phonemes; i.e. changing or removing one or more phonemes inside the recorded sentences. However, any obvious errors, such as recording the wrong sentence or the

[3]rms = Root Mean Square.

Audio file #[a]	Foreground			Background			Results - Difference
	Time started (s[b])	Time ended (s)	Average rms dB	Time started (s)	Time ended (s)	Average rms dB	Average rms dB.
MSA\01\A\01\02	0.82	3.67	-30.1	3.71	3.97	-61.7	31.6
MSA\09\D\02\42	1.00	3.22	-23.2	0.39	0.96	-63.7	40.5
MSA\05\B\02\09	0.46	2.77	-22.3	0.02	0.44	-48.0	25.7
MSA\10\A\04\06	0.35	2.64	-25.2	0.00	0.33	-54.5	29.3
MSA\07\A\05\01	0.24	1.50	-16.8	0.00	0.24	-52.7	36.0
MSA\12\H\02\02	0.23	1.00	-16.5	0.00	0.19	-49.6	33.0
MSA\02\C\02\02	0.53	2.76	-23.8	2.80	2.97	-58.2	34.4
MSA\03\A\05\10	0.50	1.67	-22.6	1.77	1.96	-55.0	32.4
GULF\01\A\01\05	0.80	1.49	-22.8	1.64	2.05	-60.7	38.0
GULF\03\C\02\12	0.36	0.85	-28.2	0.19	0.35	-55.1	26.9
GULF\10\G\01\07	0.23	0.72	-21.4	0.08	0.23	-50.5	29.1
GULF\12\D\01\07	0.36	1.08	-24.2	1.09	1.18	-53.2	29.0
GULF\02\A\03\03	0.53	1.74	-30.1	1.75	2.06	-58.3	28.2
GULF\08\D\02\05	0.41	1.42	-21.5	1.44	1.73	-59.6	38.0
GULF\06\A\01\01	0.31	1.42	-23.3	0.00	0.30	-54.4	31.1
GULF\07\E\01\03	0.31	0.85	-23.6	0.86	0.99	-45.7	22.1
LEV\01\A\01\02	0.43	3.88	-24.9	3.92	4.16	-57.9	33.0
LEV\02\C\02\02	0.60	2.24	-20.3	2.27	2.66	-59.3	39.0
LEV\04\E\01\04	0.38	0.86	-18.8	0.88	1.16	-42.9	24.0
LEV\05\H\02\08	0.60	1.23	-24.8	1.35	1.65	-50.8	26.0
LEV\06\A\01\01	0.47	1.65	-20.5	0.03	0.46	-54.3	33.8
LEV\07\B\02\02	0.44	1.79	-24.7	1.84	2.17	-60.5	35.8
LEV\08\E\01\06	0.60	1.17	-20.6	1.21	1.46	-52.8	32.2
EGY\01\A\02\02	0.96	2.59	-25.8	2.66	3.25	-69.8	44.0
EGY\04\C\02\02	0.77	2.51	-24.8	2.56	3.17	-66.9	42.1
EGY\06\E\01\04	0.64	1.21	-20.8	1.30	1.81	-69.4	48.6
EGY\09\B\02\02	0.71	2.30	-23.8	2.41	2.97	-84.1	60.3
EGY\12\H\01\01	0.73	1.13	-21.5	1.21	1.53	-59.7	38.2
EGY\15\D\02\01	1.69	2.51	-25.0	2.61	3.15	-80.1	55.1
EGY\18\B\04\13	0.67	2.18	-22.0	2.25	2.88	-75.2	53.2
Average	—	—	-23.13	—	—	-58.82	35.7

[a]See Figure 6.1 to find out how the files are organised.
[b]Seconds.

Table 6.15: Speech contrast evaluation

wrong word, or making any mistake when recording the sentences, led to them being rerecorded at the recording stage.

Most Arabic native speakers face difficulties when they try to speak using MSA. These difficulties relate to phonetics, and sometimes unfamiliar words, rather than grammar; e.g. Egyptian people pronounced many of the ث /θ/ phonemes with the س /s/ phoneme instead. Gulf people also tended to use the ظ /ðˤ/ instead of the ض /dˤ/ phoneme.

Many errors occurred in cases where the speakers were under 18 years of age. Arabic dialects are mainly spoken and not written, so younger speakers faced difficulties when reading dialect text. Most of the errors that happened were with MSA numbers for speakers of all ages. Also many errors occurred when reading city names.

6.8 Conclusions

In this chapter, we have introduced the methodology we have followed to build the parallel dialects speech corpora for Arabic. We have introduced the following points:

1. The corpora are about 32 hours in length produced by 52 speakers, and are for MSA and three other dialects, i.e. Gulf, Levantine and Egyptian.

2. We chose a specific domain, i.e. travel and tourism. This specific domain has been divided into eight sections.

3. We have started by writing MSA text containing 1291 sentences followed by the diacritisation stage, which is an important step after writing the MSA text, and it should follow the true rules of the Arabic language.

4. MSA diacritised text was translated by native dialect speakers into local dialects, where MSA is the second dialect for every Arabic native speaker. Then they have diacritised all the translated texts.

5. We were able to obtain a high-quality sound similar to a sound-proofed room by avoiding any outside noise, recording in very quiet conditions, removing recording of noise in the background, as also by using a professional microphone for recording.

6. All the speakers from different countries participated in the MSA number sections enabling us collect an MSA number speech corpus from speakers with different backgrounds.

7. We have used Transcriber software to segment the recordings to match our texts. The total number of segmented files is more than 67,000 sound files.

8. This is the first available multi-dialect Arabic speech corpus for researchers.

Chapter 7

A Comparison of Arabic Speech Recognition for Multi-Dialect vs. Specific Dialects

7.1 Introduction

Most Arabic speech recognition tasks focus on either Modern Standard Arabic (MSA) or a single specific dialect, rather than on multi-dialect tasks. There are two principal reasons for this: First, the large gap between MSA and the dialects; and second the limited availability of speech resources for Arabic dialects. Prior to the Arabic multi-dialect speech parallel corpus there were no parallel speech corpora to facilitate a comparison between dialects and MSA in Arabic Automatic Speech Recognition (ASR) systems.

The limited availability of current Arabic speech resources is an important issue affecting the accuracy of Arabic ASR tasks. Many research studies that have been conducted to improve Arabic ASR have relied on in-house corpora. Most in-house corpora have three main characteristics: first, small data size, which renders them insufficient to provide reliable quality for training ASR tasks; second, the quality of recordings is less than

the quality of the speech corpora for other languages; and third they do not follow any method of standardisation[1].

Two important factors need to be considered when dealing with dialects or languages classifiers: duration and accuracy. Time is an important issue as, according to Arslan and Hansen (1996), classification accuracy becomes higher when the test utterance length increases. However, for multi-dialect or multi-language systems we can avoid this time constraint. Although current research on dialect classifiers for Arabic dialects has proven that we can obtain a high quality result, the action to be taken if the classifier retrieves an incorrect result is unclear. This is especially pertinent in the case of languages that exhibit a large gap between dialects, such as Arabic.

Before starting to improve Pronunciation Dictionaries (PDs) and Language Models (LMs) and to reduce Out-Of-Vocabulary (OOV) in Chapters 8 and 9, in this chapter, we will check the accuracy of the speech data collected and explained in Chapter 6. We will need to apply several experiments involving multi dialect tasks to check the multi dialect task results, and to compare them to separated dialect results when checking them against the same data.

This chapter aims to answer two main questions: First, can high accuracy be achieved by building a multi-dialect Arabic speech recognition system using parallel speech corpus or, in other words, how much will training on a multi dialect corpus affect the Word Error Rate (WER) in a multi-dialect recognition system? Second, which of the following provides more accuracy: (a) pooled data in one engine or (b) data obtained separately for each dialect? If the answer is the use of individual dialect's data, it is important to determine whether the use of a dialect classifier might then be helpful. For example, if the WER for Gulf dialect is 15% when using Gulf data for training and 50% when using multi-dialect data, then it is clearly necessary to use a dialect classifier and avoid

[1]See Chapter 6 for more details.

multi-dialect data as a large difference has occurred between the two results. Issues of time elapsed to attain a result and inaccurate results are also discussed in relation to each option, in Sections 7.4 and 7.5.

This chapter is organised as follows: Section 7.2 shows the data used. Section 7.3 presents the recognition system. Section 7.4 show the results that were obtained. Section 7.5 provides a discussion of the work. Finally conclusions are presented in Section 7.6.

7.2 Data

We used MSA, Gulf, Levantine and Egyptian dialects from the multi-dialect Arabic parallel speech corpus. These sub-corpora included approximately 32 speech hours segmented into more than 67,000 files. For that research, 52 participants were divided into groups based on four dialects i.e. 12 participants for MSA and Gulf, 8 participants for Levantine and 18 participants for Egyptian. The recordings have high quality sound giving very similar results to those that could be obtained in a sound proof room. The vocabulary size is close to the 10K tokens, where the average number of tokens per dialect is close to 2.5K.

For the experiments in this research 90% of speech data was used for training and 10% for testing all dialects.

The multi-dialect Arabic speech parallel corpus includes speech files, transcriptions and the trigram language models; however, they do not contain PDs. So, we have created a baseline PD, which facilitates one-to-one auto mapping between Arabic letters and phones, as Table 7.1 shows.

7.3 Recognition system

Recognition experiments are accomplished using the CMU Sphinx (Lee et al., 1990). The results of training have been obtained using CMU Sphinxtrain v1.0.7 (Sphinxtrain, 2011).

ء	AH	ح	H7	ع	A3
آ	AE2	خ	X	غ	GH
أ	AH	د	D	ف	F
ؤ	UW	ذ	DH	ق	Q
إ	EH	ر	R	ك	K
ئ	AH	ز	Z	ل	L
ا	AE	س	S	م	M
ب	B	ش	SH	ن	N
ة	T	ص	S9	ه	HH
ت	T	ض	D9	و	W
ث	TH	ط	T6	ى	AE
ج	JH	ظ	DH6	ي	Y

Table 7.1: One-to-one auto mapping for creating a baseline PD

CMU Sphinx v3-0.8 (Sphinx, 2009) has been used for decoding and extracting the results.

7.4 Results

> Modern speech recognition algorithms are based on computing observation probabilities directly on the real-valued, continuous input feature vector. These acoustic models are based on computation of a Probability Density Function (PDF) over a continuous space. By far the most common method for computing acoustic likelihoods is the **Gaussian Mixture Model (GMM) PDFs** (Jurafsky and Martin, 2009).

Table 7.2 shows the results of a multi-dialect system that contains three main dialects i.e. Gulf, Egyptian and Levantine as well as MSA. We have set the number of densities and tied states to obtain the optimal value.

Young and Woodland (1994) suggested one of the most common algorithms, which is done by clustering and tying acoustically similar states. This method is called tied states.

Densities	Tied States	WER%
8	1000	20
8	2000	16.8
8	3000	15.9
8	4000	15.2
8	4500	14.9
4	4500	17.4
16	4500	13.7

Table 7.2: The WER variation with the tied states and the number of densities in multi-dialect system using all four corpora

Tied states is used in speech recognition tasks because the problem of data sparsity, where we need to reduce the number of tri-phone parameters that are needed to train (Jurafsky and Martin, 2009). Here is an example, "the beginning of a phone with an [n] on its left may look much like the beginning of a phone with an [m] on its left. We can therefore tie together the first (beginning) subphone of, say, the [m-eh+d] and [n-eh+d] triphones. Tying two states together means that they share the same Gaussians. So we only train a single Gaussian model for the first subphone of the [m-eh+d] and [n-eh+d] triphones. Likewise, it turns out that the left context phones [r] and [w] produce a similar effect on the initial subphone of following phones (*ibid.*)".

As can be seen from Table 7.2 the best WER that can be obtained is 13.7%[2]. It has 4500 tied states and 16 for densities, as Table 7.2 shows. We set the values for the tied states from 1000 to 4500 and the number of densities' values from 4 to 16, as these values are the most appropriate for the data size that we have used for training. We set the values of the tied states first, and then we moved them to obtain the optimal value of the number of densities.

Table 7.3 shows the best WERs for the four dialects. For these experiments, we used multi-dialect data i.e. multi-dialect acoustic and language models to extract each dialect's

[2]Optimal numbers of densities and tied states depend on the corpus. They can be determined by experiments.

Dialect	WER%
MSA	10
Gulf	17
Levantine	15.1
Egyptian	16.3
Average	14.6

Table 7.3: The best WERs for the four dialects when evaluated using multi-dialect data

Dialect	WER%	The improvement comparing to multi-dialect data results –Table 7.3
MSA	8.2	-1.8
Gulf	12.7	-4.3
Levantine	8.8	-6.3
Egyptian	11.2	-5.1
Average	10.2	-4.4

Table 7.4: The best WERs for the four dialects when evaluated against each dialect's own data

WER separately, by using the optimal values that were previously obtained. As can be seen from Table 7.3 we obtained WERs of 10%, 17%, 15.1% and 16.3% for MSA, Gulf, Levantine and Egyptian respectively.

Table 7.4 contains the best WERs that could be obtained for each dialect, using the dialect's own data rather than the multi-dialect data. Table 7.4 also shows the difference between the results of each dialect comparing its data with the results that could be obtained using multi-dialect data. The average of the WERs for the four dialects is -4.4%, as shown in Table 7.4.

Table 7.5 shows the result of an experiment that was done to extract Levantine WER using MSA acoustic models. In this experiment, we assume that we have a dialect classifier and it has inaccurately classified Levantine as MSA. The experiment has been repeated with three different LMs; MSA LM, Levantine LM and multi-dialect LM to check and compare the results in different cases. As Table 7.5 shows, the WER obtained using MSA

LM used	WER% for Levantine dialect
Levantine	47.3
Multi dialect	48.7
MSA	78

Table 7.5: The table shows Levantine dialect results when evaluated using MSA acoustic model with three different LMs

LM is 78%, and the WERs obtained when using Levantine LM and multi-dialect LM are 47.3% and 48.7%, respectively.

7.5 Discussion

As shown in Table 7.2, the best WER that could be obtained for the multi-dialect system is 13.7%. This result cannot be readily be compared to results from other research, as there are differences in the sizes of resources for speech corpora that have been used for each research. However, this WER is the lowest when compared to WERs from research focused on Arabic cross-dialects such as Kirchhoff and Vergyri (2005); Biadsy et al. (2012). In Kirchhoff and Vergyri (2005) the best WER that could be obtained is 41.4% for MSA and Egyptian dialect and in Biadsy et al. (2012) the best WER is 20.4% for Jordanian and Lebanese dialects, which are both Levantine dialects.

The usage of a specific domain, as in the multi dialect speech corpus rather than a general dictation task definitely affects the WER. Further study is required to check whether the use of parallel corpora might be useful for reducing WER or otherwise for multi-dialect tasks.

Table 7.4 shows the best WERs for the four dialects. The lowest difference in results was found in MSA, at just -1.8%. There are two reasons for this: (1) MSA is at the centre of the dialects, and although there are large differences between MSA and dialects, MSA shares more common language features with the dialects than they do between themselves;

Mean (pooled data)	14.6
Std. Deviation (pooled data)	3.17
Mean (separated data)	10.23
Std. Deviation (separated data)	2.10
P-value	0.03

Table 7.6: A Student's t-test result

and (2) there is more MSA data in the multi-dialect speech corpus where, as previously stated, the number section is spoken in MSA by all the participants.

The average between the best WER for the multi-dialect system and the best WERs for the separate dialects is -4.4%, as shown in Table 7.4. To identify the significance in differences between the using of separated data and pooled data, a Student's t-test between these two groups has been done. The t-test illustrates that the experiments results from the separated data group were significantly smaller than the results from the pooled group, where p-value for 1-tailed test < 0.05, as Table 7.6 shows.

Now, we turn to determine whether the dialect classifier failed to classify the dialect in a proper way. In other words, did the dialect classifier recognise one dialect as another dialect? Table 7.5 shows the results of the experiment that has been done. We assumed that we had a dialect classifier and that Levantine was the target dialect. However, the dialect classifier recognised the Levantine dialect as MSA. We used three different LMs with MSA acoustic model: using MSA LM, using multi-dialect LM and using Levantine LM. The result actually was not a surprise when we understand that a large gap exists between MSA and the dialects and between the dialects themselves. As Table 7.5 shows, the WER is 78% when using MSA LM. We remember that the result of Levantine dialect when it was evaluated with multi-dialect data was 15.1%. When using multi-dialect LM with MSA acoustic model we obtained a WER of 48.7% with a difference of -33.6%, when checking on multi-dialect data.

In this case, although the use of the dialect classifier increases the accuracy by an

average of 4.4%, and the state-of-the-art of Arabic dialect classifier gave an Equal Error Rate (EER)[3] of 4% (Biadsy, 2011), a large difference between dialects occurs in Arabic. This means that in a case when the dialect classifier makes a wrong classification, very low accuracy results would be obtained.

In the state-of-the-art of Arabic dialect classifier the author used 30 seconds utterances to obtain the lowest EER (Biadsy, 2011). So for example if we assume that in dialogue between different dialects speakers we need 30 seconds for each conversion from dialect to other. However, since the multi-dialect system does not require dialect classification, this is not a concern.

7.6 Conclusions

This chapter presented an Arabic dialects speech recognition system that was created to recognise MSA and three different dialects; It introduced the following points:

1. Most of the research studies that have been conducted in this area have focused on MSA or one dialect rather than on multiple dialects. One important reason for that is the shortage of available resources for Arabic speech, especially for dialects. Therefore, we have done multi dialect speech recognition system, using multi dialect speech corpora.

2. The best WER that could be obtained is 13.7% for multi-dialect system and 10%, 17%, 15.1% and 16.3% the WERs for MSA and for the Gulf, Levantine and Egyptian dialects, respectively when using multi dialect data.

3. We extracted the best WER for each dialect, by using the dialect's own data to compare the results with the multi-dialect data. The results obtained are: 8.2%, 12.7%, 8.8% and 11.2% for MSA, Gulf, Levantine and Egyptian dialects, respectively.

[3]Which is error rate when false alarm rate and miss probabilities rate are equal (Martin et al., 1997).

4. The average of the differences when checking the dialects using their own data and the best WER of the multi-dialect result is -4.4%.

5. To determine the importance of the dialect classifier in such situations, an experiment was conducted to extract the Levantine dialect result, assuming that the classifier wrongly classified the Levantine dialect as MSA. We extracted the WERs for three different LMs using an MSA acoustic model. The WERs were found to be very low compared to the WER for the Levantine dialect when evaluated against multi dialect data in all the cases that were examined.

6. The result confirms that although the state-of-the-art Arabic dialect classifier obtained a high accuracy, in the case of misclassification, a very low accuracy would be obtained for languages that have significant dialect differences like Arabic.

7. Multi-dialect systems might be judged an optimal solution when thinking that the loss of -4.4% in the average rate of accuracy and conversely saves at least 30s, according to the state-of-the-art, with each conversion between dialects of real-time dialogue systems.

Chapter 8

An Incremental Methodology for Improving Pronunciation Dictionaries for Arabic Speech Recognition

8.1 Introduction

Pronunciation Dictionaries (PDs) are one of the most important components in Automatic Speech Recognition (ASR) systems. They contain representations of phones for words. The main task of a PD in ASR system is to map words with their phones.

For the English language, many PDs have been manually created, such as the CMU Pronouncing Dictionary (Rudnicky, 2007) and the Voxforge lexicon. For Arabic it is very difficult to do this where from each single Arabic root hundreds of different full word forms can be obtained.

Modern Standard Arabic (MSA) orthography to phonology is unlike English, French or other languages which have a complex relationship between orthography and word

sound. MSA, in most cases, follows regular rules of pronunciation. However, these rules linking orthography to phonology need to be tested to establish which will improve the PD for MSA and the dialects.

In this chapter, Arabic orthography to phonology rules will be tested by the incremental methodology to improve Arabic PDs. The incremental methodology will be applied to MSA and Levantine PDs. Phonology and morphology features will be used during the incremental cycles.

This chapter is organised as follows: Section 8.2 answers the question why do we need to improve Arabic PDs; Section 8.3 focuses on the data that has been used; Section 8.4 presents the incremental methodology applied for improving Arabic PDs; Sections 8.5 and 8.6 show the recognition system and the results that have been obtained; Section 8.7 provides an evaluation of the work; and conclusions are presented in Section 8.8.

8.2 Why do we need to improve Arabic pronunciation dictionaries?

It is very difficult to manually create PDs for languages with rich morphology such as Arabic. This would be too labour-intensive and time-consuming especially for Arabic dialects, where many new stems, prefixes, suffixes and a large gap between dialects and MSA[1]. Therefore, our research objective is to find an automatic method for building and improving PDs.

As previously discussed, Arabic phonology follows regular letter-to-sound links in most cases and this would simplify the building of automatic PDs. However, there are three aspects in Arabic orthography to phonology which need to be considered: (1) there are nevertheless some irregular rules in Arabic phonology; (2) even for regular rules of phonology, more experiments are needed to check which of these rules have more effect; (3) it is

[1]See Chapter 2 for more details.

necessary to ascertain how these rules would affect the dialects.

8.3 Data

MSA and Levantine dialect data from multi dialect speech corpora were used to evaluate this work. MSA corpus has 12 speakers and Levantine corpus has 8 speakers. The speaker's recording lasts about 37 minutes in length in average, as the length would be different between different speakers. 10% of the corpora are used for testing and 90% for training.

The multi dialect speech corpora include speech files, transcriptions and the trigram language models; however, they do not contain PDs (lexicons).

8.4 Methodology

The first step required is to create a baseline PD since the multi dialect speech corpora do not have PDs. The text in the multi dialect speech corpus is diacritised, so there is no need for using the Buckwalter analyser (Buckwalter, 2002)[2], Morphological Analysis and Disambiguation of Arabic (MADA) (Habash et al., 2009)[3] or any other morphology analyser or disambiguator for diacritising or selecting the word choice. We start from diacritised lexicon to represent phones. Table 8.1 shows the baseline one-to-one auto mapping between Arabic letters and phones including diacritic representation.

Most Arabic letters have an one-to-one relationship with their phones, with few exceptions. An example of an exception is Taa Marboutah ة /T3/, which was used as a separate phone, thus separating it from ت /T/ and هـ /HH/, and differentiating between its usage as ت /T/ and هـ /HH/. Later, it was established that it is more appropriate to

[2]Buckwalter analyser lists all possible analyses of the word (Buckwalter, 2002).
[3]MADA selects the analysis that matches the current context (Habash et al., 2009).

AH	ء	DH	ذ	M	م
AE2	آ	R	ر	N	ن
AH	أ	Z	ز	HH	ه
UW	ؤ	S	س	W	و
EH	إ	SH	ش	AE	ى
AH	ئ	S9	ص	Y	ي
AE	ا	D9	ض		
B	ب	T6	ط	Diacritisation	
T3	ة	DH6	ظ	AA	Fatha
T	ت	A3	ع	AN	F. Tanween
TH	ث	GH	غ	IH	Kasrah
JH	ج	F	ف	EN	K. Tanween
H7	ح	Q	ق	UH	Dhammah
X	خ	K	ك	UN	D. Tanween
D	د	L	ل	2	Shaddah

Table 8.1: An one-to-one automapping for creating baseline PD

use Taa Marboutah; the ت /T/ phoneme, هـ /HH/ phoneme or the newly proposed phoneme /T3/, and which gives a lower Word Error Rate (WER).

An additional issue when thinking to build a PD automatically is combinations of short vowel phones. For example the word يَتَعَلَّمُ /yataʕllamu/ 'he is learning' has 5 letters. However, it is represented according to orthography and phonology rules using 10 phones Y AA T AA A3 AA L2 AA M UH. This representation includes 5 consonants i.e. Y T A3 L2 M and 5 short vowels i.e. 4 AA's and UH. The number 2 was used to refer to Shaddah 'emphasis', rather than duplicating the consonant letter, so 2 was attached

to the letter that has Shaddah 'emphasis'; for example بّ /B2/ is /B/ with Shaddah 'emphasis'. We cannot follow the same procedure for all diacritics because this will produce 13 different phones for every letters, totalling about 350 in all.

8.4.1 The pronunciation dictionary rules

Most of the orthography to phonology rules used for this work were selected from (Alghamdi et al., 2004). These selected rules affect isolated words. Furthermore, some new rules were established which are not related to orthography or phonology MSA rules. The advantage of adding these rules was to check if we can improve WER by using phonology rules other than the real rules; for example, the usage of Shaddah 'emphasis' in this work produced many new phones, therefore it follows that the WER might be improved if Shaddah 'emphasis' is removed.

In chapter 2 we have listed three Arabic phonology rules. Here we are going to recall them.

1. The Arabic letter ة /t/ or /h/ 'Ta-Marboutah letter' is used only at the end of a word. It can take one of two phonemes: when stopping it is pronounced هـ /h/, and when the speaker connects the word containing Ta-Marboutah letter with the following word, it is pronounced ت /T/. The ة Ta-Marboutah letter rule is similar to the short vowel rule for most letters at the end of the words where they take one of two phonemes; one with stopping i.e. Sukun 'no vowel', and one for continuation which is one of the thirteen different forms.

2. The /L/ phoneme in the definite article ال /AL/ can be divided into use with the 14 sun letters, and with the remaining moon letters. In the sun letters the /L/ phoneme will be ignored (it is silent). However, with the moon letters it is pronounced. As an example, the word الشمس /Alšams/

'the sun' is pronounced /aššams/ by deleting the /L/ phoneme; however, القمر /Alqamar/ 'the moon' is pronounced as it is written.

3. The Alef-Alwasl is pronounced either ء /'/ when starting with it, or ignored when it is linked to the previous word. For example, in the word البيت /'lbayt/ 'the home' we say /'/ at the beginning as it is written because we started with it. However, when we say بالبيت /bilbayt/ 'in the home' we removed the phoneme /'/ from the word البيت /'lbayt/ 'the home' because of its position.

We have tried to list the phonology rules in a specific order to avoid interference with the preceding rules. This point is very important because some rules have to be applied before or after others are added. Moreover, the effect for every rule will be relatively variable according to its place in this order. Table 8.2 shows the phonology and morphology rules in suggested order. From 11 phonology rules, 2 rules cannot be combined together; rule 9 and rule 10. A or B will be chosen, or both will be ignored If the new WER has not improved. There is one morphology rule, which is to extract the word stem for the wrong results and compare it to the original stem from the PD; in order to do this it is necessary to use the methodology of the multi dialect morphology analyser to extract the words stems.

8.4.2 The incremental methodology for improving Arabic pronunciation dictionary

Figure 8.1 shows the incremental methodology used for improving PDs. The incremental cycle for improving PDs starts by establishing the baseline PD. First, the WER for baseline, using the letter to phone map in Table 8.1, is extracted for the recognition system. By taking the first suggested rule and applying it to the PDs, the new result is obtained.

No	Linked	Description
		Phonology Rules
1	-	Remove Fatha before Hamza
2	-	Remove Fatha after Hamza
3	-	Remove Fatha before Alef Alwasl, Alef Maddah and Alef Maqsurah
4	-	Remove Kasrah before Yaa
5	-	Remove Dammah before Waw
6	-	Remove Kasrah before ﻹ /EH/ letter
7	-	Remove shadda
8	-	Remove L before sun letters
9	A	Change AE to AH at the beginning of the word
9	B	Remove Alef Alwasl at the beginning of the word
10	A	Change Taa Marbuta to /HH/
10	B	Change Taa Marbuta to /T/
11	-	Remove /AE/ phone before and after Tanween Fatha
		Morphology Rule
12	-	Compare the wrong words stems with the stems of the PD words.

Table 8.2: PD phonology and morphology rules

If the new WER has not improved, then this rule is ignored and the next suggested rule is applied. However, if the new WER has improved, then this rule is added to the PD improvement rules and the next suggested rule is applied, etc. This process is carried out for all phonology and morphology rules in both MSA and the Levantine dialect PDs.

By the end of the implementation, two PD improvement rules for both MSA and the Levantine dialect are presented.

8.5 Recognition system and baseline result

Recognition experiments are accomplished using the CMU Sphinx (Lee et al., 1990). The results of training have been obtained using CMU Sphinxtrain v1.0.7 (Sphinxtrain, 2011). Sphinx v3-0.8 (Sphinx, 2009) has been used for decoding and extracting the results.

The baseline WERs obtained were 9.9% and 11.3% for MSA and Levantine respectively, as shown in Tables 8.3 and 8.4.

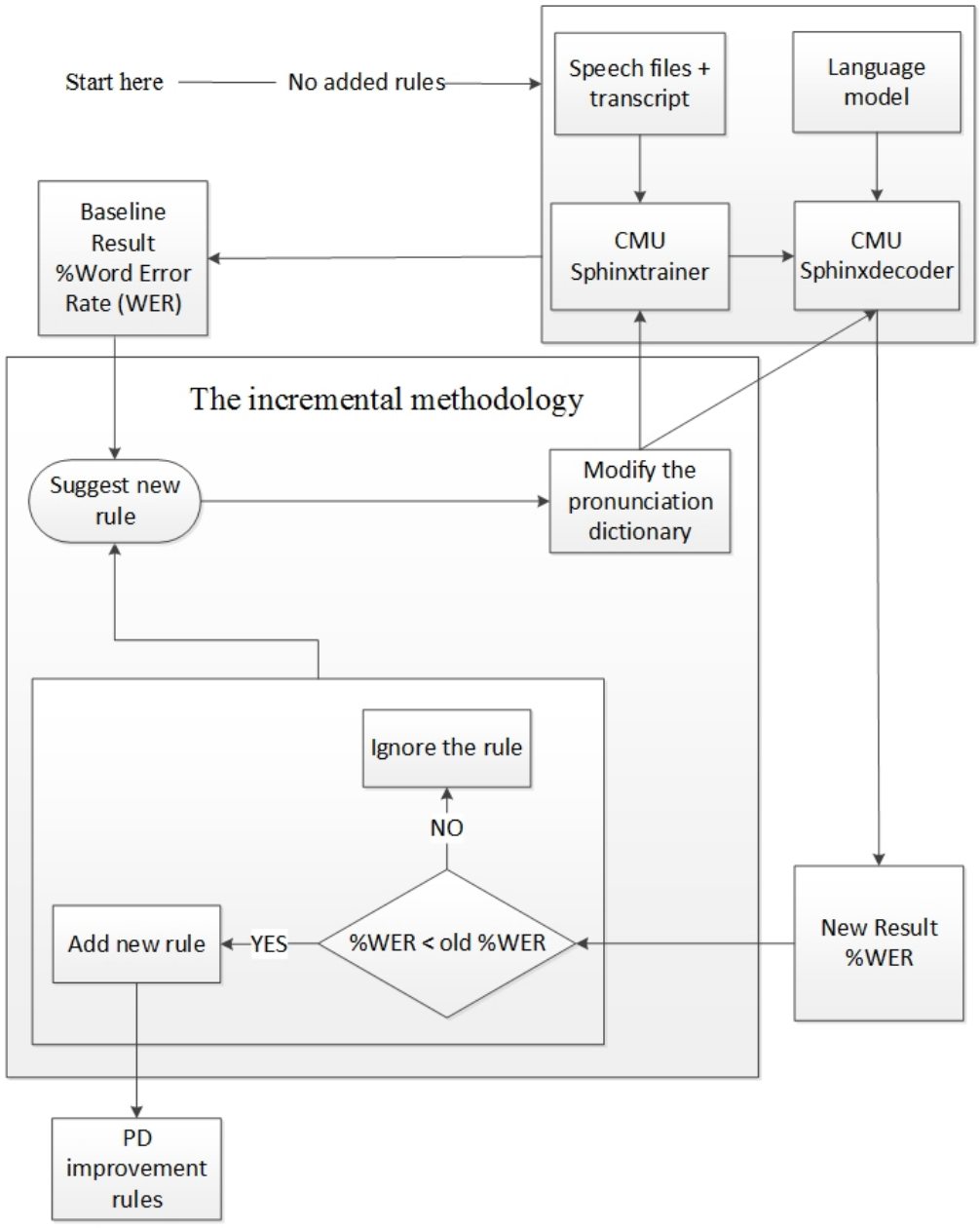

Figure 8.1: The Incremental Methodology Algorithm for improving Arabic PD in ASR

PD change rule														Result			Improvement	
1	2	3	4	5	6	7	8	9A	9B	10A	10B	11	12	Accuracy	WER[a]	RC[b]	INCR[c] IMP[d]	Total IMP
Baseline														90.1	9.9	-	-	-
?[e]														90.6	9.4	5.1	0.5	0.5
+[f]	?													91.1	8.9	10.1	0.5	1
+	+	?												91.6	8.4	15.2	0.5	1.5
+	+	+	?											92	8	19.2	0.4	1.9
+	+	+	+	?										92.1	7.9	20.2	0.1	2
+	+	+	+	+	N[g]									-	-	-	0	0
+	+	+	+	+		?								92.7	7.3	26.3	0.6	2.6
+	+	+	+	+		+	?							92	8	19.2	-0.7	1.9
+	+	+	+	+		+		?A						92.2	7.8	21.2	-0.5	2.1
+	+	+	+	+		+			?B					92.9	7.1	28.3	0.2	2.8
+	+	+	+	+		+			+B	?A				92.5	7.5	24.2	-0.4	2.4
+	+	+	+	+		+			+B		?B			**93.3**	**6.7**	**32.3**	**0.4**	**3.2**
+	+	+	+	+		+			+B		+B	?		93.2	6.8	31.3	-0.1	3.1
+	+	+	+	+		+			+B		+B		+	**95.4**	**4.6**	**53.5**	**2.1**	**5.3**

[a]Except for the morphology rule, where we extracted Stem Error Rate (STER).
[b]RC = Relative Change.
[c]INCR = Incremental.
[d]IMP = Improvement.
[e]?: check the rule through the incremental cycle.
[f]+: the rule has been added to the PD improvement rules.
[g]N: no results have been retrieved.

Table 8.3: MSA Result

8.6 Results

The rules set out in Section 8.4 are applied and evaluated through the incremental cycles described above. The systems were retrained and decoded for every iteration. Tables 8.3 and 8.4 show the process used for adding each rule. As discussed, some rules cannot be combined. Where both or one achieved positive results it was possible to use one, however, if both achieved negative results then neither was used.

As can be seen from Tables 8.3 and 8.4, each cycle checks new rule and calculates the WER: if the WER had improved, the rule is added in the following cycles; if not, it is

PD change rule														Result			Improvement	
1	2	3	4	5	6	7	8	9A	9B	10A	10B	11	12	Accuracy	WER	RC	INCR IMP	Total IMP
Baseline														88.7	11.3	-	-	-
?														88.5	11.5	-1.8	-0.2	-0.2
	?													88.9	11.1	1.8	0.2	0.2
	+	?												89.4	10.6	6.2	0.5	0.7
	+	+	?											**91.2**	**8.8**	**22.1**	**1.8**	**2.5**
	+	+	+	?										90.2	9.8	13.3	-1	1.5
	+	+	+		?									90.6	9.4	16.8	-0.6	1.9
	+	+	+			?								90.7	9.3	17.7	-0.5	2
	+	+	+				?							89.9	10.1	10.6	-1.3	1.2
	+	+	+					?A						89.1	10.9	3.5	-2.1	0.4
	+	+	+						?B					90.7	9.3	17.7	-0.5	2
	+	+	+							?A				89.9	10.1	10.6	-1.3	1.2
	+	+	+								?B			90.6	9.4	16.8	-0.6	1.9
	+	+	+									?		89.9	10.1	10.6	-1.3	1.2
	+	+	+										+	**93.7**	**6.3**	**44.2**	**2.5**	**5**

Table 8.4: Levantine Result

ignored, and the process continued. Tables 8.3 and 8.4 also show the effect of each added rule compared with the previous result and the improvement in the overall result. Table 8.3 shows the MSA results, and Table 8.4 shows the Levantine results.

As can be seen from Table 8.3, the MSA PD has improved as a result of 8 rules plus the morphology rule. The WER was improved by 3.2% by adding phonology rules and by 2.1% by adding the morphology rule. The total improvement for MSA was 5.3% which reduced the baseline WER/STER from 9.9% to 4.6%; this was the best WER/STER achieved i.e. an accuracy of 95.4%.

Table 8.4 shows that 3 phonology rules as well as the morphological rule improved the PD WER/STER for Levantine. The total improvement obtained for the Levantine recognition system was 5% in the WER, where the addition of both phonology and morphology rules reduced the WER/STER by 2.5%. The best WER/STER for the Levantine dialect was reduced from the baseline 11.3% to 6.3% i.e. an accuracy of 93.7%, as can be

seen in Table 8.4.

8.7 Evaluation

For MSA six rules resulted in an improvement of approximately 0.5% for each, and two rules yielded an improvement of just 0.1% and 0.2%. Both A's for rules 9 and 10 had negative results, and both B's resulted in improvements to the WER. Consequently, B's were chosen for rules 9 and 10 for the MSA PD.

Two negative results achieved for MSA PD. It cannot be guaranteed that any Arabic phonology rule will definitely improve a PD, so as can be noticed two rules in MSA and eight rules in Levantine dialect, have a negative or neutral impact on the WER. This illustrates the importance of the incremental methodology usage in such work, whether for real phonology rules or any new suggested rules.

The improvement obtained by adding the morphology rule was 2.1% for the MSA PD and 2.5% for the Levantine PD. This morphology rule clearly indicates the importance of using the stem rather than the full word form. Stem usage will reduce unique words. It does affect adjacent conjunctions, prefixes, and suffixes etc.

Although all of the rules mentioned thus far originated from MSA phonology rules, 3 rules resulted in improvements of 0.2%, 0.5%, and 1.8% for the Levantine dialect. The highest WER improvement of 1.8% resulted from rule 4, however, it is still not clear for us why this rule "remove Kasrah before Yaa" has such a positive effect.

Rules 2, 3 and 4 resulted in improved PDs for both MSA and the Levantine dialect. These three rules improved the Levantine dialect PD by 2.5% and the MSA PD by 1.4%.

Thus, for the automatic building of PDs for MSA, Levantine or any other dialects, there is a chance of improvement WER by introducing more rules and then checking them using an incremental cycles. One of the important observations here is that each dialect should have its own improvement rules that help when creating PD automatically.

There is a need for suggesting rules, order them and then test them in an incremental methodology.

8.8 Conclusions

In this chapter a method for improving Arabic PDs for MSA and Levantine dialect by incremental applying of phonological rules has been introduced. The chapter has introduced the following points:

1. MSA generally follows regular rules from letter to sound; however, these rules still need to be evaluated in order to ascertain which are more efficient.

2. It is worthy to suggest new rules and check them during the incremental cycles for MSA and dialects.

3. In the methodology we began by applying a new rule to the baseline system, then started a new cycle and evaluated the effect of this rule. If there was an improvement, we added this rule to the PD improvement rules for each dialect.

4. By applying phonology rules, the improvements in WER were 3.2% and 2.5% for MSA and Levantine respectively; the morphology rule also improved the STER by 2.1% and 2.5%, respectively.

5. Considerable improvements in WER, i.e. 5.3% and 5% for MSA and the Levantine, respectively, were recorded.

6. We proved that not all of the orthography to phonology rules improved WER, especially in the case of dialects.

7. There are new rules that affect WER efficiency. Those rules need to be extracted and then tested.

8. Eight phonology rules have improved the MSA PD, and three for the Levantine dialect.

9. The morphology rule illustrated the importance of using the stem rather than the full word form in ASR tasks.

Chapter 9

Morpheme-Based Language Models for Improving the Speech Recognition of Arabic Dialects

9.1 Introduction

The goal of the Language Model (LM) is to compute the probability of a sentence or sequence of words (Jurafsky and Martin, 2009). Given sequence of words w1 through wn, a LM is a set of probabilities P(W), where each sequence W = (w1, ..., wn):

$$P(W) = P(w1, w2, w3, w4, w5...wn)$$

This represents the statistical model of a word based on its history. The computation of the probabilities of the LM is challenging, due to the fact that many word contexts are observed rarely, or not at all (Kirchhoff et al., 2006).

In morphologically rich languages every stem generates hundreds of different word forms, which therefore creates considerable issues with LM, Arabic (Kirchhoff et al., 2006), Turkish (Carki et al., 2000), and Czech (Ircing et al., 2001) are all examples of those

languages with a high growth rate of vocabulary, which results in a large number of Out-Of-Vocabulary (OOV) words

In Arabic, many different prefixes and suffixes can be attached to the stem, which then produce hundreds of words for each single stem. This issue becomes more complex when discussing Arabic dialects, where there are many more affixes, and a large gap occurs in stems between Modern Standard Arabic (MSA) and dialects[1].

A common solution for OOV in morphologically rich languages is to deal with the morpheme rather than the full word form. For example by using the CALLHOME Egyptian speech corpus (Canavan et al., 1997), Billa et al. (1997) remove the definite article /il/ 'the' in the Egyptian dialect, so reducing the size of the vocabulary by 7%. This reduction is created by removing just one prefix out of many prefixes and suffixes, thus showing the fact that when using morpheme-based (rather than the full form) techniques the unique words will be fewer in number and the Word Error Rate (WER) will thus be reduced.

The LM has a considerable effect on WER in speech recognition. For example, by using an MSA LM for recognising Levantine dialect the WER obtained was 78%, which is high when compared to the 15.1% when using Levantine LM[2].

The aim of this chapter is to demonstrate the results of the experiments that have been obtained on dialects LMs and also to discuss the following four points:

1. Dealing with dialect morphemes in three different forms: stem alone, prefix+stem and stem+suffix;

2. Working on more than one dialect using parallel data, so facilitating a comparison between MSA and dialect features;

[1]See Chapter 2 for more details.
[2]See Chapter 7 for more details.

3. Using the methodology of the multi dialect morphology analyser for extracting affixes automatically (rather than manually);

4. Establishing the differences between using a closed domain LM and an open domain LM in Arabic dialects.

This chapter is organised as follows: Section 9.2 outlines the data used; Section 9.3 presents the methodology applied for dealing with Arabic dialects stems; Sections 9.4 and 9.5 demonstrate the recognition system and the results that have been obtained; Section 9.6 provides an evaluation of the work; Section 9.7 presents the resulting conclusions.

9.2 Data

MSA, Gulf and Egyptian dialects data from a multi dialect speech corpus were used as the speech data. 10% of the corpora were used for testing and 90% for training. The multi dialect speech corpora included speech files and transcriptions for each dialect. However, Pronunciation Dictionaries (PDs) were built automatically, according to the method suggested in Chapter 8.

Two types of text will be used to build the trigram language models, these being closed domain text and open domain text. The texts of speech corpora are used to create closed domain LMs. The domain of the multi dialect speech corpus is travel and tourism, including a number of sections that relate indirectly to travel and tourism, i.e. days and times, currency, global cities and numbers.

We collected a text corpus from the web[3]; it has four sub-corpus for four Arabic dialects. Fifty thousand words have been extracted as an open domain text from our corpus to represent the Gulf and Egyptian dialects. About fifty thousand tokens from the Giga word corpus (Parker et al., 2011) was chosen for MSA. Table 9.1 demonstrates the token counts for each dialect after extracting the unique words.

[3]See Chapter 5 for more details.

	Token count in				
Dialect	Text	Unique	50k text	Unique	The sources of open domain LMs
MSA	2788	1016	54105	16546	Giga word corpus
Gulf	2479	1079	50947	19054	Multi dialect text corpora
EGY	2695	1070	59006	21899	Multi dialect text corpora

Table 9.1: Description of the corpora used for creating closed and open LMs

Dialect	Full word	Stem	Reduction %	Prefix+ stem	Reduction %	stem+ suffix	Reduction %
MSA	1016	853	16	973	4	897	12
Gulf	1079	874	19	1034	4	933	14
Egyptian	1070	876	18	976	9	930	13
Average	—	—	18	—	6	—	13

Table 9.2: Reduction percentage in closed LM size- unique tokens

Tables 9.2 and 9.3 show the unique tokens' count after extracting the stems, prefixes+stems and stems+suffixes. This was done using a multi dialect analyser methodology.

The tables show, the highest reduction percentage was obtained from the web data, i.e. up to 34%. The reduction average for stem+suffix was higher than that for prefix+stem.

Table 9.4 shows the average LMs sizes used in relevant research for various morphologically rich languages. The size of the open LMs used for experiments in this work is 164K tokens.

Dialect	Full word	Stem	Reduction %	Prefix+ stem	Reduction %	stem+ suffix	Reduction %
MSA	16546	13376	19	13823	16	14721	11
Gulf	19064	12546	34	16008	16	14266	25
Egyptian	21899	14419	34	18491	16	16728	24
Average	—	—	29	—	16	—	20

Table 9.3: Reduction percentage in open LM size- unique tokens

Language/dialect (if applicable)	LM size	Reference
Czech	61K	(Ircing et al., 2001)
Turkish	87K	(Carki et al., 2000)
Arabic/Egyptian	170K	(Kirchhoff and Vergyri, 2005)
Arabic/MSA	240K	(Kirchhoff and Vergyri, 2005)
Multi language/Egyptian	160K	(Creutz et al., 2007)
Arabic	146K	(Kirchhoff et al., 2006)

Table 9.4: Comparison of LM sizes

9.3 Methodology

The multi dialect Arabic morphology analyser, that we have introduced in Chapter 4, uses a modified MSA morphology analyser in order to analyse dialect words whose affixes have been altered, and then uses the web as a corpus to extract unanalysed words according to their frequency. The hypothesis utilised by the multi dialect morphology analyser is that: (1) a large number of Arabic words have been changed according their affixes rather than their stems. (2) the usage of stems is larger than the usage of the full word forms.

For this work, the methodology of the second part of the multi dialect morphology analyser was used. The analyser required a minor modification in order to extract the prefixes+stems and the stems+suffixes.

To the author's knowledge, none of the previous studies have attempted to use more than one form of a word, i.e. combinations of morphemes. One of the benefits of using more than one form, is that in the case of similar results being obtained for different combinations; for example if the results of prefixes+stem are similar to those of the stem alone, it would be better to use prefixes+stem rather than the stem alone i.e. a larger combination. However, in the case of different results we need to determine which combination yield more accuracy and if the results agree between different dialects.

In order to enrich this research, experiments were performed on more than one dialect i.e. MSA, Gulf dialect and Egyptian dialect. When using the data from parallel dialects,

Dialect	WER for closed domain LMs	WER for an open domain LMs	Difference
MSA	13.9%	18.3%	+4.4%
Gulf	12.7%	16.2%	+3.5%
Egyptian	11.6%	18.2%	+6.6%
Average	12.7%	17.6%	+4.8%

Table 9.5: Baseline results

it will be possible to compare their parallel results, differentiations, and establish if the results of the combination of morphemes have been agreed between different parallel data for dialects or not.

Two types of LMs are also compared through the use of two different domains: (1) an open domain, and (2) a closed domain LM (i.e. travel and tourism). This gives the possibility of comparing the results of the same speech data with two different types of LMs.

9.4 Recognition system and baseline result

Recognition experiments are accomplished using the CMU Sphinx speech recognition tool (Lee et al., 1990). The results of training have been obtained using CMU Sphinxtrain v1.0.7 (Sphinxtrain, 2011). Sphinx v3-0.8 (Sphinx, 2009) has been used for decoding and extracting the results.

The baseline WERs as demonstrated in Table 9.5 result in 13.9%, 12.7% and 11.6% for MSA, Gulf and Egyptian respectively for the travel and tourism LMs, and 18.3%, 16.2% and 18.2% for the open domain LMs. The results that are shown in Table 9.5 are for the full form of the words i.e. prefix+stem+suffix. The average differences between the closed and open domain LMs is just under 5%.

	Error rate for		
LM	Stem	Prefix + Stem	Stem + Suffix
Closed domain LM	7.10%	7.80%	7.50%
The improvement in error rate	-6.80%	-6.10%	-6.40%
Relative change	48.9%	43.9%	46.0%
An open domain LM	11.80%	11.90%	11.50%
The improvement in error rate	-6.50%	-6.40%	-6.80%
Relative change	35.5%	35.0%	37.2%

Table 9.6: MSA recognition results

9.5 Automatic Speech Recognition (ASR) experiments results

Tables 9.6, 9.7 and 9.8 demonstrate the results of the experiments for MSA, Gulf and Egyptian dialects. The tables show WERs for three different cases: stem, prefix+stem, stem+suffix. They also demonstrate results for two different suggested LMs, these being a tourism and an open domain LM. The improvements in results compared to the baseline results for each dialect are also shown in these Tables.

Table 9.6 demonstrates the MSA results when the error rates[4] were reduced to between 6.1% and 6.8%. The best improvement for the closed domain LM is in the stem, i.e. an error rate of 6.8%. This result, however, is equal to the best improvement for an open domain LM that has been obtained by stem+suffix.

The WERs are reduced between 0.5% and 3.1% in the Gulf dialect. The stem alone in the Gulf dialect has obtained the best error rate for closed LM and open domain LM, i.e. error rates of 3.1% and 2.3% respectively, as demonstrated in Table 9.7.

Egyptian error rates have been reduced between 0.7% to 4.7%, as demonstrated in Table 9.8. The greatest improvements in error rates in the Egyptian dialect is in stem+suffix in both cases. They obtained 1.5% and 4.7% for closed domain and an open domain re-

[4]Including WER, Stem Error Rate (STER), Prefix + Stem Error Rate and Stem + Suffix Error Rate.

	Error rate for		
LM	Stem	Prefix + Stem	Stem + Suffix
Closed LM	9.6%	11.8%	10.0%
The improvement in error rate	-3.1%	-0.9%	-2.7%
Relative change	24.4%	7.1%	21.3%
An open domain LM	13.9%	15.7%	15.3%
The improvement in error rate	-2.3%	-0.5%	-0.9%
Relative change	14.2%	3.1%	5.6%

Table 9.7: Gulf dialect recognition results

	Error rate for		
LM	Stem	Prefix + Stem	Stem + Suffix
Closed LM	10.9%	10.4%	10.1%
The improvement in error rate	-0.7%	-1.2%	-1.5%
Relative change	6.0%	10.3%	12.9%
An open domain LM	15.2%	14.2%	13.5%
The improvement in error rate	-3.0%	-4.0%	-4.7%
Relative change	15.6%	22.0%	25.8%

Table 9.8: Egyptian dialect recognition results

spectively.

In conclusion, improvements from 0.5% to 6.8% have been obtained for all three dialects examined in these experiments. Out of six different cases, in half the stem has proved to be the most effective, while in the other half it is the stem+suffix. Although there were improvements in all prefix+stem cases, none obtained the best WER.

9.6 Evaluation

The aims of this work have been to: (1) extract prefixes and suffixes by automatic means; (2) work on more than one dialect using parallel data, in order to make a comparison between different dialects in comparable data: (3) examine dialect morphemes in three different forms: stem alone, prefix+stem and stem+suffix; (4) establish the differences between using closed domain LM and open domain LM in such tasks for Arabic dialects. There now follows a discussion of how far these aims have been achieved.

One of the important issues is that the majority of Arabic dialects do not have available lexicons with the exception of the Egyptian dialect (Abdel-Massih, 2011), and which has been used in many ASR projects, such as in Kirchhoff et al. (2006). In addition, there is no available speech tagger, or any other morphology tools, to extract affixes for multi dialect words apart from the multi dialect morphology analyser. The rich morphology of Arabic ensures that it is difficult to extract the morphemes of dialects words when such vital tools or resources are absent, apart from using manual methods, which are time-consuming. Therefore, the use of a multi dialect morphology analyser methodology has been both helpful and time saving. All the experiments that have been undertaken after using the multi dialect analyser methodology have yielded improvements in accuracy and the error rate has been reduced. This demonstrates the effectiveness of using a morphology analyser tool in such research.

There have been improvements in all eighteen results that have been completed. From the lowest WERs that have been obtained for each dialect (i.e. the best six results), three have been improved in stems, three have been improved by stem+suffix and none of the prefix+stem were given the lowest WER. For both cases of closed and open domain LMs in the Egyptian dialect, the improvements have taken place in the stem+suffix form. In relation to the Egyptian dialect, it is also notable that the improvement for even the prefix+stem is more than the stem improvements in both cases of LMs.

By dealing with morphemes instead of full word forms, fewer unique words were determined, thus leading to improved results in all three cases. However, the following question arises: Why did stem+suffix obtained the best WERs in three different cases? In order to answer this question, further research is required. However, with regard to the Egyptian dialect, a possible answer is that by working on the stem alone, the number of phonetically similar words increased, which also increased the possibility of recognition error. It may be that this phenomenon occurs in the Egyptian dialect more than in other

dialects, which would then cause greater improvement in the Egyptian dialect than in other dialects. For example, when /k/ phoneme is removed from the word سامعك /samʕk/ 'I hear you', the new word is سامع /samʕ/ 'I hear'. This word is phonetically close to the word سامح /samħ/ 'forgive'[5], and in this case the possibility of error is higher than in the case of stem+suffix. Thus, if improved (or similar) results are obtained by using stem+suffix or prefix+stem instead of the stem alone, the preference would be to use the stem with the better results of either prefix or suffix rather than the stem alone.

Given the supposition that there are the same number of prefixes and suffixes for Arabic, using prefix+stem or stem+suffix rather than prefix+stem+suffix will decrease full word forms by 50% in the lexicon for words that have affixes. In the case of similar results, this also will maintain a larger portion of the word than the use of the stem alone.

The final issue concerns the fact that work on the parallel dialect data enabled to check and compare between MSA and dialects. As previously discussed (and as can be can seen from result Tables 9.6, 9.7 and 9.8), there is no agreement between different dialects. This considerable gap between dialects themselves and MSA is behind the main reason for such a differentiation.

9.7 Conclusions

This chapter has discussed the improving of Arabic dialects speech recognition through working on LMs. We have done experiments on different types of morphemes. The chapter has outlined the following points:

[5]This word is also used as a personal name.

1. Three parallel dialects have been used i.e. MSA, Gulf and Egyptian from Arabic multi dialect speech corpora.

2. Two different LMs for each dialect have been produced: a closed domain (travel and tourism) LM and an open domain LM.

3. The methodology of the second part of a multi dialect morphology analyser has been used to extract the three suggested forms of the word; stem alone, prefix+stem and stem+suffix.

4. Six results per dialect have been extracted. This has given a total of eighteen results. The WERs in these results have been reduced between 0.5% to 6.8%.

5. In three out of six results the best WERs were in the stem, while in the other three the best WERs were in stem+suffix. None of the prefix+stem obtained the best WER.

6. Dealing with prefix+stem or stem+suffix rather than with prefix+stem+suffix will, in the case of similar results, keep a larger part of the word than the use of the stem alone.

Chapter 10

Conclusions and Future Work

10.1 Introduction

The central mission of this thesis has been to improve Arabic Automatic Speech Recognition (ASR) by dealing with Pronunciation Dictionary (PD) and Language Model (LM) for Arabic dialects. However, the only way to improve speech recognition systems is to provide access to sufficient data and to use the required tools. Therefore, before commencing working on the main aim of this thesis, important resources needed to be built and collected; i.e. a multi-dialect morphology analyser, a dialects text corpus and a dialects speech corpus. Three chapters of this thesis are devoted to detailing the methodologies of collection and building of these important resources. Section 10.2 concludes the methodologies that have been developed for building the Arabic dialects resources.

The improvement to the Arabic dialects speech recognition tasks was divided into two tasks: (1) comparing a multi dialect speech recognition task with separate dialect tasks built using the same data, and (2) improving speech processing methods for Arabic dialects according to two important components; i.e. LMs and PDs, Section 10.3 delivers conclusions about the work done in both areas.

10.2 The methodologies for building Arabic dialects resources

Chapter 4 described a multi-dialect morphology analyser, which utilises both a linguistic basis and a statistical basis to analyse Arabic dialects, loanwords and Modern Standard Arabic (MSA) words. The linguistic foundation makes use of the MSA morphology analyser and adapts it to accept affixes in dialects. The overall accuracy rate improved from 32% to 69% following adoption. One reason for this improvement was that many Arabic dialect words alter their affixes without any changes to their stems; thus, through treatment of the affixation a better result was acquired.

After this, the segmenter was then created to give four possible forms of the word. A full word form, a virtual (prefix + stem), a virtual (stem + suffix) and a virtual stem. It is possible that any one of these forms provides the correct stem, such that these examples were introduced previously in Chapter 4. By extracting the correct stem, a segmenter can also point if there is a prefix and/or suffix of the word.

By making use of the four forms created by the segmenter, the notion of using the web as a corpus was proposed. This resource made it possible to use the frequencies retrieved for each segment of the word to distinguish the stem and then extract the prefixes and suffixes where applicable. This step achieved a result of 94% accuracy over 2229 different dialect words.

The use of the web as a corpus made it possible to identify MSA words that were unrecognisable to the original MSA morphology analyser. This approach also showed that it was possible to distinguish between actual prefix and/or suffix and those letters that were similar to, but not actual prefixes/suffixes. The final advantage of this multi dialect morphology analyser was that the method was up to date, therefore any new dialect word would be detected immediately after becoming popular; i.e. if it has a frequency of

use exceeding ten thousand.

One of the shortcomings when using the web as a corpus for Arabic is that the use of Arabic search engines to search the web did not support diacritics effectively. In such work, there is also a need for more linguistic rules, as some analytical errors occurred due to the shortage of linguistic rules that could be applied to the web search portion of the task.

Chapter 5 and 6 presented the methodologies used for building text and speech corpora in Arabic dialects. The data resulting from both speech and text corpora is available for public use. Chapter 5 introduced a methodology for collecting an Arabic multi dialect based on written corpora, created by exploiting the web as a corpus. We started the work by collecting and grouping around 1500 dialect words from different Arab websites. Then, a survey was conducted with a group of people from different Arab countries to ensure that they would use only the words on their own list. The resulting corpus included four main dialects; Levantine, Egyptian, Gulf and North African. We obtained links for the resulting keywords using an API search engine, and then downloaded web pages deemed likely to have the same seed words dialects. More than 55000 web pages were downloaded.

We needed to perform a suitable cleaning and normalisation for the downloaded web pages, as the greatest concern issue was noise. It is difficult to generalise the rules for cleaning web data, as cleaning and normalisation are challenges for web data.

These corpora included four main dialects Gulf, Levantine, Egyptian and North African, which gave a result of 14.5 million, 10.4 million, 13 million and 10.1 million tokens respectively and the total number of distinct types in all corpora was more than 2 million. The results for the Zipf's law for all corpora were $\approx$ -1.

Multiple pages have mixed texts between MSA and a single dialect; this is no great concern as all dialects include some MSA words and expressions. However, it is a concern that many pages include mixed dialects, which then need to be classified in a proper way

to avoid them all appearing on one page.

In Chapter 6, we introduced the methodology we followed when building the parallel dialects speech corpora for Arabic. The resultant corpora were about 32 hours in length, and recorded from 52 participants, using MSA and three other dialects. The resulting segmented corpora included more than 67,000 wave files. A specific domain was chosen, i.e. travel and tourism before writing an MSA text containing 1291 sentences. We then diacritised the text showing short vowels that do not usually appear in Arabic text.

The MSA diacritised text was then translated by dialect native speakers into local dialects. There are additional phonemes in the dialects that are not used in MSA. Thus, we identified four new phonemes.

The recordings for the texts were made in very quiet conditions to avoid any background noise. Moreover, a professional microphone was used, i.e. a Blue Yeti microphone. The recordings obtained were of a high quality, similar to those that would be obtained in a sound proofed room.

All the speakers from different countries participated in the MSA numbered section. This made it possible to collect an MSA number speech corpus from those speakers with different backgrounds.

Arabic dialects are mainly spoken and not written; therefore, we found that many of the errors that occurred, happened with younger speakers as they faced difficulties when reading dialect text. Another issue identified was that most of the errors that happened with speakers of all ages were with MSA numbers in diacritisation, as Arabic native speakers face difficulties when trying to speak using MSA. We also found many errors when reading the names of cities.

To the author's knowledge this is the first parallel dialect speech corpus collected in Arabic. Therefore, this corpus will be useful for all researchers dealing with Arabic dialects, also for comparisons of the specific features of each dialect. In addition, it is

useful for speech-to-speech translation between MSA and dialects and between dialects themselves.

The resultant corpora have four parallel speech corpora MSA, Gulf, Levantine and Egyptian. There is also an MSA numbers speech corpus produced by native Arabic speakers from different dialect backgrounds.

10.3 Improving speech recognition for Arabic dialects

Chapter 7 presented an Arabic dialect speech recognition system created to recognise MSA and three different dialects. The majority of the research studies that have been conducted in this area have focused on MSA or a single dialect rather than on multiple dialects. An important reason for this is the shortage of available resources for Arabic speech, especially for dialects. Therefore, we aimed to build a multi dialect system before starting work on PDs and LMs to check the accuracy of the speech data collected, and also to compare multi dialect speech recognition task results, versus separated dialects results.

Of the experiments done using parallel speech data, the best Word Error Rate (WER) that could be obtained was 13.7% for multi-dialect system and 10%, 17%, 15.1% and 16.3% for MSA and for the Gulf, Levantine and Egyptian dialects, respectively when checking multi dialect data. However, when using the dialect's own data to compare the results with the multi-dialect data set, the results obtained were: 8.2%, 12.7%, 8.8% and 11.2% for MSA, Gulf, Levantine and Egyptian dialects, respectively. The average of the differences when checking the dialects using their own data and the best WER for the multi-dialect result was -4.4%.

To determine the importance of the dialect classifier in these situations, an experiment was conducted to extract the Levantine dialect, assuming that we have a dialect classifier and it has inaccurately classified Levantine as MSA. We extracted the WERs for three

different LMs using an MSA acoustic model. These were found to be very high compared to the WER for the Levantine dialect, when evaluated against multi dialect data in all the cases that were examined. The result confirms that although a state-of-the-art Arabic dialect classifier obtained a high level of accuracy, in the case of misclassification, a very low accuracy rate would be obtained for languages with significant dialect differences like Arabic.

Multi-dialect systems might be judged an optimal solution when considering a loss of 4.4% in the average rate of accuracy, conversely saves at least 30s according to state-of-the-art result with each conversion occurring between dialects in real-time dialogue systems.

Chapter 8 introduced a method for improving Arabic PDs. Arabic generally follows regular rules from letter to sound; however, these rules still need to be evaluated in order to ascertain which are the most efficient. Therefore, an incremental methodology for applying phonological rules was introduced for MSA and Levantine dialect. In the suggested incremental methodology, we began by applying a new rule to the baseline system, then started a new cycle and evaluated its effect. If there was an improvement, we added the rule to our PD improvement rules for each dialect.

Considerable improvements in WER, i.e. 5.3% and 5% for MSA and the Levantine, respectively, were recorded. The phonology rules improved the WERs by 3.2% and 2.5% for MSA and Levantine respectively; whereas the morphology rule improved the WERs by 2.1% and 2.5%, respectively. Not all the orthography to phonology rules had positive effects on WER, especially in the case of dialects. On the other hand, new rules were evaluated and found to have a positive effect on WER efficiency.

Eight phonology rules were shown to improve the MSA PD, and three for the Levantine dialect. The morphology rule illustrated the importance of using the stem rather than the full word form in ASR for rich morphologically languages tasks; this will give fewer

unique words and then the WER is likely to reduce.

Chapter 9 continued working on the stem in more details. It discussed the improving of Arabic dialects when working on LMs. We conducted experiments on three parallel dialects from Arabic multi dialect speech corpora; i.e. MSA, Gulf and Egyptian. For each dialect, two different LMs were produced: a closed domain LM and an open domain LM, to enable us to check the results against the two different LMs sizes.

The methodology of the second part of the multi dialect morphology analyser modified and then used to extract the three suggested forms of the word; stem alone, prefix+stem and stem+suffix. Six results were extracted per dialect, giving a total of eighteen results from Gulf dialect, Egyptian dialect as well as MSA. All of the experiments yielded positive results, between 0.5% to 6.8% in error rates. In three out of six cases, the best error rates were in the stem, while in the other three the best error rates were for stem+suffix. However, none of the prefix+stem obtained a best error rate.

Dealing with prefix+stem or stem+suffix rather than with prefix+stem+suffix will, in the case of similar results, ensure a larger part of the word is used, and not just the stem alone. By dealing with prefix+stem or stem+suffix rather than with prefix+stem+suffix we omitted one part of the word at any one time (i.e. prefix or suffix); therefore, we need to develop a method for turning prefix+stem or stem+suffix into the main form of the word after the recognition stage.

10.4 How this research can be extended to multi-dialect approaches to other languages

In this section, we show the methodologies of this thesis that can be carried out in other languages.

In the second and third parts of the methodology of the multi-dialect morphology analyser, which we introduced in Chapter 4, we used a special segmenter and a web-as-

corpus to retrieve the frequency of the segmented parts of the word. These two parts are language independent. They can be applied to any low resource language to build a morphology analyser, segmenter or stemmer, particularly languages that have a rich morphology or a rich combination of affixes. However, a list of combinations of prefixes and suffixes is required, which can be used to segment the words and then to check the parts through web queries to retrieve the frequency of each part. If the language does not have infixes, then the stem and the root can be distinguished and any affixes of the word can be extracted. However, the root cannot be extracted using this methodology for the languages that have infixes such as Arabic.

When there are large differences between a language and its dialects, or there are distinct words, that is, words are used in one region and are not used in others, the methodology suggested in Chapter 5 for building text corpora can then be used. There is a need to extract these distinct words, and then classify them into dialects before bootstrapping them to build a multi-dialect text corpus.

The methodologies that have been used for improving the PDs and LMs of Arabic dialects, which are described in Chapters 8 and 9, respectively, are also language independent. They can be applied to any language. However, the methodology that was used to improve PD can be applied if the language has clear orthography-to-phonology rules, which is the Arabic case. If there is a large combination of prefixes and suffixes, and the stem can be obtained by removing these affixes, the methodology of improving LM will be very useful. Any of the three forms can be the aim of the work: stem alone, prefix+stem or stem+suffix.

10.5 Future work

10.5.1 Extending the multi dialect analyser

There is still scope for further analysis in the second part of the analyser. In some cases there is more than one stem from the full word form. However, the segmenter in its current state extracts one form of the stem, and ignores the others. Therefore, it would be beneficial to add more rules to the segmenter to enable it to check other forms of the stem and extract all possible stem forms. We can take advantage of the web frequency to determine if the other stem forms that have been suggested by the segmenter are correct or not.

Another approach that would improve the analyser would be to extend the suggested approach to handle the diacritics correctly. One idea for creating a dialects diacritiser is to extend the work done by Alghamdi and Muzaffar (2007). The King Abdulaziz City for Science and Technology (KACST) diacritiser is a pure statistic diacritiser using quad-grams for dicritising MSA, and can be used for any Arabic texts if fully diacritised. The KACST diacritiser delivered encouraging results statistically without using any linguistic rules. We suggest using this technique for dialect texts, after building multi dialect text diacritised corpora.

To make the analyser more accurate there is a need to add more linguistic rules. These rules will be helpful when choosing the correct form of the stem. However, there is a need to distinguish those rules that are appropriate for dialects, and those that are appropriate for MSA words but not for dialect words as we tested these when improving PDs in Chapter 8.

10.5.2 Classifying text corpora

Many pages have mixed dialects, which causes noise when managing more than one dialect concurrently. Therefore, there is a need to undertake a dialect classification for the resulting text corpora. The classification can also be extended to include the cleaning of noise to produce a new version with less noise.

10.5.3 Producing different version of speech corpora

We are aiming in future to extend speech corpora by producing versions with different features. For example CTIMIT (Brown and George, 1995), which is cellular TIMIT, has been produced by transferring the speech database of TIMIT using the cellular network. By using a similar technique we can generate other version of our speech corpora; i.e. cellular parallel speech corpora.

10.5.4 Extending incremental methodology by applying it to other dialects

Future work will evaluate and extend the incremental methodology introduced in Chapter 8 to other Arabic dialects; the benefit of this work is to produce a list for each of the main dialects, showing how far each phonology rule affect the results positively or negatively. These new lists should also have the orthography to phonology rules that relate to connected words.

10.5.5 Returning to the original full word from prefix+stem or stem+stem after improving LMs

After obtaining improved LMs, another issue needs to be discussed in more depth. As can be seen from Chapter 9, there were improvements on the results in all experiments; i.e. prefix+stem or stem+suffix. However, results were typically better with stem+suffix

than with prefix+stem. The question raised was how best to return to the original full word form. One solution proposed, which needs to be evaluated in more detail, is that we can return to the original sentence by measuring the distance between output and original sentences in the original text. To achieve this, we need to measure any insertion between the recognised sentence and the original sentence. The minimum distance is then a pointer for the correct sentence.

10.6 Contributions

This chapter has presented proposals for future work and the conclusions to the thesis. Six parts were involved in this thesis. The summary of the contributions of this thesis is as follows:

1. We introduced a methodology for a multi dialect Arabic morphology analyser. This methodology synthesised linguistic and statistical bases. The linguistic method used an adapted MSA morphology analyser to handle the dialect affixes and then analyse the remaining word fragments. The second method involved segmenting the words and using 'the web as a corpus' to estimate the frequency of different segment combinations.

2. We presented the methodology followed when collecting Arabic dialects text corpora automatically by exploiting the web. The resultant corpus includes four sub-corpora for the four main Arabic dialects; i.e. Gulf, Egyptian, Levantine and North Africa.

3. We demonstrated the methodology utilised when collecting the Arabic parallel dialects speech corpora. The speech corpus has three dialects Gulf, Levantine, and Egyptian, as well as MSA.

4. We discussed different aspects of the building of a multi dialect speech recognition system for Arabic versus preparing a separate dialects speech recognition systems.

We discussed the results and the advantages and disadvantages affecting each system.

5. We introduced an incremental methodology for improving PDs in Arabic dialects. This methodology improved the PDs for MSA and Levantine dialect, we also used this methodology to distinguish how every letter-to-sound rule chosen affected the PD positively or negatively.

6. We introduced a morpheme-based methodology to improve the LMs in Arabic dialects speech recognition. Using this methodology, we evaluated three types of morphemes; i.e. stem alone, prefix + stem and stem + suffix. We also applied this methodology to different LMs sizes.

List of References

Abbas, M. and Berkani, D. (2006). Topic Identification by Statistical Methods for Arabic Language. *WSEAS Transactions on Computers*, 5(9):1908–1913.

Abbas, M., Smaili, K., and Berkani, D. (2011). Evaluation of Topic Identification Methods on Arabic Corpora. *Journal of Digital Information Management (JDIM)*, 9(5):185–192.

Abdel-Massih, E. T. (2011). *An Introduction to Egyptian Arabic*. MPublishing, University of Michigan Library, Michigan, United States.

Abdelali, A., Cowie, J., and Soliman, H. (2005). Building a modern standard Arabic corpus. In *Proceedings of the workshop on computational modeling of lexical acquisition.*, Split, Croatia. Citeseer.

Afify, M., Sarikaya, R., Kuo, H.-K. J., Besacier, L., and Gao, Y. (2006). On the Use of Morphological Analysis for Dialectal Arabic Speech Recognition. In *Proceedings of the Ninth International Conference on Spoken Language Processing (INTERSPEECH 2006)*, pages 277–280, Pittsburgh, Pennsylvania. IBM T.J. Watson Research Center.

Al-Sughaiyer, I. A. and Al-Kharashi, I. A. (2004). Arabic Morphological Analysis Techniques: A Comprehensive Survey. *Journal of the American Society for Information Science and Technology*, 55(3):189–213.

Al-Sulaiti, L. and Atwell, E. (2006). The design of a corpus of Contemporary Arabic. *International Journal of Corpus Linguistics*, 11(2):135–171.

Alarifi, A., Alghamdi, M., Zarour, M., Aloqail, B., Alraqibah, H., Alsadhan, K., and Alkwai, L. (2012). Estimating the size of Arabic indexed web content. *Scientific Research and Essays*, 7(28):2472–2483.

Alghamdi, M., Alhargan, F., Alkanhal, M., Alkhairy, A., Eldesouki, M., and Alenazi, A. (2008). Saudi Accented Arabic Voice Bank. *Journal of King Saud University - Computer and Information Sciences*, 20:43–58.

Alghamdi, M., Almuhtasib, H., and Elshafei, M. (2004). Arabic phonological rules. *King Saud University Journal: Computer Sciences and Information (in Arabic)*, 16:1–25.

Alghamdi, M. and Muzaffar, Z. (2007). KACST Arabic Diacritizer. In *Proceedings of the First International Symposium on Computers and Arabic Language (ISCAL07)*, pages 25–28, Riyadh, Saudi Arabia. King Abdulaziz City for Science and Technology (KACST).

Almeman, K. and Lee, M. (2012). Towards Developing a Multi-Dialect Morphological Analyser for Arabic. In *Proceedings of the Fourth International Conference on Arabic Language Processing (CITALA12)*, pages 19–25, Rabat, Morocco.

Almeman, K. and Lee, M. (2013a). A Comparison of Arabic Speech Recognition for Multi-Dialect vs. Specific Dialects. In *Proceedings of the Seventh International Conference on Speech Technology and Human-Computer Dialogue (SpeD 2013)*, Cluj-Napoca, Romania.

Almeman, K. and Lee, M. (2013b). An Incremental Methodology for Improving Pronunciation Dictionaries for Arabic Speech Recognition. In *Proceedings of the Seventh International Conference on Speech Technology and Human-Computer Dialogue (SpeD 2013)*, Cluj-Napoca, Romania.

Almeman, K. and Lee, M. (2013c). Automatic Building of Arabic Multi Dialect Text Corpora by Bootstrapping Dialect Words. In *Proceedings of the First International Conference on Communications, Signal Processing, and their Applications (ICCSPA13)*, pages 1–6, Sharjah, UAE.

Almeman, K. and Lee, M. (2013d). Automatic Multi-Dialect Analysis of Arabic. *Linguistica Communicatio: International journal of Arabic language engineering & General Linguistics*, 5:95–108.

Almeman, K. and Lee, M. (2013e). Building a Multi-Dialect Morphological Analyser for Arabic. *The International Journal of Computer Science and Engineering in Arabic (IJCSEA)*, 5(1):74–92.

Almeman, K., Lee, M., and Almiman, A. A. (2013). Multi Dialect Arabic Speech Parallel Corpora. In *Proceedings of the First International Conference on Communications, Signal Processing, and their Applications (ICCSPA13)*, pages 1–6, Sharjah, UAE.

Alorifi, F. S. (2008). *Automatic Identification of Arabic Dialects Using Hidden Markov Models*. PhD thesis, University of Pittsburgh, Pittsburgh, USA.

An-Nahar (2000). The An-Nahar Lebanon Newspaper Text Corpus. Technical report, ELRA. `http://catalog.elra.info/product_info.php?products_id=767` [accessed 11 March 2014].

Anumanchipalli, G. K., Oliveira, L. C., and Black, A. W. (2012). INTENT TRANSFER IN SPEECH-TO-SPEECH MACHINE TRANSLATION. In *Proceedings of the Spoken Language Technology Workshop (SLT)*, pages 153–158. IEEE.

Appen Pty Ltd (Gulf) (2006). Gulf Arabic Conversational Telephone Speech, Transcripts. Technical report, Linguistic Data Consortium (LDC), University of Pennsylvania, Philadelphia, USA. LDC Catalog No: LDC2006T15, http://catalog.ldc.upenn.edu/docs/LDC2006T15/0readme.txt [accessed 12 March 2014].

Appen Pty Ltd (Iraqi) (2006). Iraqi Arabic Conversational Telephone Speech, Transcripts. Technical report, Linguistic Data Consortium (LDC), University of Pennsylvania, Philadelphia, USA. LDC Catalog No: LDC2006T16, http://catalog.ldc.upenn.edu/docs/LDC2006T16/0readme.txt [accessed 12 March 2014].

Arslan, L. M. and Hansen, J. H. (1996). Language accent classification in american english. *Speech Communication*, 18(4):353–367.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1):5–22.

Beesley, K. R. (1998). Arabic Morphology Using Only Finite-State Operations. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages (Semitic '98)*, pages 50–57, Stroudsburg, PA, USA. Association for Computational Linguistics.

Beringer, N., Schiel, F., and Regel-Brietzmann, P. (1998). German Regional Variants-A Problem for Automatic Speech Recognition? In *Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP 1998)*, pages 85–88, Sydney, Australia.

Biadsy, F. (2011). *Automatic Dialect and Accent Recognition and its Application to Speech Recognition*. PhD thesis, Columbia University.

Biadsy, F., Habash, N., and Hirschberg, J. (2009). Improving the Arabic Pronunciation Dictionary for Phone and Word Recognition with Linguistically-Based Pronunciation Rules. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 397–405. Association for Computational Linguistics.

Biadsy, F., Moreno, P. J., and Jansche, M. (2012). Google's cross-dialect Arabic voice search. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*, pages 4441–4444. IEEE.

Billa, J., Ma, K. W., McDonough, J. W., Zavaliagkos, G., Miller, D. R., Ross, K. N., and El-Jaroudi, A. (1997). Multilingual Speech Recognition: The 1996 Byblos Callhome System. In *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*, pages 363–366, Rhodes, Greece.

Bokova, I. (2012). World Arabic Language Day. http://www.unesco.org/new/en/unesco/events/prizes-and-celebrations/celebrations/international-days/world-arabic-language-day/. [accessed 18 November 2013].

Boudelaa, S. and Marslen-Wilson, W. D. (2010). Aralex: a lexical database for Modern Standard Arabic. *Behavior Research Methods*, 42(2):481–487.

Boudlal, A., Lakhouaja, A., Azzeddine, M., and Abdelouafi, M. (2011). Alkhalil Morpho Sys1: A Morphosyntactic analysis System for Arabic texts. In *Proceedings of the International Arab Conference on Information Technology (ACIT'2010)*, pages 1–6, Riyadh, Saudi Arabia.

Brown, K. L. and George, E. B. (1995). CTIMIT: A speech corpus for the cellular environment with applications to automatic speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1995)*, volume 1, pages 105–108. IEEE.

Buckwalter, T. (2002). Buckwalter Arabic morphological analyzer version 1.0. Technical report, Linguistic Data Consortium (LDC), University of Pennsylvania, Philadelphia, USA. LDC Catalog No: LDC2002L49, http://catalog.ldc.upenn.edu/LDC2002L49 [accessed 13 March 2014].

Caballero, M., Moreno, A., and Nogueiras, A. (2009). Multidialectal Spanish acoustic modeling for speech recognition. *Speech Communication*, 51(3):217–229.

Canavan, A., Zipperlen, G., and Graff, D. (1997). CALLHOME Egyptian Arabic Speech. Technical report, Linguistic Data Consortium (LDC), University of Pennsylvania, Philadelphia, USA. LDC Catalog No: LDC97S45, http://catalog.ldc.upenn.edu/LDC97S45 [accessed 13 March 2014].

Carki, K., Geutner, P., and Schultz, T. (2000). Turkish LVCSR: Towards better Speech Recognition for Agglutinative Languages. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, volume 3, pages 1563–1566. IEEE.

Carter, R., McCarthy, M., Mark, G., and O'Keeffe, A. (2011). *English grammar today: An AZ of spoken and written grammar*. Cambridge University Press, The Edinburgh Building, Cambridge CB2 8RU, UK.

Chengalvarayan, R. (2001). Accent-Independent universal HMM-based speech recognizer for American, Australian and British English. In *Proceedings of the Seventh European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 2733–2736, Aalborg, Denmark. ISCA.

Chiang, D., Diab, M. T., Habash, N., Rambow, O., and Shareef, S. (2006). Parsing Arabic Dialects. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pages 369–376, Trento, Italy.

Chiang, D. and Rambow, O. (2006). The hidden tag model: synchronous grammars for parsing resource-poor languages. In *Proceedings of the Eighth International Workshop on Tree Adjoining Grammar and Related Formalisms*, pages 1–8. Association for Computational Linguistics.

CIA (2013). The World Factbook. `https://www.cia.gov/library/publications/the-world-factbook/`. [accessed 11 March 2014].

Clive, H. (2004). *Modern Arabic: Structures, Functions and Varieties*. Georgetown Classics in Arabic Languages and Linguistics series. Georgetown University Press, Washington, DC, USA, revised edition.

Cole, A., Graff, D., and Walker, K. (2001). Arabic Newswire Part 1 Corpus (1-58563-190-6). Technical report, Linguistic Data Consortium (LDC), University of Pennsylvania, Philadelphia, USA. LDC Catalog No: LDC2001T55, http://catalog.ldc.upenn.edu/LDC2001T55 [accessed 13 March 2014].

Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar, M., and Stolcke, A. (2007). Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing (TSLP)*, 5(1):1–29.

Creutz, M. and Lagus, K. (2005). Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Technical Report A81, Publications in Computer and Information Science, Helsinki University of Technology.

Darwish, K. (2002). Building a shallow Arabic Morphological Analyzer in one day. In *Proceedings of the workshop on Computational approaches to semitic languages (SEMITIC '02)*, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Diab, M., Ghoneim, M., and Habash, N. (2007). Arabic diacritization in the context of statistical machine translation. In *Proceedings of the Machine Translation Summit (MT-Summit)*, pages 143–149, Copenhagen, Denmark.

El-Desoky, A., Gollan, C., Rybach, D., Schluter, R., and Ney, H. (2009). Investigating the use of morphological decomposition and diacritization for improving Arabic LVCSR. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, pages 2679–2682, Brighton, United Kingdom.

Elmahdy, M., Gruhn, R., and Minker, W. (2012). *Novel techniques for dialectal arabic speech recognition*. Springer.

Elshafei, M., Al-Muhtaseb, H., and Alghamdi, M. (2006). Statistical methods for automatic diacritization of Arabic text. In *Proceedings of the Saudi Eighteenth National Computer Conference*, volume 18, pages 301–306, Riyadh, Saudi Arabia.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the LREC*, pages 2544–2547.

Ethnologue (17th ed., 2013). Arabic, Standard. `http://www.ethnologue.com/language/arb`. [accessed 11 March 2014].

Farghaly, A. and Shaalan, K. (2009). Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4):14:1–14:22.

Ferguson, C. A. (1971). *Language structure and language use: essays*, volume 1. Stanford University Press.

Finegan, E. (2008). *Language: Its structure and use*. Hedge series. Michael Rosenberg, fifth edition edition.

Gadalla, H., Kilany, H., Arram, H., Yacoub, A., El-Habashi, A., Shalaby, A., Karins, K., Rowson, E., MacIntyre, R., Kingsbury, P., Graff, D., and McLemore, C. (1997). CALLHOME Egyptian Arabic Transcripts. Technical report, Linguistic Data Consortium (LDC), University of Pennsylvania, Philadelphia, USA. LDC Catalog No: LDC97T19, http://catalog.ldc.upenn.edu/LDC97T19 [accessed 13 March 2014].

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (1993). TIMIT acoustic-phonetic continuous speech corpus. Technical Report 5, Linguistic Data Consortium (LDC), University of Pennsylvania, Philadelphia, PA, USA. LDC Catalog No: LDC93S1, http://catalog.ldc.upenn.edu/LDC93S1 [accessed 13 March 2014].

Ghani, R., Jones, R., and Mladenić, D. (2001). Mining the web to create minority language corpora. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 279–286, Atlanta, Georgia. ACM.

Goweder, A. and De Roeck, A. (2001). Assessment of Significant Arabic Corpus. In *Proceedings of the Arabic NLP workshop at ACL/EACL*, Toulouse, France. Citeseer, ELSNET.

Haak, M. (1996). *The Arabic Verb. A Functional Grammar approach to verbal expressions in Classical and Modern Arabic*. Ph.d. dissertation, University of Amsterdam.

Habash, N. (2010). *Introduction to Arabic natural language processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Habash, N., Eskander, R., and Hawwari, A. (2012). A morphological analyzer for egyptian arabic. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON2012)*, pages 1–9, Montreal, QC, Canada. Association for Computational Linguistics.

Habash, N. and Rambow, O. (2006). MAGEAD: a morphological analyzer and generator for the Arabic dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 681–688, Stroudsburg, PA, USA. Association for Computational Linguistics.

Habash, N., Rambow, O., and Roth, R. (2009). Mada+ tokan: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, pages 242–245, Cairo, Egypt. The MEDAR Consortium.

Habash, N., Soudi, A., and Buckwalter, T. (2007). *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, volume 38 of *Text, Speech and Language Technology*, chapter On Arabic Transliteration, pages 15–22. Springer.

Hammo, B., Abu-Salem, H., and Lytinen, S. (2002). Qarab: A question answering system to support the arabic language. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, pages 1–11. Association for Computational Linguistics.

Heintz, I. (2010). *Arabic Language Modeling with Stem-Derived Morphemes for Automatic Speech Recognition*. Ph.d. dissertation, The Ohio State University.

Hudson, G. (1986). Arabic root and pattern morphology without tiers. *Journal of Linguistics*, 22(1):85–122.

Ircing, P., Krbec, P., Hajic, J., Khudanpur, S., Jelinek, F., Psutka, J., and Byrne, W. (2001). On large vocabulary continuous speech recognition of highly inflectional language - Czech. In *Proceedings of the Seventh European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 487–490, Aalborg, Denmark. ISCA.

Jurafsky, D. and Martin, J. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Series in Artificial Intelligence. Pearson, second edition.

Kain, A., Hosom, J.-P., Ferguson, S. H., and Bush, B. (2011). Creating a speech corpus with semi-spontaneous, parallel conversational and clear speech Tech Report: CSLU-11-003. Technical report, Center for Spoken Language Understanding, Oregon Health & Science University.

Kilany, H., Gadalla, H., Arram, H., Yacoub, A., El-Habashi, A., and McLemore, C. (1997). Egyptian colloquial Arabic lexicon. Technical report, Linguistic Data Consortium (LDC), University of Pennsylvania, Philadelphia, USA. LDC Catalog No: LDC99L22, http://catalog.ldc.upenn.edu/LDC97S45 [accessed 13 March 2014].

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the Special Issue on the Web as Corpus. *Computational linguistics*, 29(3):333–347.

Kirchhoff, K., Bilmes, J., Das, S., Duta, N., Egan, M., Ji, G., He, F., Henderson, J., Liu, D., Noamany, M., Schone, P., Schwartz, R., and Vergyri, D. (2003). Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins Summer Workshop. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, volume 1, pages 344–347, Missouri, USA.

Kirchhoff, K. and Vergyri, D. (2005). Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition. *Speech Communication*, 46(1):37–51.

Kirchhoff, K., Vergyri, D., Bilmes, J., Duh, K., and Stolcke, A. (2006). Morphology-based language modeling for conversational Arabic speech recognition. *Computer Speech and Language*, 20(4):589–608.

LaRocca, C. S. A. and Chouairi, R. (2002). West point Arabic speech corpus. Technical report, Linguistic Data Consortium (LDC), University of Pennsylvania, Philadelphia, USA. LDC Catalog No: LDC2002S02, http://catalog.ldc.upenn.edu/LDC2002S02 [accessed 13 March 2014].

Lee, K. F., Hon, H. W., and Reddy, R. (1990). An overview of the SPHINX speech recognition system. *Acoustics, Speech and Signal Processing*, 38(1):35–45.

Lee, Y.-S., Papineni, K., Roukos, S., Emam, O., and Hassan, H. (2003). Language model based arabic word segmentation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL03)*, volume 1, pages 399–406, Stroudsburg, PA, USA. Association for Computational Linguistics.

Liu, Y. and Fung, P. (2006). Multi-accent chinese speech recognition. In *Proceedings of the Ninth International Conference on Spoken Language Processing (INTERSPEECH 2006)*, pages 133–136, Pittsburgh, Pennsylvania. ISCA.

Maamouri, M., Bies, A., Jin, H., and Buckwalter, T. (2003). Arabic treebank: Part 1 v 2.0. Technical report, Linguistic Data Consortium (LDC), University of Pennsylvania, Philadelphia, USA. LDC Catalog No: LDC2003T06, http://catalog.ldc.upenn.edu/LDC2003T06 [accessed 13 March 2014].

Maamouri, M., Buckwalter, T., Graff, D., and Jin, H. (2007). Fisher Levantine Arabic Conversational Telephone Speech, Transcripts. Technical report, Linguistic Data Consortium (LDC), University of Pennsylvania, Philadelphia, USA. LDC Catalog No: LDC2007S02, http://catalog.ldc.upenn.edu/LDC2007S02 [accessed 13 March 2014].

Maamouri, M., Graff, D., and Cieri, C. (2006). Arabic Broadcast News Speech. Technical report, Linguistic Data Consortium (LDC), University of Pennsylvania, Philadelphia, USA. LDC Catalog No: LDC2006S46, http://catalog.ldc.upenn.edu/LDC2006S46 [accessed 13 March 2014].

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. Massachusetts Institute of Technology (MIT) Press, Massachusetts, United States, sixth edition.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M. (1997). The det curve in assessment of detection task performance. In *Proceedings of the European Conference on Speech Communication and Technology*.

Mazzoni, D. and Dannenberg, R. (2006). Audacity [Computer software]. `http://audacity.sourceforge.net/`. [accessed 25 April 2013].

Microsoft (2011). Bing search API. `http://msdn.microsoft.com/en-us/library/dd900818.aspx`. [accessed 10 January 2013].

Nakanishi, A. (1980). *Writing systems of the world: alphabets, syllabaries, pictograms*. Tuttle Publishing, revised (9 jan 1998) edition.

Nguyen, L., Ng, T., Nguyen, K., Zbib, R., and Makhoul, J. (2009). Lexical and phonetic modeling for Arabic automatic speech recognition. In *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH 2009)*, pages 712–715, Brighton, United Kingdom.

Nikkhou, M. and Choukri, K. (2004). Survey on Industrial Needs for Language Resources. Technical report, NEMLAR Network for Euro-Mediterranean Language Resources. [accessed 13 March 2014].

Olive, J., Christianson, C., and McCary, J. (2011). *Handbook of natural language processing and machine translation*. Springer Publishing Company, Incorporated, first edition.

Parameswarappa, S., Narayana, V., and Bharathi, G. (2012). A novel approach to build Kannada web Corpus. In *Proceedings of the International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6, Hyderabad, India. IEEE.

Parker, R., Graff, D., Chen, K., Kong, J., and Maeda, K. (2011). Arabic Gigaword, Fifth Edition. Technical report, Linguistic Data Consortium (LDC), University of Pennsylvania, Philadelphia, USA. DC Catalog No: LDC2011T11, http://catalog.ldc.upenn.edu/LDC2011T11 [accessed 13 March 2014].

Pérez, A., Alcaide, J. M., and Torres, M. I. (2012). EuskoParl: a speech and text Spanish-Basque parallel corpus. In *Proceedings of the 13th International Conference on Spoken Language Processing (INTERSPEECH 2012)*, pages 2362–2365, Portland, Oregon, United States.

Rabiner, L. and Juang, B.-H. (1986). An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL)*, pages 527–534, Maryland, USA. Association for Computational Linguistics.

Riesa, J. and Yarowsky, D. (2006). Minimally supervised morphological segmentation with applications to machine translation. In *Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas (AMTA06)*, pages 185–192, Cambridge, Massachusetts, USA.

Rudnicky, A. (2007). The CMU pronunciation dictionary, release 0.7a. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`. [accessed 5 June 2013].

Ryding, K. C. (2005). *A reference grammar of modern standard Arabic*. Cambridge University Press, Cambridge, UK.

Salloum, W. and Habash, N. (2011). Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21, Edinburgh, Scotland, UK.

Sarikaya, R., Afify, M., and Gao, Y. (2007). Joint morphological-lexical language modeling (jmllm) for arabic. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, volume 4, pages 181–184, Honolulu, Hawaii, USA.

Soltau, H., Saon, G., Kingsbury, B., Kuo, J., Mangu, L., Povey, D., and Zweig, G. (2007). The IBM 2006 Gale Arabic ASR System. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, volume 4, pages 349–352, Honolulu, Hawaii, USA. IEEE.

Soudi, A., Neumann, G., and van den Bosch, A. (2007). *Arabic computational morphology: knowledge-based and empirical methods*, volume 38 of *Text, Speech and Language Technology*. Springer.

Sphinx (2009). Sphinx 3.0.8 [software]. `http://sourceforge.net/projects/cmusphinx/files/sphinx3/0.8/`. [accessed 11 March 2014].

Sphinxtrain (2011). Sphinxtrain 1.0.7 [software]. `http://sourceforge.net/projects/cmusphinx/files/sphinxtrain/1.0.7/`. [accessed 11 March 2014].

Staal, J. F. (1967). *Word order in Sanskrit and universal grammar*, volume 5 of *Foundations of Language Supplementary Series*. Springer, 1st edition.

Tawileh, A. and Alghamedi, M. (2011). A Corpus Linguistics-based Approach for Estimating Arabic Online Content. In *Proceedings of the Conference on Human Language Technology for Development (HLTD 2011)*, pages 02–05, Alexandria, Egypt.

Torres-Carrasquillo, P. A., Gleason, T. P., and Reynolds, D. A. (2004). Dialect identification using Gaussian mixture models. In *Proceedings of the Speaker and Language Recognition Workshop (ODYSSEY04)*, pages 297–300, Toledo, Spain.

Vergyri, D., Mandal, A., Wang, W., Stolcke, A., Zheng, J., Graciarena, M., Rybach, D., Gollan, C., Schlüter, R., Kirchhoff, K., A., F., and N., M. (2008). Development of the SRI/Nightingale Arabic ASR system. In *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH 2008)*, pages 1437–1440, Brisbane, Australia.

Versteegh, K. (2001). *The Arabic Language (Islamic Surveys)*. Edinburgh University Press, Edinburgh, UK.

Watson, J. C. (2007). *The phonology and morphology of Arabic*. Phonology of the World's Languages. Oxford University Press.

WCAG2.0 (2008). Web content accessibility guidelines (wcag) 2.0. `http://www.w3.org/TR/WCAG20/`.

Xiang, B., Nguyen, K., Nguyen, L., Schwartz, R., and Makhoul, J. (2006). Morphological decomposition for Arabic broadcast news transcription. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, volume 1, pages 1089–1092, Toulouse, France. IEEE.

Yeti (2011). Blue Yeti Microphone, user manual. Technical report, Blue Yeti, USA. `http://cdn.bluemic.com/pdf/yetipro/YetiPro_manual_English.pdf` [accessed 5 April 2013].

Young, S. J. and Woodland, P. (1994). State clustering in hidden Markov model-based continuous speech recognition. *Computer Speech & Language*, 8(4):369–383.

Zemanek, P. (2001). CLARA (Corpus Linguae Arabicae): An Overview. In *Proceedings of the ACL/EACL Workshop on Arabic Language*, pages 111–112, Toulouse, France.

Zipf, G. K. (2012). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Martino Fine Books, reprint of 1949 edition.