Faculty of Engineering and Technology

Master in Computing

“Towards Building a Corpus for Palestinian Dialect”

”نحو بناء مدونة للعامية الفلسطينية“

By: Diyam Fuad Akra

Supervisor: Dr. Mustafa Jarrar

Committee:

Dr. Nizar Habash

Dr. Mahdi Arar

Dr. Abualseoud Hanani

This Thesis was submitted in partial fulfillment of the requirements for the Master’s Degree in Computing from the Faculty of Graduate Studies at Birzeit University, Palestine

28-01-2015

Faculty of Engineering and Technology

Master in Computing

"Towards Building a Corpus for Palestinian Dialect"

"نحو بناء مدونة للعامية الفلسطينية"

By: Diyam Fuad Akra

Committee:	**Signature**
Dr. Mustafa Jarrar (Chairman)	……..
Dr. Nizar Habash (Member)	……...
Dr. Mahdi Arar (Member)	……..
Dr. Abualseoud Hanani (Member)	……..

الاهداء

إليكَ يا نور الحياة، لولاك ما كنت وما كان هذا الانجاز، معك أمشي بخطوات ثابتة وبثقة تطاول عنان السماء، إليك أبي الحبيب

إليكِ يا سر البسمة وفرحة العمر، يا من صنعتني على عينك فدونك أنا لا شيء، إليكِ أمي الحبيبة.

إلى شريك العمر، إلى أيامي القادمة ، إليكَ شادي.

Acknowledgment

Special thanks to my supervisor Dr. Mustafa Jarrar for his great support and valuable discussions and feedback.

I would like to also thank Dr. Nizar Habash for his valuable ideas during this research, and the other committee members Dr. Mahdi Arar and Dr. Abualseoud Hanani for their valuable comments.

Special thanks to my colleagues at Sina Institute, specially Faeq Alrimawi and Nasser Zalmout, for their input, support, and feedback.

Many special thanks to Faisal Alshargi and Owen Rambow and Ramy Eskander, from Columbia University, for their cooperation and for providing me with tools to support this research.

Special thanks to my sisters and my brothers, for helping in many ways.

This research is part of the Cuuras project, at Sina Institute, which is funded by the Ministry of Higher Education – Scientific Research Council.

Abstract

The number of users of the internet is increasing exponentially every year; most of these users are using social networks, blogs and forums. When users in the Arab world need communicate with each other, they often use their colloquial dialect Arabic instead of the Modern Standard Arabic (MSA). To retrieve information about a specific topic, Modern Standard Arabic (MSA) Informational Retrieval will retrieve only MSA relevant data, but maybe there is helpful information published in Dialect Arabic, also if we need to directly translate from Dialect to English, this may be done by translating the Dialect to standard Arabic then to English.

This thesis aims to build an annotated corpus for Palestinian Dialect with Relevant Meta Data. The proposed methodology includes studying linguistic facts about Palestinian dialect and comparing it with Modern Standard Arabic in terms of morphology, orthography and lexical. As well as collecting Palestinian written text from different resources, then analyzing and annotating the corpus by using resources designed for

Egyptian dialect, after that annotating manually a list of words that can't be analyzed by existing resources, finally start using the existing annotation tool to double check over annotated corpus.

ملخص

يتزايد استخدام الانترنت ووسائل التواصل الاجتماعي بشكل ملحوظ يوما بعد يوم،

وكما هو معروف فإن الناس عادة تستخدم العاميات في حياتها اليومية بدلا من اللغة الفصحى، وبالضرورة فإن هذه العادة انتقلت للعالم الافتراضي و وسائل التواصل الاجتماعي، ومن هنا يتبادر إلى أذهاننا التساؤل التالي: ماذا لو أردت أن ابحث عن كلمة أو معلومة معينة عبر الانترنت، هل أستطيع أن استرجع ما نشر حول هذه الكلمة أو المعلومة بالعامية أيضا، وماذا لو خطر ببالنا أننا نود الترجمة من العامية الى الانجليزية مباشرة مثلا.

من هذه التساؤلات برزت الفكرة المقدمة في هذه الرسالة ألا وهي بناء مدونة للعامية الفلسطينية تتضمن العديد من الكلمات الفلسطينية موسومة بمجموعة من الصفات النحوية

والمعجمية مثل أصل الكلمة والسوابق واللواحق وتصنيف الكلمة(فعل،اسم...الخ) ومعناها ومقابلها في اللغة العربية الفصحى.

تمت عملية بناء المدونة من خلال عدة مراحل وهي : دراسة الفروقات النحوية والشكلية والمعجمية بين العامية الفلسطينية واللغة العربية الفصحى ومن ثم جمع مادة فلسطينية مكتوبة من مصادر مختلفة، ثم استخدام الوسائل التي صممت للعامية المصرية لتحليل النص الفلسطيني، وأخيرا توسيم ما لم يتم تحليله من قبل الوسائل المصرية يدويا واستخدام برنامج صمم لتسهيل عملية التوسيم لتدقيق ما تم تحليله من قبل الوسائل المصرية.

Table of Contents

الاهداء I

Acknowledgment II

Abstract III

ملخص V

Introduction 1

1.1 Scope and Motivation 1

1.2 Problem Statement and Thesis Goals 9

1.3 Summary of Contributions 9

1.4 Summary of Structural Work 10

Literature Review 12

2.1 Corpus Collection and Annotation 12

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

2.1.2 Arabic Dialects Parser 14

2.1.3 Cross Lingual Arabic Blog Alerts (COLABA) 18

2.2 Dialectal Orthography 25

2.3 Dialectal Morphological Annotation 31

2.3.1 MADAMIRA 32

2.4 Remarks on the related works and literature 38

Corpus Collection 40

3.1 Corpus 40

3.1.1 Corpus collection methodology 40

3.1.2 Corpus Statistics 41

3.2 Processing and Storing collected text 50

3.2.1 Storing methodology 50

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

List of Figures

Figure 2.1: Morphological/Part-of-speech/gloss (MPG) tagging [2]……..13

Figure 2.2: "انا مش عارفة" sentence Treebank [2]……..13

Figure 3.3: "رايح اشترك فيه" sentence Treebank [2] 14

Figure 2.4:Sample of Levantine-MSA dictionary [1] 15

Figure 2.5: tree pair from MSA-dialect synchronous grammar [1][3].....18

Figure 2.6: Sample of Dialect Arabic queries [5] 19

Figure 2.7: Typo identification and Fixing interface [4].....21

Figure 2.8: Specifying word dialect interface [4] 22

Figure 2.9: Lemma creation interface [4].....23

Figure 2.10: Morphological Profile creation interface [4] 23

Figure 2.11: EGY CODA Example[8] 30

Figure 2.12: TUN CODA Example [9] 30

Figure 2.13: MADA output sample[15]……..33

Figure 2.14: MADAMIRA architecture[16].....36

Figure 2.15: MADAMIRA Demo online interface[16] 37

Figure 3.1: Sample from Facebook Document…….…….....42

Figure 3.2: Sample from Twitter Document…….....43

Figure 3.3: Sample from "عبد الحميد عبد العاطي" blog…….....44

Figure 3.4: Sample from Palestinian Stories…….....45

Figure 3.5 : Sample from Forum …….....46

Figure 3.6: Sample from Palestinian terms…….....46

Figure 3.7 : one of "وطن ع وتر" episodes…….....47

Figure 3.8 : Sample from Dictionary……..…….....48

Figure 3.9 : Sample 2 from Dictionary…….....49

Figure 3.10 : Sample 3 from Dictionary……..…..…49

Figure 3.11: Database ER model.…….……51

Figure 3.12: Document Table……..…….…….....52

Figure 3.13: Palestinian Dialect words Table…….…53

Figure 3.14: Bigram Palestinian Dialect words Table……..…….……54

Figure 3.15: Trigram Palestinian Dialect words Table……..….54

Figure 3.16: Fourgram Palestinian Dialect words Table…….55

Figure 4.1: Sample of MADAMIRA_EGY Result……..68

Figure 4.2: Sample of Manual Annotated List……...……..68

Figure 4.3: DIWAN Interface……...…….…….69

Figure 4.4: DIWAN Output Sample……...70

Figure 4.5 : Annotators differences Sample ……74

List of Tables

Table 3.1: Data base tables' statistics 55

Table 4.1: Sample of Exceptional Palestinian words 60

Table 4.2: Annotation Specification Example 63

Table 4.3:MADAMIRA EGY Result 65

Table 4.4: MADAMIRA MSA Result 65

Table 4.5: experiment result 65

Table 4.6: inter-annotator agreement result 73

Chapter One

Introduction

This chapter gives a quick introduction to the thesis. Section 1.1 presents the scope and the motivation of our research. The problem statement and the thesis goals are presented in section 1.2. Section 1.3 summarizes our main contributions, and section 1.4 gives a structural overview of the thesis.

1.1 Scope and Motivation

The need for processing the Arabic language texts recently became an important issue due to the necessity for many application types, such as

search engines, spelling checker, morphological analyses, machine translation, among others.

The Arabic language is spoken by 300 million people all over the world [4]. Arabic language has two forms: Modern Standard Arabic (MSA) and Dialectal Arabic; MSA is the superordinate of language which is used particularly -the in media, education, and formal communications. MSA is not used in daily life expressions; it is also more written rather than spoken. On the other hand, Dialectal Arabic is an informal colloquial, which is used mainly in daily communications among Arabs. Besides that, it is a spoken language rather than written language. So the term dialect is used to describe differences in speech that are associated with different regions or different social groups, as you wander around, you can find variations in speech that are associated with their place of residence (urban or rural). In general, Dialectal Arabic is classified into five main categories [4]: *Egyptian Dialect, which* is used in Egypt and Sudan, *Gulf Arabic, which* is used in Gulf countries, *Moroccan* Arabic, which is used in Western Arab countries (Morocco, Libya, Tunisia, Algerian), *Iraqi Arabic, which* is used in Iraq, and ultimately *Levantine*

Arabic which is used in Levant Countries (Lebanon, Syria, Jordan, and Palestine).

Dialectal Arabic is found as a verbal language among people only, and recently in these days Arab peoples start using their dialectal Arabic in web area. The emergence of social networks, blogs and forums, users tend to use the Arabic dialect language in their communications, therefore, the need for processing the Arabic dialect is getting urgently more important day by day.

The need to deal with these Arabic Dialectal data is becoming significantly important for the upcoming uses. However, using Natural Language Processing (NLP) tools which are basically designed for Standard Arabic to deal with Dialectal Arabic will not solve the problem, but it will lead to lower performance due to the fact that [4]:

- Standard Arabic has different morphological, lexical, orthographical, and phonological aspects rather than dialectal Arabic;

- Dialectal Arabic is less controlled; because it is used by people in daily life, which means that there are no standard orthographic rules for it.
- Lack of annotated dialectal Arabic corpus and resources.

Our Next Step is to explore some differences between MSA and dialects.

Arabic Variants

The forms of Standard Arabic and its dialects are called *Arabic Variants* and classified in [6, 10, 11, 12, and 13] as follows:

Phonological level variants, which is the sound of character; Arabic people utter same character in different sounds. For example: In Standard Arabic, "ق" is varied between Arabic dialects such as in Levantine is replaced by "أ", "ك", "چ"; in Egyptian is replaced by "أ", "چ" in Gulf is replaced by "چ." This variant is coming over phonological level. Another phonological level is "ث" which becomes as "س","ت" in the Levantine and Egyptian; "ذ" which has variants "ز","د"; "ج" which has variants "ي","چ".

Orthographic level variant, which is the way of writing/spelling a character. Arabic people write the same character in different spellings. Same phonological level variants also occur in orthographic level because each phoneme may be written in different forms as it is spoken. Another issue in orthographic level is lack of standard of orthographic, so as an example the sentence "لم أقل لك" in standard Arabic may be written in Levantine Arabic as follows: "ما بقلكش", "م بقلكش", "مبقلكش", "ما بألكش", "م بألكش", "مبألكش", "ما بكلكش", "م بكلكش", "مبكلكش". Also, sometimes people write the same word in different forms, as an example "الضو", "الضوء".

Lexicon variants, which meaning that lexical term between the variants is completely different, examples of lexical variants: The word "فقط" in Standard Arabic is replaced by the word "بس" in Levantine Arabic. The word "اريد" in Standard Arabic is replaced by the word "بدي" in Levantine Arabic, "عاوز" in Egyptian Arabic, "ابي" in Gulf Arabic, "بغيت" in Moroccan Arabic, and "اريد" in Iraqi Arabic. Reference names, as "هؤلاء" becomes "هدول" in Levantine Arabic.

Morphological level variants, which meaning that there is a difference in the morphology of the same word, examples of morphological variants: present progressive is different between Arabic variants, as an example the verb "يلعب" in Standard Arabic is replaced by " بيلعب" in the Levantine and Egyptian Arabic, and "ديلعب" In Iraqi Arabic, and "كيلعب" in Moroccan Arabic. Future also differs, as an example "سيلعب" in Standard Arabic became "حيلعب" in Levantine, "هيلعب" in Egyptian, "رح يلعب" in Iraqi, and "غيلعب" in Moroccan.. Furthermore, pronouns are different among Arabic variants, such as "كم" in MSA may be replaced by "كوا" In Levantine.

Sentence structure level variants, which means that the structure for same sentence differs, examples of syntactic variants: "شركة محمد" in MSA sometimes becomes "الشركة تبعت محمد" in Levantine; word order, as "هذا الولد" becomes "الولد دا" as in Egyptian Dialect. MSA Negations forms such as "لم" which may replaced by "ما" before a verb and "ش" at the end of the verb; Example: "لم أقل لك" in MSA equivalent to "ما قلتلكش" in Levantine.

So, because of these Arabic variants using existing MSA NLP tools for Arabic dialects may lead low accuracy results [1,14]. This becomes a problematic issue and there is a need for tools that have the ability to deal with dialectal Arabic. So from here comes the importance of building an annotated corpus of Palestinian dialect, which we are going to work on this thesis.

The importance of building an annotated corpus considers as the foundation stone of many applications, some of them are as follows:

- *Information Retrieval and Extraction applications*, there are many applications, which are responsible for retrieving or extracting information about something. Building annotated corpus will help these applications to go through dialectal Arabic data not only over Standard Arabic.
- *Search engines*, through which users can search the required things using dialectal Arabic words, and using an annotated corpus to get its equivalent from Standard Arabic.

- *Machine Translation*, by converting Dialectal Arabic words to its equivalent or synonym Standard Arabic then translate them to the corresponding languages.
- *AutoComplete* with possible equivalent Standard Arabic word; In other words, if a user starts entering some Dialectal Arabic word, then -an application starts giving him/her possible Standard Arabic word
- *Part-Of-speech* tagging applications.
- Dialectal Arabic Parsers.

According to the variants for Arabic, which are discussed above, there is a need to deal with each Arabic language dialect separately, for this reason, this thesis will focus on the Palestinian dialect, Palestinian dialect which is considered a sub-dialect of Levantine Arabic.

1.2 Problem Statement and Thesis Goals

This thesis aims to *collect a corpus of written Palestinian dialect and annotate it with relevant metadata*. More specifically, we did this by:

- Collecting Palestinian dialect texts manually from different resources such as blogs, forums, social networks, dialect dictionaries, and TV series.
- Annotating each word in the corpus manually using the DIWAN tool, which is linked with the Egyptian MADAMIRA [15, 16]. This also includes writing each word in the CODA (Conventional Orthography for Dialectal Arabic) standard [8, 17].

1.3 Summary of Contributions

Our contributions can be summarized as follows:

- A corpus of Palestinian dialect text (43K words, 15K of them unique) was manually collected from different resources, like blogs, forums, Facebook, Twitter, books and Palestinian series.

- The corpus was parsed automatically and stored in a proper database schema.
- A list of 1234 unique corpus words is annotated manually.
- 18743 words (with 500 unique words) from this corpus were selected and fully annotated using the DIWAN tool. The POS of each word was specified, as well as it was written according to the CODA rules [8,17]. The rest of the corpus will be future work.
- It is worth noting that this work was done in cooperation with researchers from the Curras project, and a joint paper [20] was published about it (see appendix 1), as well as a technical report [17]. The achievements listed above are our own contributions, but following the project's methodology, tools, and resources.

1.4 Summary of Structural Work

Chapter Two presents the literature review of related work done in this area. A special focus given to work done on conventional orthography for dialectal Arabic (CODA) [8,9]. The work was done on developing morphological analyzer for Arabic and its dialects (MADAMIRA)

[15,16,18,19]. The work was done on developing web application for Dialectal Arabic Text annotation [5]. Finally The work was done on Parsing Arabic Dialects [1,2,3,4].

Chapter Three described our corpus. It will cover our methodology that we followed to collect our corpus, it covers corpus statistics such as type of documents, number of documents, number of threads in each document, and number of words in each document.

Chapter Four presented approaches, methods and tools used in the annotation process. It will cover annotation concept, annotation methodology which includes a list of meta data that each word will be annotated with it. The benefits of using MADAMIRA in our annotation process, a sample of our manual annotation list, and finally it covers how DIWAN tool will speed up the annotation process.

Chapter Five This chapter discussed the results we -have gotten until now, and it also presented the future work that can be done in this area. And finally covered using of DIWAN tool to complete our annotated corpus.

Chapter 2

Literature Review

This chapter gives a quick review over works done on Arabic Dialects. Section 2.1, presents work done on corpus collection and annotation. Section 2.2, presents work done on dialectal orthography. Section 2.3, presents work done on dialectal morphological annotation.

2.1 Corpus Collection and Annotation

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

Levantine Arabic Treebank (LATB) contains 33,000 words and it is built for development and testing purposes. LATB is built using Jordanian Levantine Conversational Telephone speech.

Developing Levantine Arabic Treebank approach begins by analyzing corpus data morphologically; so for each word in corpus generates manually a tag called Morphological/Part-of-speech/gloss (MPG) tag. Figure 2.1 shows an example of this tagging.

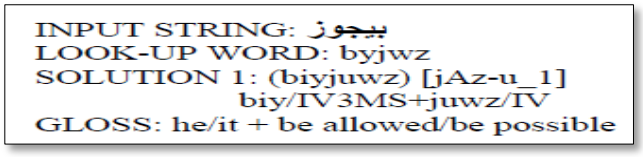


Figure 2.1: Morphological/Part-of-speech/gloss (MPG) tagging [2]

Then analyzing corpus data syntactically; which is done manually and raises many issues such as detecting conversational speech effects and disfluencies, active Participles which may be adjectival or verbal, according to context; Figures 2.2, and 2.3 show examples when it is adjectival, or verbal

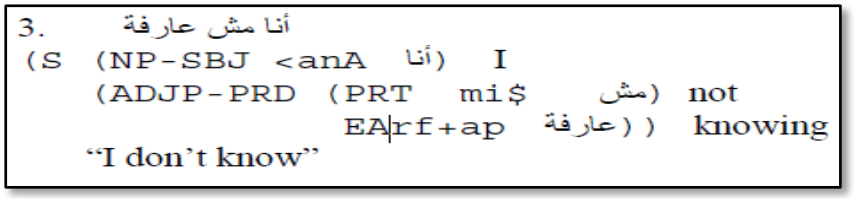


Figure 2.2: "انا مش عارفة" sentence Treebank [2]

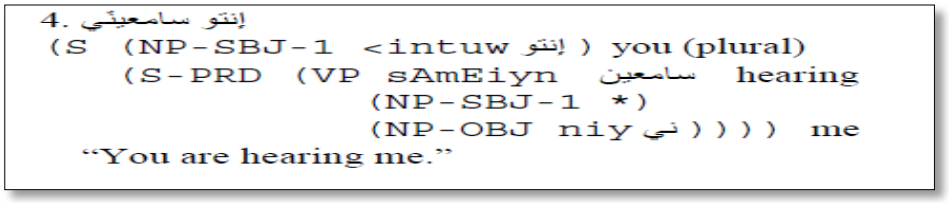


Figure 2.3: "رايح اشترك فيه" sentence Treebank [2]

After developing Levantine Treebank, the authors apply automatic tree tool over it which are: Cat-tree which is responsible for detecting inconsistency and unwell-formed sentences and it separates multiple trees, Clean-trees which are responsible for removing resulting null sentences (unfinished sentences) and null constituents, Tregex and Tsurgeon which are responsible for traversing constituents and either transform it or remove it. Applying these tools reduced Levantine Arabic Treebank from 6639 trees to 3979 trees.

2.1.2 Arabic Dialects Parser

In [1] and [3] authors' presented three approaches to building a parser for Arabic Dialects, but before starting work on these approaches, the authors need to prepare linguistic resources that will be used in these approaches. The first linguistic resource is Levantine-MSA dictionary, this dictionary will be used to translate Levantine sentence to Standard

Arabic sentence and to convert Standard Arabic Treebank to Levantine Treebank. Each dictionary entry contains Levantine word with its Part-of-Speech, equivalent MSA word and English gloss as shown in Figure 2.4. The dictionary is relatively small. It contains 2201 words.

Levantine		POS	MSA		English
إيه	<yh	UH	نعم	nEm	yes
إنتو	<ntwA	PRP	أنتم	<ntm	you (pl.)
+كي	+ky	PRP	+ك	+k	her
كمان	kmAn	RB	ايضاً	AyDAF	also
كمان	kmAn	RB	كذلك	k*lk	also
اللي	Ally	WP	الذي	Al*y	who
اللي	Ally	WP	التي	Alty	who
شو	$w	WP	ماذا	mA*A	what
كيف	Kyf	WRB	كيف	kyf	how
شلون	$lwn	WRB	كيف	kyf	how
بحكي	bHky	VBP	أتكلم	>tklm	I speak
أحكي	>Hky	VBP	أتكلم	>tklm	that I speak
منحكي	mnHky	VBP	نتكلم	Ntklm	we speak
حكيت	Hkyt	VBD	تكلمت	Tklmt	I spoke
العيلة	AlEylp	NN	العائلة	AlEA}lp	the family
عيلة	Eylp	NN	عائلة	EA}lp	Family

Figure 2.4: Sample of Levantine-MSA dictionary [1]

The Second linguistic resource is part-of-speech tagging for corpus data, and the basic idea here is to assume Standard Arabic (MSA) tagger as a baseline then make adoption on it. So the first step is to run the MSA tagger over Levantine data, the accuracy of applying MSA tagger over Levantine was 69%. According to accuracy result a number of adaption making on MSA tagger were added, which are: Adoption using probabilities and normalization after applying this step the accuracy of

MSA tagger became 73%, replace Levantine words that have an entry in the dictionary presented previously with MSA word, and run the MSA tagger another time then the accuracy became 80%, Manually tagged part of Levantine corpus, and used the tagged set as training set, the accuracy after applying this step reached to 80%.

After preparing linguistic resources, the authors started applying parsing approaches. The first approach is sentence transduction which depends on translating Levantine sentence to MSA sentence, then parsed the MSA translated sentence using MSA parser. This approach has many disadvantages such as lack of resources for translation from Levantine Arabic to MSA, and sometimes two words in Levantine Arabic are Translated to the same word in MSA مثل مِن-مين تترجم إلى مِن و مَن"نفس الكلمة". Another approach is Treebank Transduction which depends on converting Standard Arabic Treebank into Levantine-Like Treebank using linguistic knowledge, then train a statistical parser on converted Treebank, then test parsing performance over new Levantine corpus data. There are a lot of transformations that were done over Standard Arabic Treebank to convert it to Levantine-like Treebank such as unifying two

blocks with the same meaning in one and separating nested node, replacing all negations in Arabic Treebank with Levantine negation forms (“مش\ما” precedes verb, “ش” after verb), replacing (Verb-Subject-Object) order in Standard Arabic Treebank with Levantine order which is (Subject-Verb-Object), replacing (demonstrative pronoun – noun) order in Standard Arabic Treebank with Levantine order which is (noun-demonstrative pronoun), replacing every lexical in Arabic Treebank with its equivalent in Levantine using dictionary that described above, replacing every “احتاج\اريد” verb and all derivatives in Arabic Treebank with “بد” and its derivatives, adding prefix “ب” to every verb that has “VBP” tag, Converting “ليس” in Arabic Treebank to particle.

Final approach used in parsing Arabic dialects is Grammar transduction which depends on developing an MSA-dialect synchronous grammar. This grammar contains pairs of elementary trees which combine MSA elementary tree with corresponding Levantine elementary tree. Figure 2.5 shows a tree pair from this grammar.

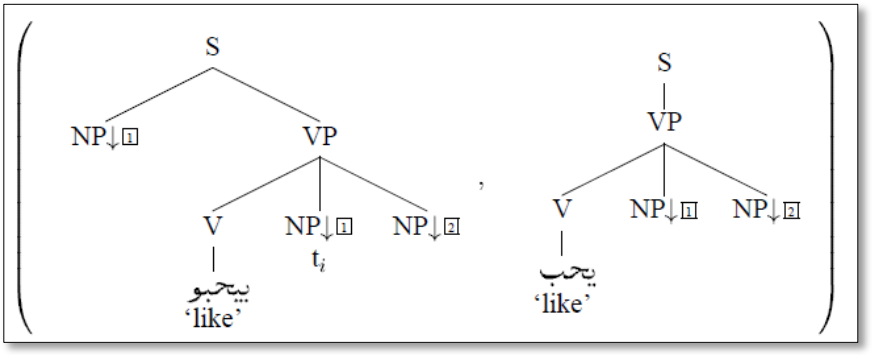


Figure 2.5: Tree pair from MSA-dialect synchronous grammar [1] [3]

This grammar developed by extracting Standard Arabic elementary trees from Arabic Treebank then translates it using handwritten rules to corresponding Levantine elementary trees.

2.1.3 Cross Lingual Arabic Blog Alerts (COLABA)

In this section, we present COLABA (Cross Lingual Arabic Blog Alerts) project, which is done by Diab, Habash, Rambow, Altantawy, and Benajiba in [5], and aims to processing Arabic social data on the web, through COLABA project many tools were developed in order to achieve COLABA goal which is to retrieve all dialect Arabic Blogs data and all MSA blogs data that are related to the required MSA query.

COLABA goal is achieved by applying many stages; In the first stage, COLABA project asked 25 annotators to design a dialect query that is

responsible for harvesting large amount of dialectal data from the web, annotators designed 40 queries with its dialect, corresponding MSA translation, and English translation. Figure 2.6 shows a sample of these dialect queries.

DA Query	DA	MSA	English
الطلاق بقى ظاهره	EGY	الطلاق اصبح ظاهرة	divorce became very common
راح احجيلكم	IRQ	سوف اروي لكم	I will tell you a story
راح دوز ع قبر بيه	LEV	ذهب فورا الى قبر ابيه	He went directly to visit his father's tomb
ما زال شاد فراسو	MOR	لا زال في وضع جيد	he is still in good shape

Figure 2.6: Sample of Dialect Arabic queries [5]

Then COLABA Project goes through annotation stage, which is responsible for removing HTML markup, spam, advertisements, encoding issues, and every Meta data from blogs data that's retrieved in the first stage.

After that, COLABA project gives each blog initial rank according to its degree of dialectness. This was done through a simple module called Dialect Identification pipeline which is responsible for determining the degree to which a text includes Dialect Arabic words. It works as follows: it takes an input text, and then analyzes each word in the input text by Buckwalter MSA morphological analyzer (BAMA), if BAMA

returns a result; then a word is MSA. Otherwise, it is potentially a dialect word, and then it gives rank to a blog according to the number of words that aren't classified by BAMA.

Then COLABA project Applies typographical clean up over blogs with high ranking. This is done by correcting every MSA word with non-standard orthography by writing it in standard orthography, remove speech effects from words, add missing spaces between words, and then applies Dialect Identification pipeline another time.

Then COLABA project applies COLABA Conventional Orthography (CCO), which is responsible for providing orthography for each word in Dialects, as an example the word "باب" has Levantine orthography which is "be:b", and Egyptian orthography which is "ba:b".

After that, COLABA project applies Dialect annotation over highly ranked blogs, this step is done using a COLANN_GUI web application which was presented by Benajiba and Diab in [4], and it is a web application used by annotators to annotate text. It is browser independent, it uses PHP to interact with a server database and JavaScript for GUI, it contains three types of users: Annotator which is responsible for

annotation, Lead Annotator which is responsible for assign annotation tasks to annotators, Super User Which is the administrator of the system.

Annotation process using COLANN_GUI passes through many steps. The first step, if there is a misspelling in a word, an annotator must correct it, and if there is also a speech effect, annotators must remove it as well remove the missing spaces, then the annotator must do sentence boundary detection. Figure 2.7 shows the interface that annotators used in this step.

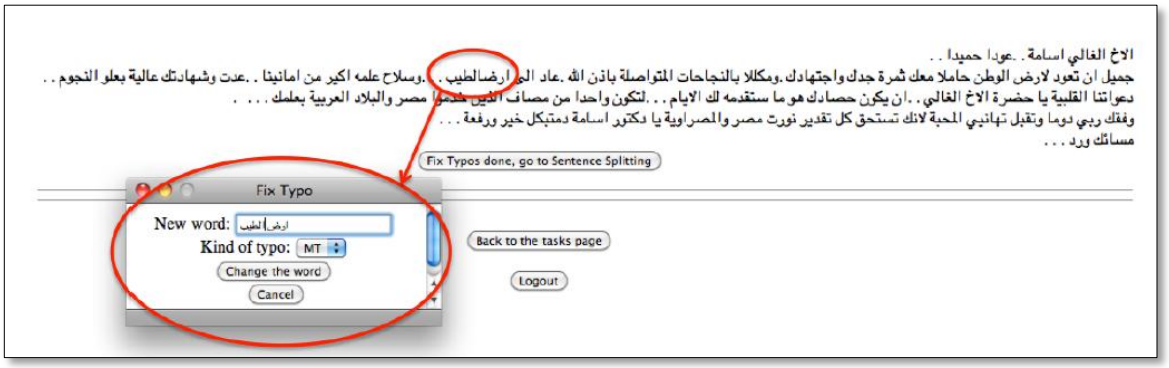


Figure 2.7: Typo identification and Fixing interface [4]

Second, the annotator is asked to choose the dialect of a word, level of dialectalness, and enters phonetic transcription of the word. Figure 2.8 shows the interface that annotators used it in this step.

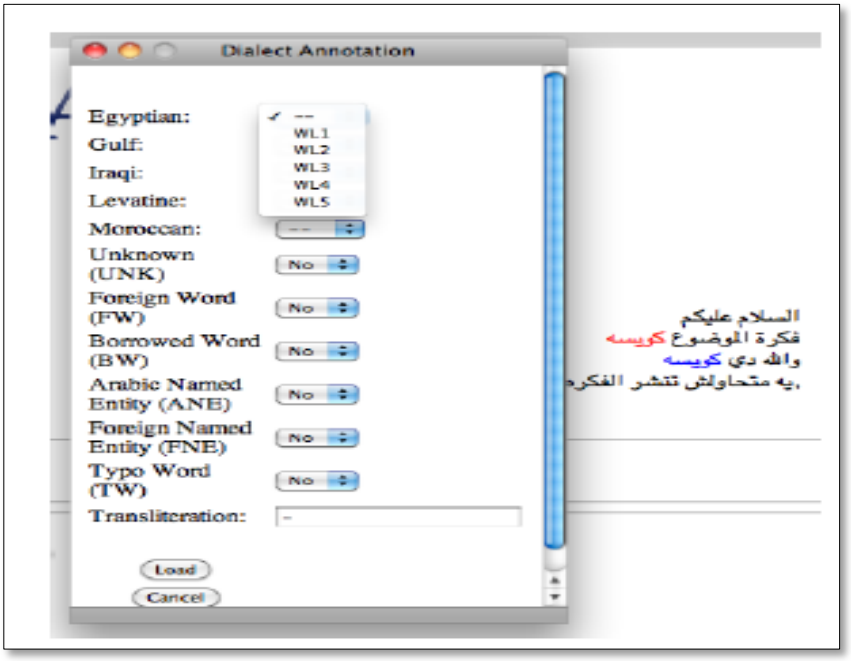


Figure 2.8: Specifying word dialect interface [4]

Third; annotators must enter the underlying lemma form of each word (derived from), in this step the application provides an annotator with a Dialect Arabic word and a list of usage examples from blogs data, then asks an annotator to provide a corresponding lemma, MSA equivalent, English equivalent, dialect Id for each word in usage examples, and associate each lemma with its usage example. Figure 2.9 shows the interface that annotators used in this step.

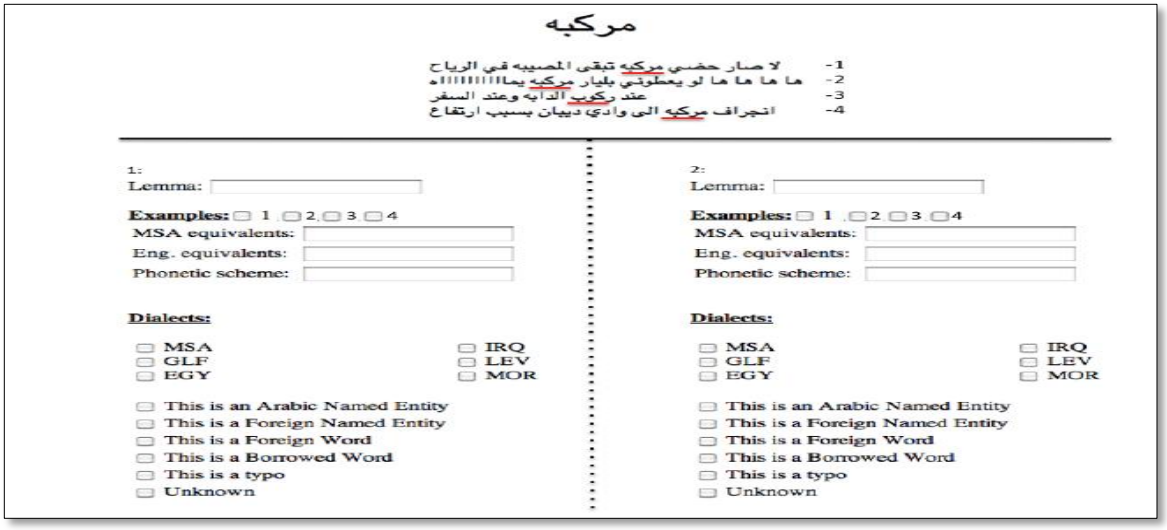


Figure 2.9: Lemma creation interface [4]

Finally; the application asks annotators to enter POS tag for it for each lemma created in the previous step. Figure 2.10 shows the interface that annotators used in this step.

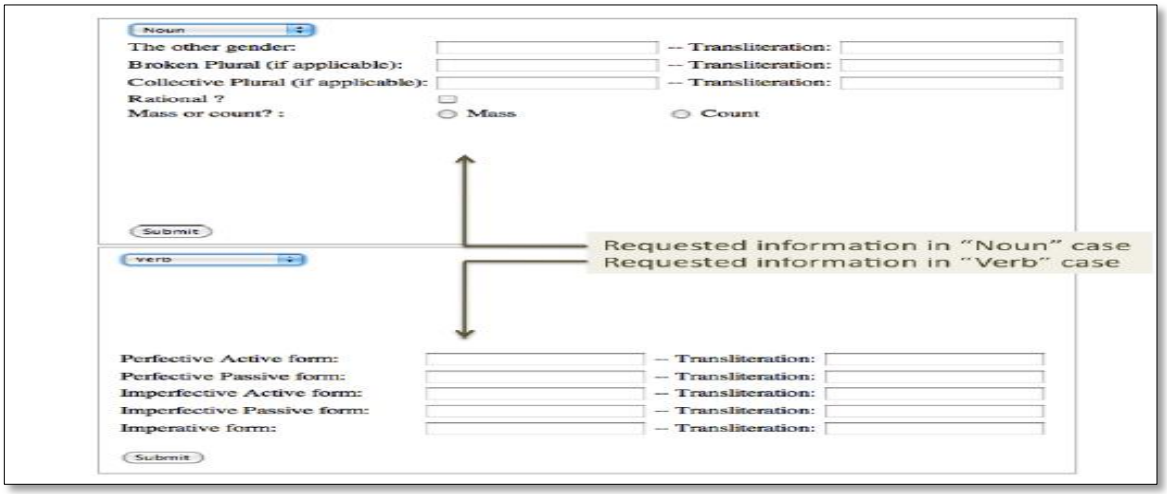


Figure 2.10: Morphological Profile creation interface [4]

All information entered by annotators using the application is stored in the database. This database contains 22 relational tables that are responsible for saving basic information, annotation information, assignment information, user permissions information, and user connection information.

The Final Stage in COLABA project is using DIRA (Dialectal Information Retrieval for Arabic) which is responsible for retrieving all MSA and dialect data for a word, DIRA works as follows: It takes a verb as input, then generates three surface forms for it; these surface forms are: MSA inflected forms, as example of the verb "أصبح", it generates "يصبح", "سيصبح","أصبحنا" …etc, MSA with dialectal morphemes, as example for verb "أصبح", it also generates "بيصبح"."حيصبح","هيصبح", "حيصبحوا"…etc, MSA lemma is translated to Dialect lemma, as example of the word "أصبح", it translated to "بقى"with all morphemes such as "حيبقى", "بيبقى", Then DIRA re-injected all generated forms into original query to retrieve Dialectal data.

2.2 Dialectal Orthography

This section reviews works done on building Conventional Orthography for Dialectal Arabic (CODA) [8,9, 17]. The most important problem that has been noted in all works done under Dialectal Arabic area is the lack of conventional orthography; in Arabic dialect, there are many ways to writing the same words, and there are also a lot of commons that are shared among all Arabic dialects. So there is a need for writing in a standard way over all Arabic dialects. From these issues come the needs for CODA, which is a conventional orthography for dialectal Arabic. CODA aimed to satisfy many Goals such as; consistent and coherent standards for writing Arabic Dialects, build for computational purposes, unified framework for writing all Arabic Dialects, and Save a level of uniqueness over each dialect. CODA builds conventions according to similarities between Modern Standard Arabic and its dialect.

In order to achieve these goals, CODA teams made many design decisions such as; using Arabic Script, unique orthographic form for each dialect; this form represents dialect phonology, morphology, and lexical

semantics, Orthographic decisions are displayed in MSA-like format such as spelling the Definite article "ال" morphologically.

Generally speaking, CODA tries to preserve phonology as it in word; e.g. راجل، كتاب ، اتكتب , but there is some exceptions, such as a word that contains /q/ which differs between dialects, then the word is written in MSA form; e.g.: قصر, and if the word in dialect has short vowel that comes from long vowel in MSA then it is written in MSA Form; such as **طبور** will be written **طابور**.

CODA also tries to preserve the morphology as it in word, But there is an exception here if there is a negation or pronoun that comes with a verb CODA separates it from the word; e.g. **ما كتبتلوش** will write **ما كتبت لوش** in CODA.

Also CODA team decides to keep the word order in a sentence as in the original sentence, to develop CODA in a way that is easy to learn and write, and to build the unique CODA Map for each dialect that contains rules and exceptions.

Until now, CODA has been applied for Egyptian Dialect (EGY) and Tunisian dialect (TUN). So, according to the CODA design principles a word is written as pronounced unless there is an exception on; this exception maybe:

- Morphological exception; in EGY there are many morphological exceptions, such as; in noun and adjectives ة CODA teams decided to -writ as ة not ,هand if the word has clitics then it becomes either ت or ا, ين suffix (dual & plural) does not vary according to its case as MSA, add vowels for geminates in verbs such as مديت not مددت, the plural suffix (توا-وا) spelled with silent Alif in CODA, in affixes that relate to feminine, there must be an extra ي at the end. e.g. اكتبيif particles is single then according to CODA it will be attached to the word such as بالطول, also in CODA pronominal pronouns and negations part must be attached at the end of word, ال التعريف written as is, in some cases we removed some letters from the original word such as : ل+ال written as ة ,لل becomes ا or ت e.g. معلمة+هم written as معلمتهم or معلماهم according to context, ا in واو الجماعة is removed e.g. كتبوا+ها is

written as **كتبوها**, **ى** becomes **ا** or **ي**, and finally there are dialectal clitics such as ,progressive part **ب** this progressive part add to original verb such as **ب+اكتب** written as **باكتب**, Future part **ح** e.g. such as **حامشي**,Second person plural **كو** e.g. **شافوكو**, and Feminine clitics **ي** e.g. **شافوكي**. Also in TUN there are many exceptions, such as; interrogation proclitic **شي**, e.g. **وشريتوهاشي**, Negation part **ش**, and Single litter colitis **ع**, **م**

- Phonological exception, in EGY there are the following phonological exceptions such as; CODA EGY team decided that there are 9 consonants that spelled differently from their phonology in DA if these constants have a dialect Arabic root radical and its dialect Arabic root has MSA root ; these consonant are :/**ء**/ spelled /**ق**/ ; /**ت**/-/**س**/ spelled /**ث**/; /**ز**/-/**د**/ spelled /**ذ**/; /**ز**/-/**ظ**/ ; /**ص**/ spelled /**ص**/-/**س**/; /**ض**/-/**ذ**/-/**ز**/ spelled /**ظ**/ ; /**ض**/ spelled /**ض**/-/**د**/ /**ط**/ spelled /**ط**/-/**ت**/, preserve long vowels as is such as "**كتبين** which - written in CODA as **كاتبين**", if there is multiple long vowels that shorten to one in EGY; then CODA team decided to write two long vowels as MSA; e.g. **قنون** is written as **قانون** in CODA,

different writing for ي and و in Arabic script according to context. E.g. دور that maybe /du:r/ or /do:r/ according to context it appears on it, CODA team decided to keep Hamza as it; in other words if peoples write word without Hamza or with Hamza CODA team decided to be as is, for example if CODA team face "**مايل**" in text, then will preserve it as is, and if they face "**مائل**" will also keep as is -, and finally ي،ى in the end of words in CODA will be spelled correctly, as example"**الكلام ده علي مين** " will be written in CODA" **الكلام ده على مين**". Also in TUN there are the following phonological exceptions, such as same exception in 9 consonants as EGY with two special cases in TUN which are /ق/, /غ/ spelled /ق/ as e.g. /bagra/ is written in CODA as /baqrah/, and Consonants with multiple punctuations are written in a form closes to MSA if there is MSA, and Finally TUN CODA adds ن after some numerals in some cases such as **خمسطاشن راجل**.

- Dialectal lexical exception, such as **برضه** not **برضو**, **ده** not **ذه** in EGY, and **هذاك** not **هاذاكة**, **عالسلامة** not **عسلامة**, and Foreign words

that use non-Arabic phoneme /g/, /v/ and /p/ will be written in CODA as q, f and b. e.g. **فيسته, قازوز** in TUN.

Figure 2.11 & 2.12 shows examples of EGY CODA and TUN CODA

Raw Text	هه هه والة العظيم فطست من الضحك اية ياعبير المواضيع الجامدة دى دحنا كدة بقى عندنا بنك كامل متكامل على العموم انا اجتهدت وجبت شوية نكت واكيد طبعا منقولين بس يارب يعجبوكوا اسيبكوا مع النكت . هنا رقد الرجل على فراشة يغالب الغيبوبة وكلما افاق وجد زوجتة بجانبة وتنظر الية بحنان فامسك بيديها قائلا : لما اترفدت وقفتى معايا . ولما شركتى فلست كنتى جمبى . ولما بيتنا اتحرق كنتى جمبى . ودلوقتى انتى بردة جمبى . مش عارف لية انا حاشش انك نحس *hh hh wAllh AlçĎym fTst mn AlDHk Ayh yAçbyr AlmwADyç AljAmdh dý dHnA kdh bqý çndnA bnk kAml mtkAml çlý Alçmwm AnA Ajthdt wjbt šwyh nkt wAkyd TbçA mnqwlyn bs yArb yçjbwkwA AsybkwA mç Alnkt . hnA rqd Alrjl çlý frAšh yγyAlb Alγybwbh wklmA AfAq wjd zwjth bjAnbh wtnĎr Alyh bHnAn fAmsk bydyhA qAŷlA : lmA Atrfdt wqftý mçAyA . wlmA šrktý flst kntý jmbý . wlmA bytnA AtHrq kntý jmbý . wdlwqtý Antý brdh jmbý . mš çArf lyh AnA HAšš Ank nHs*
CODA	هه هه واله العظيم فطست من الضحك إيه يا عبير المواضيع الجامدة دي داحنا كده بقى عندنا بنك كامل متكامل على العموم انا اجتهدت وجبت شوية نكت وأكيد طبعا منقولين بس يا رب يعجبوكو اسيبكو مع النكت . هنا رقد الرجل على فراشه يغالب الغيبوبة وكلما أفاق وجد زوجته بجانبه وتنظر إليه بحنان فأمسك بيديها قائلا : لما اترفدت وقفتي معاي . ولما شركتي فلست كنتي جنبي . ولما بيتنا اتحرق كنتي جنبي . ودلوقتي انتي برضه جنبي . مش عارف ليه انا حاسس انك نحس *hh hh* **wAllh** *AlçĎym fTst mn AlDHk* **ĂyhyA** **çbyr** *AlmwADyç AljAmdh* **dy dAHnA kdh** *bqý çndnA bnk kAml mtkAml çlý Alçmwm AnA Ajthdt wjbt šwyh nkt* **wÂkyd** *TbçA mnqwlyn bs* **yA rb yçjbwkw Asybkw** *mç Alnkt . hnA rqd Alrjl çlý* **frÂšh** *yγAlb Alγybwbh wklmA* **ÂfAq wjd zwjth bjAnbh** *wtnĎr* **Ălyh** *bHnAn* **fÂmsk** *bydyhA qAŷlA : lmA Atrfdt* **wqfty mçAy** . *wlmA* **šrkty** *flst* **knty jnby** . *wlmA bytnA AtHrq* **knty jnby** . **wdlwqty Anty brDh jnby** . *mš çArf* **lyh** *AnA* **HAss** *Ank nHs*
English	ha ha [.] I swear to God [.] I died from laughter [.] Abeer [.] what cool topics [!] we now have a complete comprehensive bank [.] any way [.] I put some effort and got some jokes that are of course copied [.] but hopefully you will like them [.] I leave you with the jokes . [MSA] There lied a man on his bed coming in and out of a coma [;] and every time he woke up he found his wife by his side looking at him lovingly [.] so he held her hands and said [/MSA]: when I got fired [.] you stood by me . And when my company went bankrupt you were by my side . And when our house burnt down you were by my side . And now also you are by my side . I don't know why I feel you're bad luck [.]

Figure 2.11: EGY CODA Example [8]

Raw Text	مساء الخير مرحبا بكم في المباشر في ناس نسمة سبسيال اليومة في ساعتنا الثانية باش نحكيو على تطورات جديدة . فمة تطورات في موضوع تشكيل الحكومة . نحكيو عليها مع التحالف الديموقراطي . نشوفو وين وصلت الأمور معاهم ، ونشوفو شنو سار في الأحداث متاع مقبرة زلاز . و بش نحكيو زادة على الزيادات في اسوام القاز و نسهلو ع انعكاسات متاعو ع المواطن . *msA' Alxyr mrHbA bkm fy AlmbAšr fy nAs nsmħ sbsyAl Alywmħ fy sAçtnA AlθAnyħ bAš nHkyw çlý tTwrAt jdydħ . fmħ tTwrAt fy mwDwç tškyl AlHkwmħ . nHkyw çlyhA mç AltHAlf AldymwqrATy . nšwfw wyn wSlt AlÁmwr mçAhm , wnšwfw šnw sAr fy AlÁHdAθ mtAç mqbrħ zlAz . w bš nHkyw zAdħ çlý AlzyAdAt fy AswAm AlgAz w nshlw ç AnçkAsAt mtAçw ç AlmwATn .*
CODA	مسا الخير مرحبا بكم في المباشر في ناس نسمة سبسيال اليوم في ساعتنا الثانية باش نحكيوا على تطورات جديدة . ثمة تطورات في موضوع تشكيل الحكومة . نحكيوا عليها مع التحالف الديموقراطي . نشوفوا وين وصلت الامور معاهم ، ونشوفوا شنوة صار في الاحدا متاع مقبرة جلاز . وباش نحكيوا زادة على الزيادات في اسوام القاز ونسألوا عالانعكاسات متاعه عالمواطن . *msA Alxyr mrHbA bkm fy AlmbAšr fy nAs nsmħ sbsyAl Alywm fy sAçtnA AlθAnyħ bAš nHkywA çlý tTwrAt jdydħ . θmħ tTwrAt fy mwDwç tškyl AlHkwmħ . nHkywA çlyhA mç AltHAlf AldymwqrATy . nšwfwA wyn wSlt AlAmwr mçAhm , wnšwfwA šnwħ SAr fy AlAHdAθ mtAç mqbrħ jlAz . wbAš nHkywA zAdħ çlý AlzyAdAt fy AswAm AlqAz wnsÂlwA çAlAnçkAsAt mtAçh çAlmwATn .*
English	Good evening. Hello. You are on the air with "Nesma People Special Program". Today in our second hour, we'll talk about some new developments. There are developments on the subject of forming the government. We will discuss it with the Democratic Alliance. We will see where they have reached on this topic and we will discover what happened in the events at Jlaz cemetery. We will talk also about the increase in gas prices and question its impact on the citizen.

Figure 2.12: TUN CODA Example [9]

So as you have noticed in Figures 2.11 & 2.12, the first part of the figure contains text as appeared; in the second part, the text rewritten in CODA; the last part of the figure contains the corresponding English text.

CODA guidelines will be extended to cover PAL in this thesis, as discussed in chapter 4.

2.3 Dialectal Morphological Annotation

Most of the previous work that is done under the area of Morphology in Arabic focused on MSA, but in Dialects the works are relatively small or it depends on MSA to deal with Dialects. But as stated previously available MSA tools cannot be easily extended or transferred to work properly for Dialects according to varying between MSA and its Dialects. So it is important to develop annotated and morpheme-segmented resources and morphological analysis tools to deal with Dialects. One of the most recent contributions that deals with Dialects is CALIMA which is a morphological analyzer for EGY [19]. CALIMA and MSA analyzer SAMA are also used in EGY morphological tagger MADA-ARZ and in MADAMIRA.

2.3.1 MADAMIRA

MADAMIRA is a morphological analyzer for Arabic and its dialect, MADAMIRA combines two systems for Arabic processing. These systems are:

- MADA

A morphological analysis and disambiguation for Arabic (MADA) depends on presenting many analyses for each word, then selecting the suitable analysis according to context. In order to do this, MADA contains 19 orthogonal features that suitable analysis is selected depends on it. These features are: Part_Of_Speech(POS/pos); such as Noun, Verb, …etc, presence of conjunction(CNJ/conj); such as w, and f, presence of a particle clitics(PRT/part); such as b, k, and l, presence of pronominal clitics(PRO/clitic); such as object, and possessive, presence of definite article(DET/art) ; such as Al, gender(GEN/gen) ; such as FEM, and MASC, number(NUM/num) ; such as SG ,DU, and PL, Person(PER/per) ; such as 1,2, and 3, voice(VOX/voice) ; such as PASS, and ACT, Aspect(ASP/aspect) ; such as IV, CV, and PV, mood(MOD/mood) ; such as I, S, J, and SJ, presence of nunation (NUN/def) ; such as DEF, and

INDEF, construct state(CON/idafa) ; such as POSS, and NOPOSS, and Finally case(CAS/case) ; such as ACC, GEN, and NOM

Each of these features is weighted. So when a new word comes, MADA provides a list of potential analyses by using a Buckwalter Arabic morphological analyzer (BAMA), and then the analysis that gains most of the features will be selected by MADA. Figure 2.13 shows a sample of MADA outputs

INPUT	wsynhY	Alr}rys	jwlth	bzyArp	AlY	trkyA	.
GLOSS	and will finish	the president	tour his	with visit	to	Turkey	.
ENGLISH	The president will finish his tour with a visit to Turkey.						

```
;;; SENTENCE wsynhY Alr}ys jwlth bzyArp AlY trkyA .
;;WORD wsynhY
;;MADA: wsynhY art-NA aspect-IV case-NA clitic-NO conj-YES def-NA mood-I num-SG part-NO per-3 pos-V
       voice-ACT
*0.930061 wasayunohiy=[>anohaY_1 POS:V +IV s+ MOOD:I +S:3MS w+
       BW:wa/CONJ+sa/FUT+yu/IV3MS+nohiy/IV+(null)/IVSUFF_MOOD:I]=complete/finish/communicate
^0.780654 wasayanohaY=[nahaY-i_1 POS:V +IV s+ MOOD:I +S:3MS w+
       BW:wa/CONJ+sa/FUT+ya/IV3MS+nohaY/IV+(null)/IVSUFF_MOOD:I]=forbid/restrain
_0.739338 wasayunohaY=[>anohaY_1 POS:V +IV s+ +PASS MOOD:I +S:3MS w+
       BW:wa/CONJ+sa/FUT+yu/IV3MS+nohaY/IV_PASS+(null)/IVSUFF_MOOD:I]=be_completed/be_communicated
[ ... 7 additional options omitted ...]
```

Figure 2.13: MADA output sample [15]

Then MADA was extended to MADA ARZ which is the Egyptian version of MADA, the main major changes in this extension occurred in morphological analyzer used. In MADA ARZ, the morphological

analyzer that used is CALIMA; CALIMA is a Morphological analyzer for Egyptian Arabic;

Building CALIMA was passed through many stages. Firstly, choose Egyptian Colloquial Arabic Lexicon (ECAL) which contains 27K verbs, 36K nouns and adjectives, 1.5K proper nouns, and 1K closed class; Each ECAL entry consists of phonological form, undiacritized orthography, lemma, and morphological features, then diacritrize each ECAL entry, then write rules for converting from diacritized form to CODA and rules for converting from ECAL morphology to LDC EGY POS tags, then update lemma according to lemma Specification in SAMA, after that each mapped ECAL entry was converted to SAMA-like representation.

Applying approach described above generates six tabled for CALIMA. These tables are: complex prefixes, complex suffixes, complex stem, prefix-stem, prefix-suffix, and stem suffix and each table was also extended to contain non CODA variants.

CALIMA has 100K stem corresponding to 36K lemmas, 2421 complex prefixes, 1179 complex suffixes.

- **AMIRA**

AMIRA is a tool that is used for processing Arabic; it is based on supervised learning with no dependency on any morphological knowledge, AMIRA toolkit contains: Clitic token (TOK), Part of Speech Tagger (POS), and Base phrase chunker (BPC).

So when an input text is entered to MADAMIRA, it is cleaned and converted to Buckwalter representation schema, then MADAMIRA builds all possible analysis for each word using either SAMA analyzer for MSA or CALIMA analyzer for EGY Arabic, after that MADAMIRA gives predictions for the word's morphological features and then gives score for each word analysis and sorts the analyses according to the score, after that MADAMIRA tokenizes the top score analysis according to the schema requested by the user, then MADAMIRA divides the input text into chunks, Finally MADAMIRA passes the input text to Named Entity Recognizer in order to mark and categorize the named entities. Figure2.14 shows MADAMIRA architecture

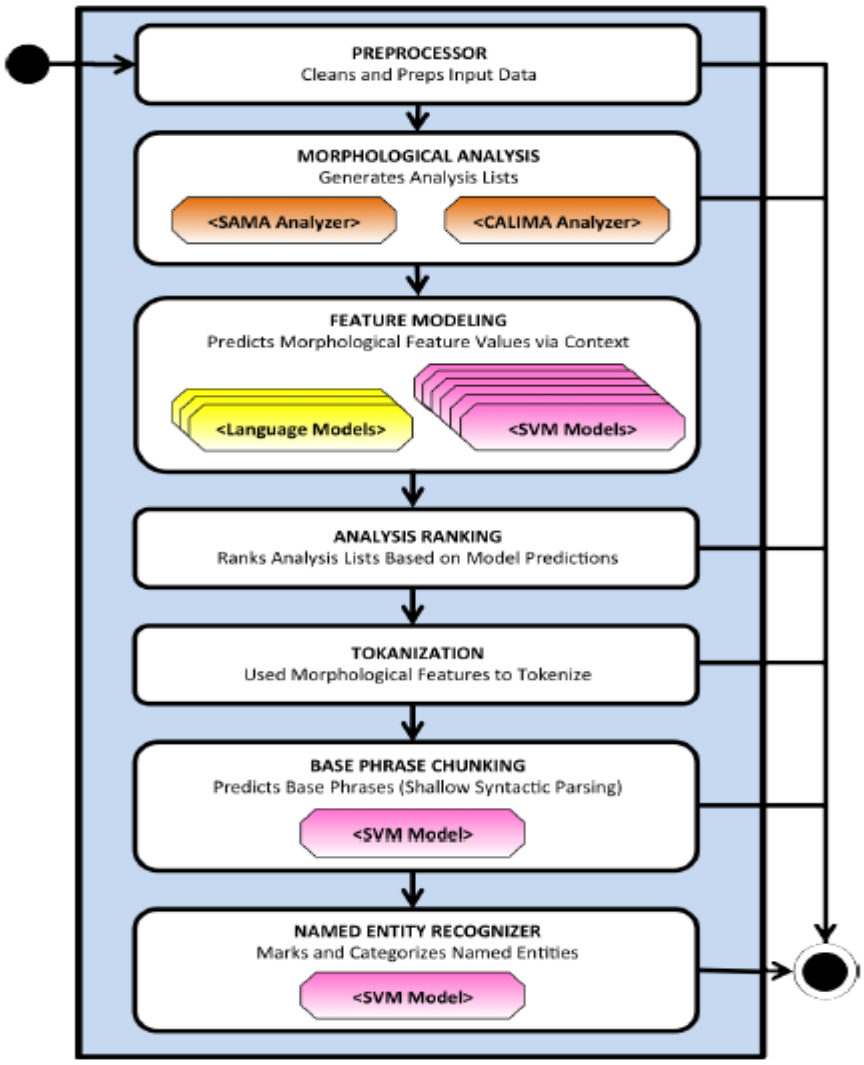


Figure 2.14: MADAMIRA architecture [16]

In order to evaluate MADAMIRA ; authors choose a set of 25,000 words for standard Arabic and 20,000 words for EGY Arabic then ask MADAMIRA to analyze them and compare MADAMIRA results with gold annotation list ; the results are : the percentage of words that

diacritized correctly in MADAMIAR is 86.3% for MSA and 83.2% for EGY, the percentage of words that MADAMIRA gives its lemma correctly is 96.0% for MSA and 87.8% for EGY, the percentage of words that MADAMIRA gives its part-of-speech tag correctly is 95.9% for MSA and 92.4% for EGY, and the percentage of words that MADAMIRA gives all morphological features (match gold entry exactly) correctly is 84.1% for MSA and 77.3% for EGY. Figure2.15 shows MADAMIRA interface

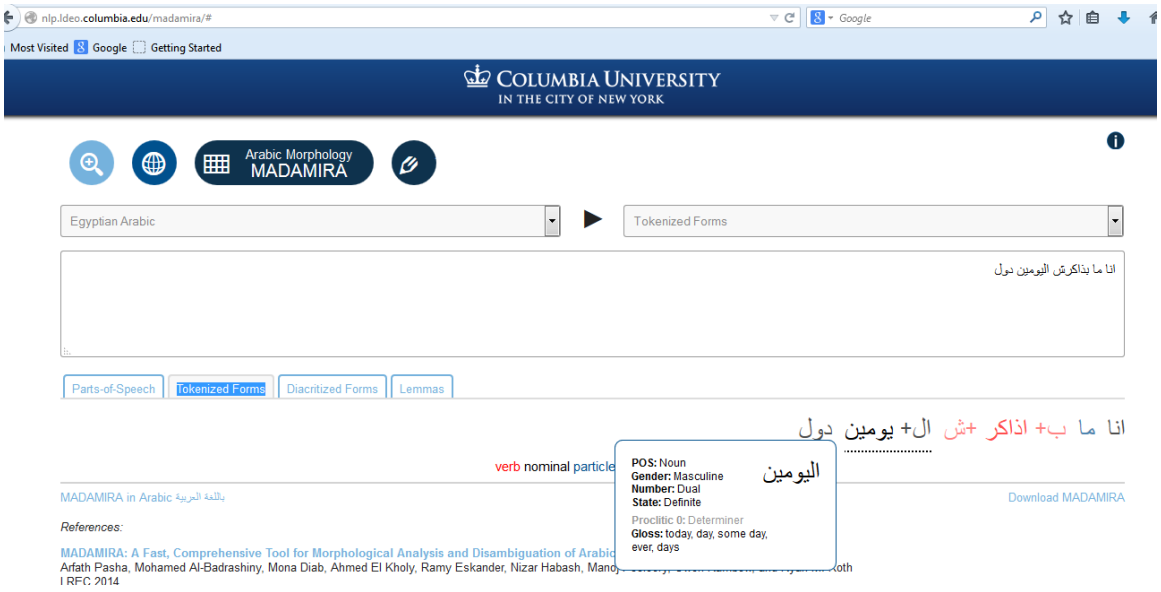


Figure 2.15: MADAMIRA Demo online interface [16]

2.4 Remarks on the related works and literature

As we seen previously there are many works done in dialects but most of them have a lot of issues. These issues may be related to small size of dialectal data as work presented in sub-section 2.1.1 which has only 33.000 words, subsection 2.1.2 which depends in too small dictionary, also numbers of queries that designed to collect data from web in subsection 2,1,3 is relatively small because there are a huge data published on the web and 40 queries couldn't collect all of it.

Also, There is another issue comes from that some works were depended on Standard Arabic to deal with dialects which didn't lead to significant results as work presented in subsection 2.1.2.

Also, all works presented in section 2.1 didn't take into account the lack of orthographic and inconsistency in dialectal data.

These previous efforts start being on the train when CODA, and MADAMIRA which presented in section 2.2 and 2.3 start, but also these two work have many issues such as CODA need to be extended to cover more dialects, also MADAMIRA need to be trained on more data in both

MSA and EGY because sometimes it couldn’t give results for MSA word.

Chapter 3

Corpus Collection

This chapter gives a detailed description about the corpus. Section 3.1 presents corpus collection process and corpus statistics. Section 3.2 presents corpus database

3.1 Corpus

3.1.1 Corpus collection methodology

As we stated previously, dialects data are founded as oral data, not as written data, so it is too difficult to find resources for written dialectal content. Dialect data have also a lot of noise and inconsistency due to the lack of orthographic standards for dialects. Hence, the same word may be in different formats in dialects. So lack of resources and noise led to

many challenges in the collection of high-coverage and high-accuracy dialect corpora. We decided to focus on precision and variety more than on size in our corpus, so when we collected our corpus, we tried to cover a variety of topics and contexts, localities and sub-dialects, including the social class and gender of the speakers and writers [20].

3.1.2 Corpus Statistics

As we stated above, we are collecting our corpus manually from different resources, and different context. The most important part of our corpus is the famous Palestinian series "Watn E Watr وطن ع وتر". Our corpus documents are:

- **Facebook**

The text has been manually collected from different Palestinian pages on Facebook. This is done by crawling many Palestinian Facebook pages such as: "يما بديش أتجوز" page (https://www.facebook.com/mombdesh2tjwaz?ref=br_tf); this page contains several Palestinian jokes in different Palestinian sub-dialects and it collected in June, 2013, and "شبكة فلسطين للحوار" page (https://www.facebook.com/paldf); this page contains several political

discussions in Palestinian dialect. Total number of threads in this document is 35, total number of word tokens are 3120 and total number of word types is 1985. Figure 3.1 shows a sample of Facebook document.

Figure 3.1: Sample from Facebook document

- **Twitter**

The text has been manually collected from accounts of Palestinian peoples who are tweet in Palestinian dialects; those peoples come from different sub-dialects. Accounts chosen here are reflecting different cultures and different sub-dialects such as: "مروة الخطيب" account(https://twitter.com/Marwa_101); this account is for girl from umm-alfahm, "هبة عبد السلام الحايك" account(https://twitter.com/Hebahayek); this account is for girl from Gaza, and "Tamer Hammam" account

(https://twitter.com/itamer83); this account is for young man from Gaza. Total number of threads in this document is 38, total number of word tokens is 3541 and total number of word types is 2133. Figure 3.2 shows a sample of Twitter document.

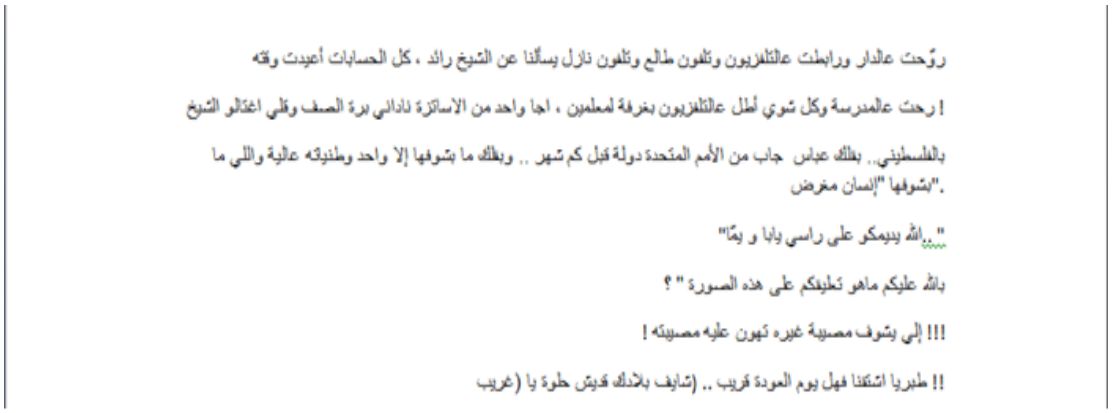
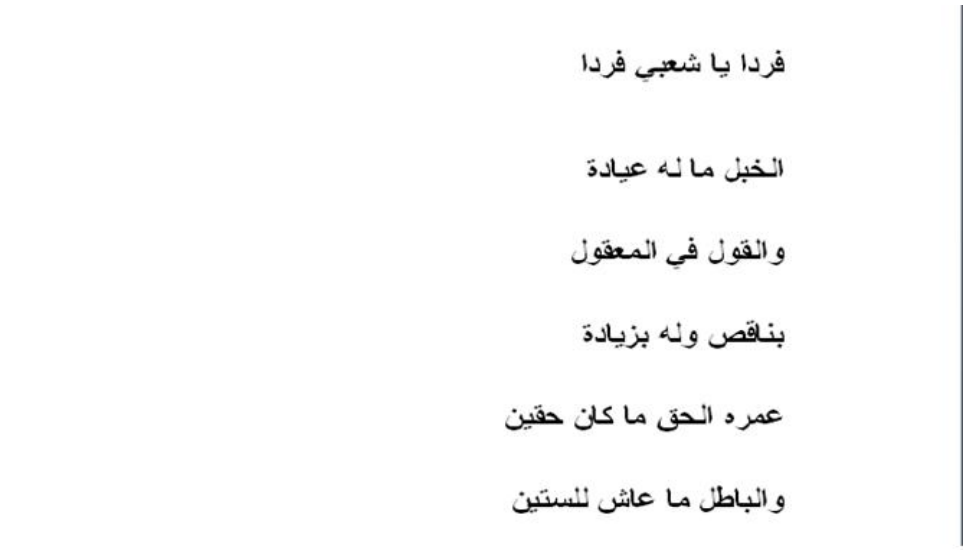


Figure 3.2: Sample from Twitter document

- **Blogs**

The text has been manually collected from "عبد الحميد عبد العاطي" blog, it contains many of poets that have written in Palestinian dialect. Total number of threads in this document is 37, total number of word tokens is 8748 and total number of word types is 4454. Figure 3.3 shows a sample of Blogs document


فردا يا شعبي فردا

الخبل ما له عيادة

والقول في المعقول

بناقص وله بزيادة

عمره الحق ما كان حقين

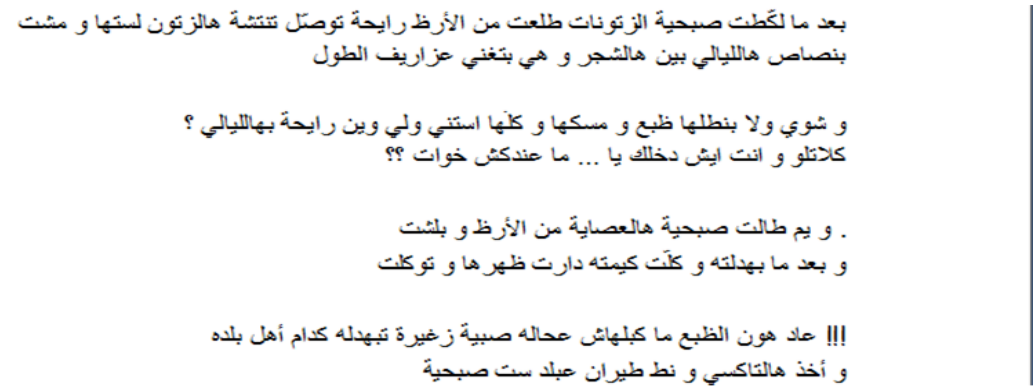
والباطل ما عاش للستين

Figure 3.3: Sample from "عبد الحميد عبد العاطي" blog

- **Palestinian Stories**

The text has been manually collected from different forums, it contains six stories that written in Palestinian dialect such as: "قصة صبحية و الظبع" (http://www.paledream.com/vb/showthread.php?t=16157); this story is written in rural Palestinian sub-dialects, "قصة جبينة" (http://www.omaniyat.com/vb/showthread.php?t=21995); this story is also written in rural Palestinian sub-dialects, "قصة سندريلا"(http://forum.sedty.com/t428456.html) ; which is written in rural Palestinian dialect, "قصة الملك و بناته الثلاث" which is written in standard Palestinian dialect, "قصة الغرة و الجرة"(

http://pulpit.alwatanvoice.com/articles/2012/05/10/260140.html) which is written in Jaffa rural Palestinian dialect, and finally a dialogue between number of Palestinian rural women. The total number of word tokens is 2407 and total number of word types is 1422. Figure 3.4 shows a sample of Stories document


بعد ما لكّطت صبحية الزتونات طلعت من الأرظ رايحة توصّل تنتشة هالزتون لستها و مشت
بنصاص هالليالي بين هالشجر و هي بتغني عزاريف الطول

و شوي ولا بنطلها ظبع و مسكها و كلّها استني ولي وين رايحة بهالليالي ؟
كلاتلو و انت ايش دخلك يا ... ما عندكش خوات ؟؟

. و يم طالت صبحية هالعصاية من الأرظ و بلشت
و بعد ما بهدلته و كلّت كيمته دارت ظهرها و توكلت

!!! عاد هون الظبع ما كبلهاش عحاله صبية زغيرة تبهدله كدام أهل بلده
و أخذ هالتاكسي و نط طيران عبلد ست صبحية

Figure 3.4: Sample from Palestinian story

- **Forums**

The text has been manually collected from "منتديات شبكة فلسطين للحوار" (https://www.paldf.net/forum/index.php), it contains many of the discussions in Palestinian Dialect. Total number of threads in this document is 33, total number of word tokens is 1092 and total number of word types is 798. Figure 3.5 shows a sample of forums document

i. كل اتنين طلعوا في الشارع بدنا نعمل لهم متابعة ؟؟!!

2. 2-أعمال عنف تتعامل معها الحكومة بكل توازن فلم يقتل شخص واحد على الرغم من ان المتظاهرين اعتدوا ودمروا مؤسسات عامة وخاصة يا ريت نظام الشبيحة وأعوانه يتعلموا كيف يتعاملوا مع البشر مثل أردوغان , بعدين الموضوع مش كبير لحتى نفتح له محور للمتابعة كل القصة متنزه الحكومة بدها تشيلوا وتبني مكانوا مجمع تجاري وين المشكلة يعني الموضوع مش موضوع حزب له متربع على الحكم 40 حكومة تركيا الشعب اختارها ثلاثة فترات انتخابية متتالية لما وجده من نجاح , سنة وبيمنع الناس يتنفسوا لحزب العدالة والتنمية

3. 3- اتنين

شكلك مش عايش بالدنيا

فوق ال100 ألف روح شوف التلفزيون

Figure 3.5: Samples from Forum

- **Palestinian Terms**

It is one document that contains 556 Palestinian terms with its meaning; these terms are collected manually from the web (https://www.paldf.net/forum/showthread.php?t=119188&page=2). Figure 3.6 shows a sample from these terms.

مصطلحات :

بنغل نغل

انخم ، انطم ،سد بوزك ، حط هوى في تمك و اخرس ، اخرس

برة = طلطميس // يعني واحد ما بيفهم شي

ابو هطل ////برضو واحد هبيلة يعني

...عارض = مش عارف كيف أعبرلكم عنها .. بس يعني مرات بنقول عارض ياخدك ايش هوة العارض مش عارف

الكورية = راس الشارع = لغة آخر الشارع

Figure 3.6: Samples from Palestinian terms

- **Series "وطن ع وتر"**

Episodes of "وطن ع وتر" series, which was broadcast on the "الفلسطينية" channel, that have been obtained from those in charge of the series in the"مسرح حياة", and the idea in series based on a critique of the conditions of Palestinian society in Comic way. Total number of episodes is 41, total number of word tokens is 23423 and total number of word types is 8459. Figure 3.7 shows one of "وطن ع وتر" episodes

بنزين

مشهد الكازية

عامل في كازية يخاطب مجموعة من الناس (يفضل الكل حامل غلن بنزين)

العامل: ما فيش بنزين ولا سولار للسيارات ولا للماتورات، البنزين والسولار بس لطلاب التوجيهي عشان يشغلوا الماتور ويدرسوا، اللي معش توجيهي الله معاه

يدخل الطالب

الطالب: عبيلي

العامل: انت توجيهي

الطالب: اه

العامل: علمي ولا ادبي ولا تجاري؟

الطالب: ادبي

العامل: اتسوف بطاقة الجلوس ورقم المقعد (**الطالب يعطيه البطاقة**)

العامل: استنى بدي احكي مع مدير المدرسة اتأكد، لانو صورتك معش واضحة (يتصل) الو ابو خالد عندك طالب توجيهي اسمو محمد سمعان شعراتو سبايكي؟ شاطر في المدرسة، ساقط 3 امتحانات لهلقيت، ماشي سلام

Figure 3.7: one of "وطن ع وتر" episodes

- **Dictionary of Palestinian Vocabularies and Loan words"** معجم العامي و الدخيل"

It contains Palestinian Dialect words from "أ" to "ي", and for each word, author was listed the meanings of a word, and where to use it, along with its corresponding Standard Arabic word, it also contains 53 morphological rules for Palestinian Dialect. Total number of pages in this book is 646, and total number of interpreted terms as (معس ، إجر، طوطح... الخ) is 5595. The Author of Book is Hussien Ali Lubany, and Publisher of The Book is Lebanon Library. Figures 3.8 -3.10 show samples from the dictionary

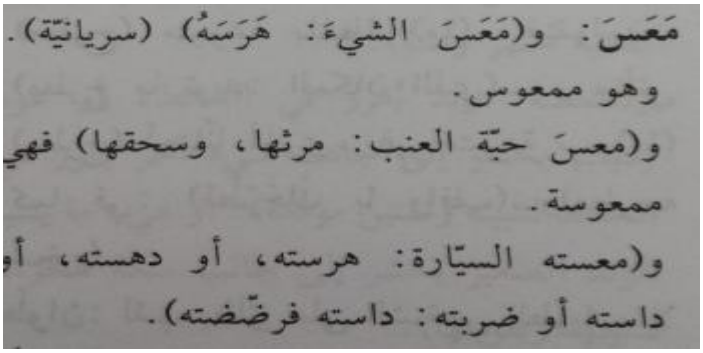
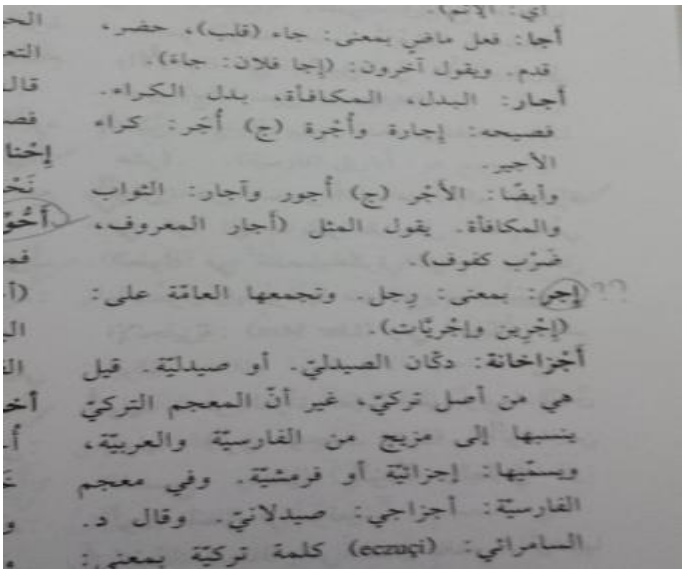

مَعَسَ: و(مَعَسَ الشيءَ: هَرَسَهُ) (سريانيّة).
وهو ممعوس.
و(معسَ حبّة العنب: مرثها، وسحقها) فهي
ممعوسة.
و(معسته السيّارة: هرسته، أو دهسته، أو
داسته أو ضربته: داسته فرضّضته).

Figure 3.8: Sample from Dictionary



أجا: فعل ماضٍ بمعنى: جاء (قلب)، حضر، قدم. ويقول آخرون: (إجا فلان: جاء).

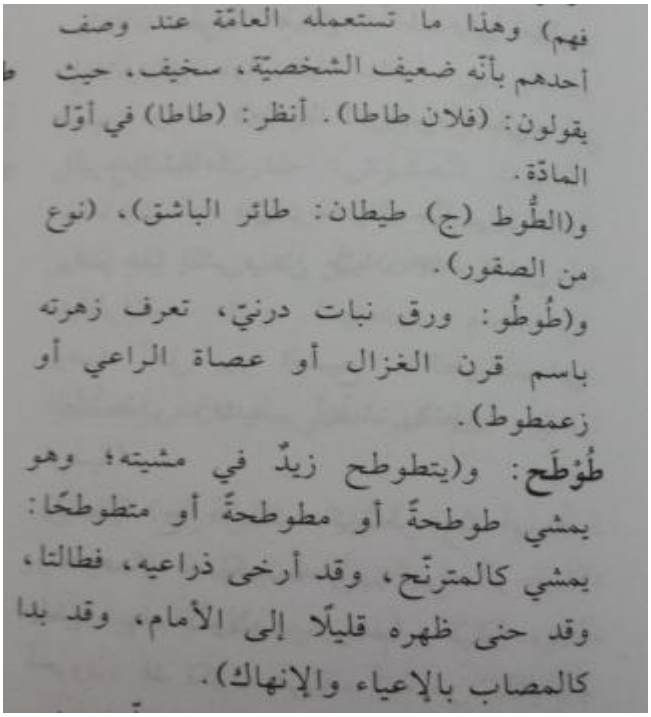
أجار: البدل، المكافأة، بدل الكراء. فصيحه: إجارة وأجرة (ج) أجَر: كراء الأجير.

وأيضًا: الأجر (ج) أجور وآجار: الثواب والمكافأة. يقول المثل (أجار المعروف، ضرب كفوف).

إجر: بمعنى: رجل. وتجمعها العامّة على: (إجرين وإجريّات).

أجزاخانة: دكّان الصيدليّ. أو صيدليّة. قيل هي من أصل تركيّ، غير أنّ المعجم التركيّ ينسبها إلى مزيج من الفارسيّة والعربيّة، ويسمّيها: إجزائيّة أو فرمشيّة. وفي معجم الفارسيّة: أجزاجي: صيدلانيّ. وقال د. السامرائي: (eczaçi) كلمة تركيّة بمعنى:

Figure 3.9: Sample 2 from Dictionary



فهم) وهذا ما تستعمله العامّة عند وصف أحدهم بأنّه ضعيف الشخصيّة، سخيف، حيث يقولون: (فلان طاطا). أنظر: (طاطا) في أوّل المادّة.

و(الطُّوط (ج) طيطان: طائر الباشق)، (نوع من الصقور).

و(طُوطُو: ورق نبات درنيّ، تعرف زهرته باسم قرن الغزال أو عصاة الراعي أو زعمطوط).

طُوطَح: و(يتطوطح زيدٌ في مشيته؛ وهو يمشي طوطحةً أو مطوطحةً أو متطوطحًا: يمشي كالمترنّح، وقد أرخى ذراعيه، فطالتا، وقد حنى ظهره قليلًا إلى الأمام، وقد بدا كالمصاب بالإعياء والإنهاك).

Figure 3.10: Sample 3 from Dictionary

3.2 Processing and Storing collected text

This section presents our work on parsing and storing collected texts in N-gram models. This work is conducted at the beginning of the research, but it turned out that is not needed as DIWAN tool was used to annotate the corpus. It is presented here as we think it is important when extending our work to for example annotating phrases, rather than only words in the corpus.

3.2.1 Storing methodology

Storing the corpus data is a preliminary step to achieve our goals, so we retrieve and manipulate the dialect words easily. To maximize the dialect words that we could extract from the available resources we decided to store the data using N-gram model (considering N $\in$[1-4]), this - due to the fact that -people -do not only use- single words but also phrases to express every things. We also decided to store the position of the word in -the document; this position can be used later to generate an equivalent document in Modern Standard Arabic (MSA) after translating

the dialect words into their correspondence MSA words. Figure 3.11 shows an ER model for corpus database

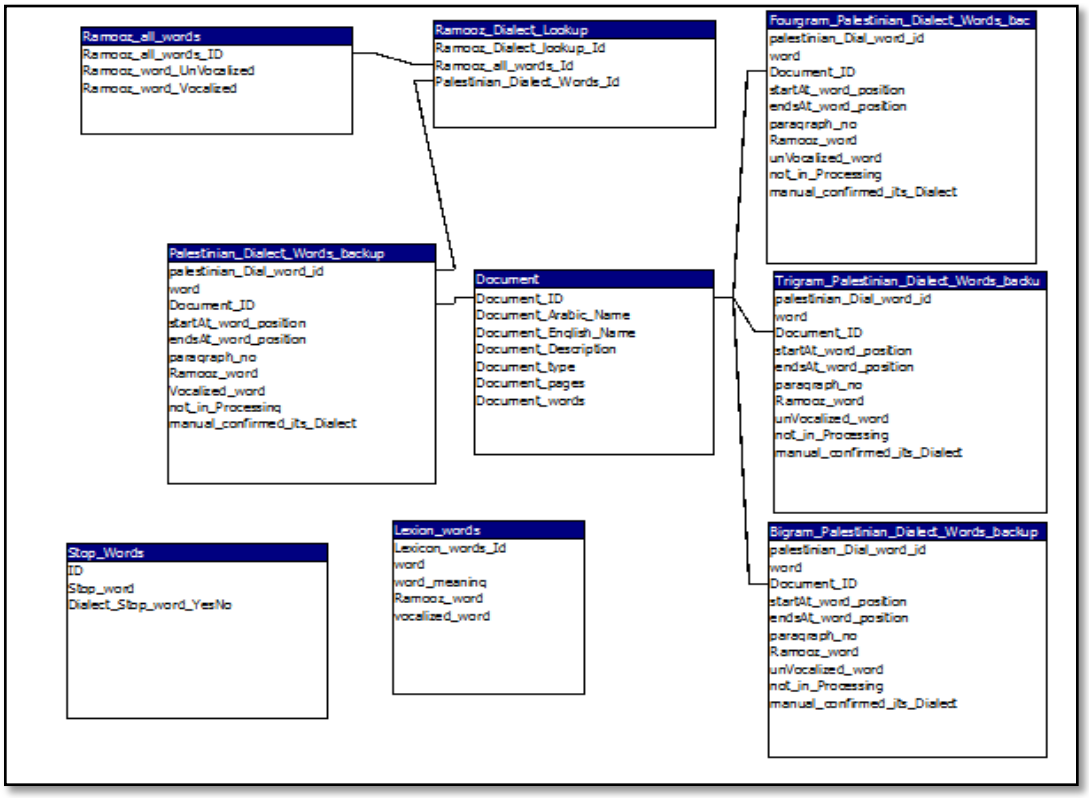


Figure 3.11: Database ER model

As shown in Figure 3.12, “Document” table is used to store the information about each document in the corpus; those information –are: document name, document description, number of pages, and number of words.

Figure 3.12 shows a sample of data in Document Table.

Document

Document_ID	Document_English_Name	Document_Arabic_Name	Document_Description	Document_typ	Document_	Document_words
1	Palestinian Dialect-Facebook	اللهجة الفلسطينية-فيس بوك	عدد من الصفحات على الفيس بوك	Microsoft Word	27	3947
2	Palestinian Dialect-Twitter	اللهجة الفلسطينية -تويتر	عدد من الأشخاص المفتوحة للكل)	Microsoft Word	15	3812
3	AbdAlhamid blog	مدونة عبد الحميد عبد العاطي	غزة يكتب فيها بالعامية الفلسطينية	Microsoft Word	42	9087
4	Palestinian Stories	قصص بالفلسطينية	القصص مروية باللهجة الفلسطينية	Microsoft Word	11	2772
5	Palestinian Forums	منتديات شبكة فلسطين للحوار	فلسطين للحوار بتاريخ 1-6-2013	Microsoft Word	7	863
6	Palestinian Vocubalries	مصطلحات فلسطينية	تديات شبكة فلسطين للحوار بتاريخ	Microsoft Word	7	1260
7	Watn3Water-SaddestArabic	وطن ع وتر -اتعس عربي	المجتمع الفلسطيني بشكل كومبدي	Microsoft Word	4	1058
8	Watn3Water-UnemployedSong	وطن ع وتر -اغنية نحنا ماعنا شباب تتوظف بشهادتها	المجتمع الفلسطيني بشكل كومبدي	Microsoft Word	2	299
9	Watn3Water-Program3	وطن ع وتر-البرنامج3	المجتمع الفلسطيني بشكل كومبدي	Microsoft Word	3	532
10	Watn3Water-Program4	وطن ع وتر-البرنامج4	المجتمع الفلسطيني بشكل كومبدي	Microsoft Word	4	973
11	Watn3Water-ProgramFinal	وطن ع وتر-البرنامج نهائي	المجتمع الفلسطيني بشكل كومبدي	Microsoft Word	3	587
12	Watn3Water-Program5	وطن ع وتر-البرنامج5	المجتمع الفلسطيني بشكل كومبدي	Microsoft Word	4	1015
13	Watn3Water-palestininanDream	وطن ع وتر-الحلم الفلسطيني	المجتمع الفلسطيني بشكل كومبدي	Microsoft Word	3	747
14	Watn3Water-Taxes	وطن ع وتر-الضرائب	المجتمع الفلسطيني بشكل كومبدي	Microsoft Word	3	546
15	Watn3Water-Family	وطن ع وتر-العائلة	المجتمع الفلسطيني بشكل كومبدي	Microsoft Word	3	803
16	Watn3Water-ProgramMovies	وطن ع وتر-برنامج البرنامج / افلام	المجتمع الفلسطيني بشكل كومبدي	Microsoft Word	3	574
17	Watn3Water-ProgramOfProgramP	وطن ع وتر-برنامج البرنامج من فلسطين	المجتمع الفلسطيني بشكل كومبدي	Microsoft Word	3	743
18	Watn3Water-gasoline	وطن ع وتر-بنزين	المجتمع الفلسطيني بشكل كومبدي	Microsoft Word	3	606
19	Watn3Water-Takhareef	وطن ع وتر-تخاريف	المجتمع الفلسطيني بشكل كومبدي	Microsoft Word	4	718
20	Watn3Water-Turkish	وطن ع وتر-تركي	المجتمع الفلسطيني بشكل كومبدي	Microsoft Word	4	794

Figure 3.12: Document Table

Tables (Palestinian_Dialect_words, Bigram_Palestinian_Dialect_words, Trigram_Palestinian_Dialect_words, and Fourgram_Palestinian_Dialect_words) are used to store N-gram model; each of these tables has: N-gram data (1, 2, 3, or 4 words), document that N-gram data related to, paragraph that N-gram data located on it, and position of N-gram data in the document. Figures 3.13-3.16 show samples from these tables

Document | Palestinian_Dialect_Words_backup

palestinian_	word	Document_I	startAt_word_position	endsAt_word_position	paragraph_no
1	ما	1	0	2	0
2	هو	1	3	5	0
3	حال	1	6	9	0
4	النت	1	10	14	0
5	عندكن	1	15	20	0
6		1	24	26	0
7	للازكياء	1	27	35	1
8	فقط	1	36	39	1
9	ما	1	40	42	1
10	هو	1	43	45	1
11	الجواب	1	46	52	1
12		1	56	58	1
13	اي	1	59	61	2
14	اكتر	1	62	66	2
15	شي	1	67	69	2
16	بتستخدموا	1	70	79	2
17	لتروحوا	1	80	87	2
18	على	1	88	91	2
19	المدرسة	1	92	99	2
20	او	1	100	102	2
21	الجامعة	1	103	110	2
22	شو	1	117	119	3
23	اسم	1	120	123	3

Figure 3.13: Palestinian Dialect words Table

Bigram_Palestinian_Dialect_Words_backup

palestinian_	word	Document_I	startAt_word_position	endsAt_word_position	paragraph_no
1	ما هو	1	0	5	0
2	هو حال	1	3	9	0
3	حال النت	1	6	14	0
4	النت عندكن	1	10	20	0
5	للازكياء فقط	1	27	39	1
6	فقط ما	1	36	42	1
7	ما هو	1	40	45	1
8	هو الجواب	1	43	52	1
9	اي اكتر	1	59	66	2
10	اكتر شي	1	62	69	2
11	شي بتستخدموا	1	67	79	2
12	بتستخدموا لتروحوا	1	70	87	2
13	لتروحوا على	1	80	91	2
14	على المدرسة	1	88	99	2
15	المدرسة او	1	92	102	2
16	او الجامعة	1	100	110	2
17	شو اسم	1	117	123	3
18	اسم هاد	1	120	127	3
19	هاد السلاح	1	124	134	3
20	السلاح عندكن	1	128	140	3
21	اربع رجال	1	152	161	4
22	رجال فتحوا	1	157	167	4
23	فتحوا سوبر	1	162	172	4

Figure 3.14: Bigram Palestinian Dialect words Table

Trigram_Palestinian_Dialect_Words_backup

palestinian_	word	Document_I	startAt_word_position	endsAt_word_positic	paragraph_no
1	ما هو حال	1	0	9	0
2	هو حال النت	1	3	14	0
3	حال النت عندكن	1	6	20	0
4	للازكياء فقط ما	1	27	42	1
5	فقط ما هو	1	36	45	1
6	ما هو الجواب	1	40	52	1
7	اي اكتر شي	1	59	69	2
8	اكتر شي بتستخدموا	1	62	79	2
9	شي بتستخدموا لتروحوا	1	67	87	2
10	بتستخدموا لتروحوا على	1	70	91	2
11	لتروحوا على المدرسة	1	80	99	2
12	على المدرسة او	1	88	102	2
13	المدرسة او الجامعة	1	92	110	2
14	شو اسم هاد	1	117	127	3
15	اسم هاد السلاح	1	120	134	3
16	هاد السلاح عندكن	1	124	140	3
17	اربع رجال فتحوا	1	152	167	4
18	رجال فتحوا سوبر	1	157	172	4
19	فتحوا سوبر ماركت	1	162	178	4
20	سوبر ماركت وحطوا	1	168	184	4
21	ماركت وحطوا كل	1	173	187	4
22	وحطوا كل دور	1	179	191	4
23	كل دور فيه	1	185	195	4

Figure 3.15: Trigram Palestinian Dialect words Table

Fourgram_Palestinian_Dialect_Words_backup

palestinian_	word	Document_I	startAt_word_position	endsAt_word_position	paragraph_n
1	ما هو حال التنت	1	0	14	0
2	هو حال التنت عندكن	1	3	20	0
3	للازكياء فقط ما هو	1	27	45	1
4	فقط ما هو الجواب	1	36	52	1
5	اي اكتر شي بتستخدموا	1	59	79	2
6	اكتر شي بتستخدموا لتروحوا	1	62	87	2
7	شي بتستخدموا لتروحوا على	1	67	91	2
8	بتستخدموا لتروحوا على المدرسة	1	70	99	2
9	لتروحوا على المدرسة او	1	80	102	2
10	على المدرسة او الجامعة	1	88	110	2
11	شو اسم هاد السلاح	1	117	134	3
12	اسم هاد السلاح عندكن	1	120	140	3
13	اربع رجال فتحوا سوبر	1	152	172	4
14	رجال فتحوا سوبر ماركت	1	157	178	4
15	فتحوا سوبر ماركت وحطوا	1	162	184	4
16	سوبر ماركت وحطوا كل	1	168	187	4
17	ماركت وحطوا كل دور	1	173	191	4
18	وحطوا كل دور فيه	1	179	195	4
19	كل دور فيه كاشير	1	185	201	4
20	دور فيه كاشير دخل	1	188	205	4
21	فيه كاشير دخل زبون	1	192	210	4
22	كاشير دخل زبون بدو	1	196	214	4
23	دخل زبون بدو جاج	1	202	218	4

Figure 3.16: Fourgram Palestinian Dialect words Table

3.2.2 Storing Schema Statistics

Database statistics are shown in Table 3.2

Table Name	**Contents**
Document Table	50 documents
Palestinian_Dialect_words Table	43090 words
Dictionary of Palestinian Vocabularies and Loan words" **معجم**	5595 terms

"العامي و الدخيل	
Bigram_Palestinian_Dialect_words Table	37043 pair of words
Trigram_Palestinian_Dialect_words Table	32022 triple of words
Fourgram_Palestinian_Dialect_words Table	27684 four of words

Table 3.1: Data base tables statistics

Finally, The total Number of the distinct words in the corpus is 25402

Chapter 4

Corpus Annotation

This chapter gives a detailed description about annotation approach followed in this thesis. Section 4.1 presents the annotation specification. Section 4.2 discusses the annotation process that followed. Section 4.3 discusses the importance of using annotation tool.

4.1 Annotation Methodology

Annotation is a process that aims to annotate each word with relevant Meta data; these relevant Meta data are helpful in many applications such as translation, morphological analyses; these Meta data -could be part-of-speech tagging, stem, gloss, equivalent MSA… etc. The annotation will

be more helpful if it is done in a context because the same word may have different analysis and different meanings in different contexts.

So imagine that we have a word with its Metadata such as an equivalent word in MSA then we can use it in application that translate from dialect to MSA, on the other side we can use English meanings in application that translate from Dialect to English. We followed the annotation methodology provided in [20], which requires annotating every word in the corpus with Metadata, such as prefix, suffix, stem, lemma, part-of-speech, gloss and MSA lemma, defined in [20, 17] as the following:

- ***Word:*** it is the row data as it appeared in the input data; and it is represented in Arabic
- ***Word*** **(Buckwalter):** it is the row data as it appeared in the input data; but it is represented in the Buckwalter transliteration [22]. Examples of Buckwalter transliteration: letter "ط" is written as "T" in Buckwalter, letter "ذ" is written as "*", the letter "أ" is written as ">", and letter "ص" is written as "S".
- **Surface (Unicode):** It is written in Arabic characters, and it is reflecting the word written as CODA specification presented in

chapter 2, and there is a work in progress to extending it to cover Palestinian Dialect; So As we stated in [17] Palestinian dialect does not have a standard orthography and we can't also use the MSA orthography because there are many differences in phonological, morphological and lexical. Also, Palestinian people write the dialect in different ways that reflect the differences in phonology; for example the word "دقيقة" -might be written in four ways, such as "دقيقة","دئيئة","دكيكة","دغيغة". Similarly, if a word has many long vowels, then it may be written in different ways, for example the word "مساكين" maybe also written as "مسكين"by shorting the first vowel. PAL also has some qualities that do not exist in MSA, which may be written in different ways, such as the Palestinian future particle ح which is written attached to the verb that followed it such as "حاروح" or separate from the verb that follows it such as "ح اروح". Finally, there are words in the Palestinian dialect may be written in different forms such as "برضه","برضو", and "برضة".

All of these problems can be covered by applying the guidelines of EGY version of CODA, but there are unique Palestinian problems that we need to be deal with, such as: adding the letter "ك" to the list of letters that spelling in different forms in dialect because it may be spelling as "ك/k" or as "/tš/, adding the non-EGY to the list of clitics such asdemonstrative proclitic ه+h ; e.g.: هالشغل ، هالحياة ، هالبنت and Conjunction proclitic ت+t 'so as to', e.g., تيشوف ، لتروحوا،تيوكل, and Finally extending the list of exceptional words to cover additional Palestinian words; an example of these words is word "هيو" which corresponding to "ها هو" in Standard Arabic, Table 4.1 provides a sample from it

	CODA	Non-CODA Variants	English
Demonstrative Pronouns	هاذا	هاظا-هادا	this, that [3ms]
	هاذي	هاظي-هادي	this, that [3fs]
	هذول	هدول-هظول	these, those [3p]
	هذولاك	هدولاك-هظولاك	these, those [3p]
	هيو	هذهو-هظهو-هيوتو-هيتا	There he, it is

Table 4.1: Sample of Exceptional Palestinian words [17]

- **Surface (Buckwalter):** The Buckwalter transliteration of the Surface.

- **Lemma:** The lemma is the dictionary entry that abstracts all inflectional morphology. The lemma for the verbs is the third person masculine singular perfective form and for nouns is the masculine singular form if available or feminine singular form if masculine is not available. The lemma of the word is written in Buckwalter transliteration. E.g.: for the word "یرى" the lemma is "رأى", for the word "أولاد" the lemma is "ولد", for the word "سيارات" the lemma is "سيارة", and for the word "بيجي" the lemma is "أجا"
- **Buckwalter POS:** it is the fullest part-of-speech (POS) of the word; it combines prefixes, stem, and suffixes of a word with their tags. Tags that used here are the tags that define in[21], these tags are classified to many groups such as Nouns which have basic tags as for example NOUN and NOUN_PROP and also have related tags that come as prefix e.g. "DET" for "ال" or as suffixes e.g."NSUFF_FEM_PL" for "ات"(جمع المؤنث) , another group is Pronouns group which have many tags(PRON-1S e.g. "أنا" ,DEM_PRON e.g. "ه"(هالشغل),REL_PRON e.g. "اللي", pronouns that reflect objects of the verb e.g. PVSUFF_DO: 1S for "ي" in

“عزمني”, pronouns that reflect subjects of the verb e.g. IVSUFF_SUBJ: 1P for “نا” in “حيشوفنا”…etc), another group is group for adjectives which have tags(ADJ,ADJ_COMP), verbs group which have tags (IV for imperfective verb”فعل مضارع”, PV for perfective verb “فعل ماضي” ,CV for command verb “ فعل أمر”…etc), Adverbs group that have tags(ADV, REL_ADV), particles group that has tags(NEG_PART for negations e.g. "ما","ش”, PROG_PART for progressive part in imperfective verbs e.g. “ب” in “بيحكي”…etc), and Finally there is a tag “PREP” for prepositions as ,"من"CONJ for conjunctions such as “و”, INTERJ for interjections such as "اوكي", PSEUDO_VERB such as “ياريت”, TYPO such as “اخص”, and VERB such as “حاشا”.

Examples of pB:

- صفحة→ SfH/NOUN+ p/NSUFF_FEM_SG
- أتجوز→ A/IV1S+ tjwz/IV
- يما→ ymA/NOUN
- بديش→ bd/IV+ y/IVSUFF_SUBJ:1S+$/NEG_PART

- ***MSA lemma:*** the equivalent lemma in Modern Standard Arabic (MSA). e.g.: for lemma "أجا" in the Palestinian dialect; the corresponding lemma in MSA is "جاء", for lemma "بد" the corresponding lemma in MSA is "أراد".
- **Gloss** it is the corresponding meaning of the lemma in English.
- **Annotator** A specification of the source of the annotation; e.g.: diwan_approved meaning that annotation approved by annotator using DIWAN tool, MADA meaning that annotated automatically By MADAMIRA

word_id	word	word_E	Prefix	Stem	Suffix	Surface	lex	BW POS	MSA	golss	ANNO
1	صفحة	SfHp		SfH/NOUN	p/NSUFF_FEM_SG	SfHp	SafoHap_1	bw:+SfH/NOUN+p/NSUFF_FEM_SG	SafoHap_1	page;leaf	diwan_approved
2	يما	ymA		ymA/NOUN		ymA	yamA_1	bw:+ymA/NOUN+	>my_1	my	diwan_approved
3	بديش	bdy$	(null)/IV1S	bd/IV	y/IVSUFF_SUBJ:1S+$/NEG_PART	bdy$	bd_1	bw:(null)/IV1S+bd/IV+y/IVSUFF_SUBJ:1S+$/NEG_PART	>arAd_1	want;desire;intend	diwan_approved
6	هو	hw		hw/PRON_3MS		hw	huwa_1	bw:+hw/PRON_3MS+	huwa_1	it/he	diwan_approved
7	حال	HAl		HAl/NOUN		HAl	HAl_1	bw:+HAl/NOUN+	HAl_1	situation;condition;	diwan_approved
8	النت	Alnt	Al/DET	nt/NOUN_PROP		Alnt	Alnit*_1	bw:Al/DET+nt/NOUN_PROP+	Alnit*_1	internet_[CALIMA]	diwan_approved
41	عندكن	Endkn		End/NOUN	kn/POSS_PRON_2P	Endkn	Einod_1	bw:+End/NOUN+kn/POSS_PRON_2P	Einod_1	with/at_[SAMA]	diwan_approved

Table 4.2: Annotation Specification Example

One of the decisions that was taken in the annotation methodology [20] is to discard diacritizion when write Buckwalter POS, this decision was taken in order to minimize the load on the human annotator, and also because EGY and MSA generates morpheme and lexical items that differ

from PAL only in the short vowels. Finally, in the methodology considers diacritized Lemma only because it has a critical role in generating suitable gloss.

4.2 Annotation process

4.2.1 Using MADAMIRA

Now, after collecting our corpus, and specifying our annotation specification, the question up to mind is: how can we annotate our corpus? shall we annotate it manually or by using certain tool then annotate it automatically? In order to answer these questions we decided to make an experiment and then according to the experiment's results we decided to use DIWAN tool to annotate our corpus which internally annotate the words using MADAMIRA which presented in chapter 2. This experiment starts by choosing randomly an episode of the PAL TV show “Watan Aa Watar” (460 words), then entering it through both MADAMIRA-MSA and MADAMIRA-EGY to analyze it, after that we analyzed the output that we get from both systems to determine if it's

usable to annotate our corpus or not. Tables 4.3 and 4.4 show samples from experiment results for both EGY and MSA, and Table 4.5 shows experimental results

	A	B	C	D	E	F	G
1	Word	Diac	lex	pos	bw	gloss	Status
2	bdhA-بدها	:bud~hA-بُدِّهَا	:bud~_1-بُدّ	:noun	:bud~/NOUN+hA/POSS_PRON_3FS	:escape;avoiding_[SAMA]	Wrong
3	$HAl-تحال						No ANALYSIS
4	m$ynA-مشينا	:ma$~aynA-مَشَّيْنَا	:ma$~aY_1-مَشَّى	:verb	:ma$~/PV+aynA/PVSUFF_SUBJ:1P	:make_walk;adjust;be_made_to_walk;be_adjusted_[CALIMA]	Correct
5	nAs-ناس	:nAs-ناس	:nAs_1-ناس	:noun	:nAs/NOUN	:people_[CALIMA]	Correct
6	qd-قد	:qado-قَدّ	:qado_1-قَدّ	:part_verb	:qado/VERB_PART	:may,_might,_perhaps,_maybe,_possibly,_probably_[SAMA]	Wrong
7	ybArk-يبارك	:yibArik-يِبارِك	:bArik_1-بارِك	:verb	:yi/IV3MS+bArik/IV	:congratulate;bless_[CALIMA]	Correct
8	kyf-كيف	:kayof-كَيْف	:kayof_1-كَيْف	:adv_interrog	:kayof/INTERROG_ADV	:how;what_degree_[CALIMA]	Correct
9	djAj-دجاج	:dajAj-دَجاج	:dajAj_1-دَجاج	:noun	:dajAj/NOUN	:chickens;poultry;fowl_[SAMA]	Correct
10	AlHmAr-الحمار	:AlHumAr-الحُمار	:HumAr_1-حُمار	:noun	:Al/DET+HumAr/NOUN	:donkey;donkeys;jackass;fol;stupid;dolt;lever;vulgarity_[CALIMA]	Correct
11	Axdt-اخدت	:>axadit-أَخَذِتَ	:Aaxad_1-أَخَذ	:verb	:>axad/PV+it/PVSUFF_SUBJ:3FS	:take;begin;be_taken;convey;treat;chose;receive_and_gain;marry_[CALIMA]	Correct
12	rqm-رقم	:raqam-رَقَم	:raqam_1-رَقَم	:noun	:raqam/NOUN	:number_[CALIMA]	Correct
13	wnS-ونص	:winuS~-وِنُصّ	:nuS~_1-نُصّ	:noun_quant	:wi/CONJ+nuS~/NOUN_QUANT	:half_[CALIMA]	Correct
14	AlAkl-الاكل	:Al>akol-الأُكُل	:>akol_1-أُكُل	:noun	:Al/DET+>akol/NOUN	:eating;consumption;food;meal_[CALIMA]	Correct
15	ltngng-لتنغنغ						No ANALYSIS

Table 4.3: MADAMIRA EGY Result

	A	B	C	D	E	F
1	Word	Diac	lex	bw	gloss	Status
2	bdhA-بدها	:bud~ihA-بُدِّهَا	:bud~_1-بُدّ	:bud~/NOUN+i/CASE_DEF_GEN+hA/POSS_PRON_3FS	:escape;avoiding	Wrong
3	$HAl-تحال	:$HAl-تَحال	:$HAl-تحال			No ANALYSIS
4	m$ynA-مشينا	:ma$oyanA-مَشَيْنَا	:ma$oy_1-مَشَى	:ma$oy/NOUN+a/CASE_DEF_ACC+nA/POSS_PRON_1P	:going;walking	Wrong
5	nAs-ناس	:nAs-ناس	:nAs_1-ناس	:nAs/NOUN	:people	Correct
6	qd-قد	:qado-قَدّ	:qado_1-قَدّ	:qado/VERB_PART	:may,_might,_perhaps,_maybe,_possibly,_probably	Wrong
7	ybArk-يبارك	:yubAriku-يُبارِكُ	:bArak_1-بارَك	:yu/IV3MS+bArik/IV+u/IVSUFF_MOOD:I	:bless;approve;congratulate	Correct
8	kyf-كيف	:kayofa-كَيْفَ	:kayofa_2-كَيْفَ	:kayofa/REL_ADV	:how	Correct
9	djAj-دجاج	:dajAjK-دَجاجٍ	:dajAj_1-دَجاج	:dajAj/NOUN+K/CASE_INDEF_GEN	:chickens;poultry;fowl	Correct
10	AlHmAr-الحمار	:AlHimAru-الحِمارُ	:HimAr_1-حِمار	:Al/DET+HimAr/NOUN+u/CASE_DEF_NOM	:donkey	Correct
11	Axdt-اخدت	:Axdt-اخدتَ	:Axdt-اخدت			No ANALYSIS
12	rqm-رقم	:raqomu-رَقْمُ	:raqom_1-رَقْم	:raqom/NOUN+u/CASE_DEF_NOM	:number;numeral	Correct
13	wnS-ونص	:wanaS~a-وَنَصٌّ	:naS~-u_1-نَصّ	:wa/CONJ+naS~/PV+a/PVSUFF_SUBJ:3MS	:stipulate;specify	Correct
14	AlAkl-الاكل	:Al>akola-الأَكْلَ	:>akol_1-أَكْل	:Al/DET+>akol/NOUN+a/CASE_DEF_ACC	:eating;consumption	Correct
15	ltngng-لتنغنغ	:ltngng-لتنغنغ	:ltngng-لتنغنغ			No ANALYSIS

Table 4.4: MADAMIRA MSA Result

	EGY		MSA	
	Number	Percentage	Number	Percentage
NO_Analysis	43	9.32%	82	17.78%
wrong _Analysis	80	17.35%	91	19.70%
correct _Analysis	338	73.31%	288	62.47%

Table 4.5: Experiment result

As we noticed in Table 4.5, Egyptian dialect looks close to Palestinian dialect; this is because we focus here on written data not in oral data. Palestinian and Egyptian people write, for example, "قلتلك" in the same way, although both pronounce it differently.

The last column "Status" in tables 4.3, and 4.4 isn't generated automatically by MADAMIRA, we added this column manually according to analysis that generated by MADAMIRA; we gave "No ANALYSIS" to words that MADAMIRA fails on it, also we gave "Wrong" to words that MADAMIRA returned wrong analysis in somewhere on it such as gloss, POS, lemma, and finally we gave "Correct" to words that MADAMIRA returned correct analysis for it in all parts(correct gloss, correct POS, correct lemma).

Words with status No Analysis in Tables 4.3, and 4.4 refer to the words that the morphological analyzer couldn't analysis it. There are many reasons that cause failure in analysis in MADAMIRA such as the word is totally Palestinian (it is used in the Palestinian dialect only), e.g. word in entry #2 "شحال", and word in entry #15 "لتتغنغ" in Table 4.3.

While the wrongly analyzed words are words MADAMIRA gives it incorrect part-of-speech (POS) or incorrect lemma or incorrect gloss. An example of wrong analysis is the word in entry#1 "بدها", which MADAMIRA-EGY analyze it as noun while it is a verb and it is a totally Palestinian word. In general, the results show that we can use MADAMIRA-EGY to annotate our corpus because almost it gives correct analysis or give analysis that need some modifications.

4.4.2 Manual Annotation

After annotating the corpus using MADAMIRA-EGY the total Number of words annotated By MADAMIRA-EGY is 55586 (16,334 unique); number of words that MADAMIRA_EGY returned NO_ANALYSIS is 1689 word from 55586; these words which equivalent to 1244 unique words are annotated manually. Figure 4.1 shows sample from MADAMIRA result

word_id	diac	lex	pos	bw	gloss
1	SafoHap	SafoHap_1	noun	SafoH/NOUN+ap/NSUFF_FEM_SG	page;leaf;pages;leaves_[CALIMA]
2	yam~aA	yam~ap_1	noun	yam~/NOUN+ap/NSUFF_FEM_SG	side;part;direction;sectors;offices;institutions;officials;inc
3	bad~aY$	bad~aY_1	verb	bad~aY/PV+(null)/PVSUFF_SUBJ:3MS+$/NEG_PART	TBA_[CALIMA]
4	Aatojaw~iz	Aitojaw~iz_1	verb	Aa/IV1S+tojaw~iz/IV	marry;get_married_[CALIMA]
5	mA	mA_1	part_focus	mA/FOCUS_PART	TBA_[CALIMA]
6	huw~a	huw~a_1	pron	huw~a/PRON_3MS	it;he_[CALIMA]
7	HAl	HAl_1	noun	HAl/NOUN	situation;condition;case;situations;conditions;cases_[CALI
8	Aln~it~	nit~_1	noun	Al/DET+nit~/NOUN	immediately;at_once;immediately_after;boiling;boil;right
9	Einodkun~a	Einod_1	noun	Einod/NOUN+kun~a/POSS_PRON_2FP	with/at_[SAMA]
10	?	?_0	punc	?/PUNC	?
58	kA$yr	kA$yr_0	NOUN_NUM	NO_ANALYSIS	noun_num
41226	$HAl	$HAl_0	ADJ_COMP	NO_ANALYSIS	adj_comp
11	?	?_0	punc	?/PUNC	?
12	?	?_0	punc	?/PUNC	?
13	?	?_0	punc	?/PUNC	?
14	lil>azokiyA'	zakiy~_1	noun	li/PREP+Al/DET+>azokiyA'/NOUN	pure;blameless_[SAMA]
15	faqaT	faqaT_1	adv	faqaT/ADV	only_[CALIMA]
16	mA	mA_1	conj_sub	mA/SUB_CONJ	that;if;unless;whether;as_long_as;as_soon_as_[CALIMA]
17	huw~a	huw~a_1	pron	huw~a/PRON_3MS	it;he_[CALIMA]
18	AljawAb	jawAb_1	noun	Al/DET+jawAb/NOUN	answer;answers;letter;letters_[CALIMA]

Figure 4.1: Sample of MADAMIRA_EGY Result

Figure 4.2 shows a sample from words that annotated manually

WORD_BW	WORD_AR	LEMMA__AR	LEMMA_BW	POS	CODA	POS BW
syArtyn	سيارتين	سيارة	sayArap_1	noun	سيارتين	sayArat/NOUN+iyn/NSUFF_MASC_DU
$ykt	شيكت	شيك	$~ayak_1	verb	شيكت	$~ayak/PV+t/PVSUFF_SUBJ:1S
*bHtwnA	ذبحتونا	ذبح	*abaH_1	verb	ذبحتونا	*abaH/PV+tw/PVSUFF_SUBJ:2P+nA/PVSUFF_DO:1P
>bEt*r	أبعتذر	اعتذر	AiEota*ar_1	verb	باعتذر	b/PROG_PART+A/IV1S+Eota*ir/IV
>bqdr$	أبقدرش	قدر	qidir_1	verb	باقدرش	b/PROG_PART+A/IV1S+qadar/IV+$/NEG_PART
>fTrtwA	أفطرتوا	أفطر	>afoTar_1	verb	أفطرتوا	>afoTar/PV+twA/PVSUFF_SUBJ:2P
>gAnykm	أغانيكم	أغنية	>ugoniyap_1	noun	أغانيكم	>agAniy/NOUN+kum/POSS_PRON_2P
>kbr	أكبر	أكبر	>akobar_1	adj_comp	أكبر	>akobarr/ADJ_COMP
>lgAz	ألغاز	لغز	luguz_1	noun	ألغاز	>alogaz/NOUN
>sbtlkm	أسبتلكم	أثبت	>avobat_1	verb	أثبت لكم	>avobit/PV+l/PREP+kum/PRON_2P
>wrdgAn	أوردغان	أوردغان	>wrdgAn_1	noun_prop	أوردغان	>wrdgAn/NOUN_PROP
>wh	أوه	أوه	>wh_1	interj	اوه	>wh/INTERJ
A$fTy	اتشفطي	شفط	$afaT_1	verb	اتشفطي	Au$ofuT/CV+y/CVSUFF_SUBJ:2FS
A$krwA	اتشكروا	شكر	$akar_1	verb	اتشكروا	Ai$okur/CV+wA/CVSUFF_SUBJ:2P
A$rHlwA	اترحلوا	شرح	$araH_1	verb	اترح له	A/IV1S+$oraH/IV+l/PREP+uh/PRON_3MS
AHsn	احسن	احسن	AaHosan_1	adj_comp	احسن	AaHosan/ADJ_COMP
ATlE	اطلع	طلع	TiliE_1	verb	اطلع	A/IV1S+TolaE/IV
AlY	الى	إلى	<ilA_1	prep	إلى	<ilA/PREP
AETytwny	اعطيتوني	اعطى	>aEoTY_1	verb	اعطيتوني	>aEoTy/PV+tw/PVSUFF_SUBJ:2P+niy/PVSUFF_DO:1S
AHbhA	احبها	حب	H~ab_1	verb	أحبها	A/IV1S+Hib/IV+hA/IVSUFF_DO3:MS
ASly	اصلي	أصلي	>aSoliy_1	noun	أصلي	>aSoliy/NOUN

Figure 4.2: Sample of Manual Annotated List

4.2.3 Correct errors and fill in gaps using DIWAN

After we annotated NO_ANALYSIS words manually; we started with the next step which is to make a double check over other words that

annotated by MADAMIRA_EGY to verify it if it is analyzed correctly or not, then to correct words that are annotated wrongly. A decision was made to do this step using DIWAN tool. DIWAN is a tool developed at Columbia University for text annotating. DIWAN takes a text as input; in this case a text is a MADAMIRA_EGY output file, then DIWAN shows every word with its annotation for annotator; annotator may choose to save it as is or to make some updates on it. Figure 4.3 shows DIWAN Interface.

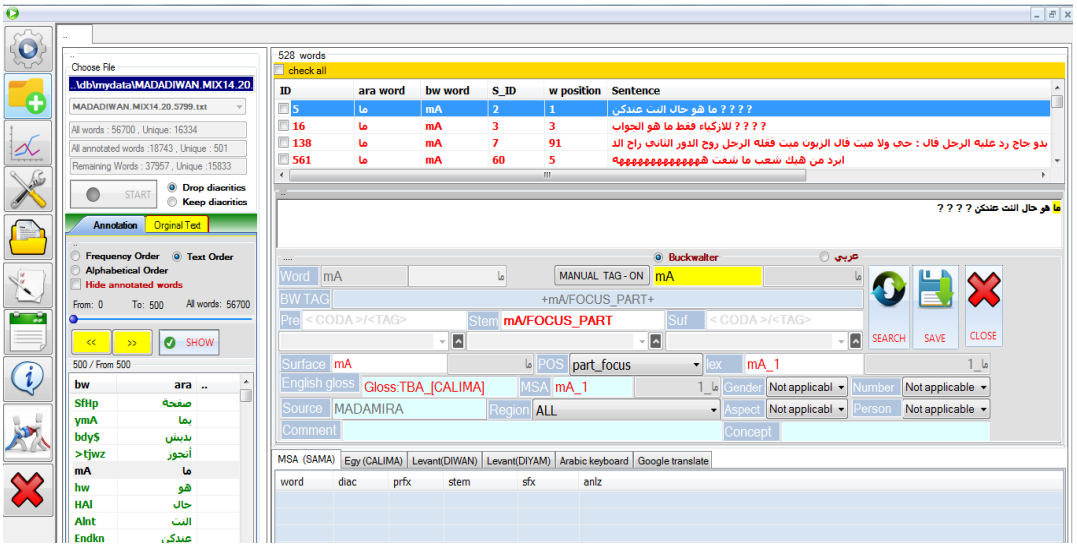


Figure 4.3: DIWAN Interface

So as shown in Figure 4.3, annotator chooses the word, then clicks on it, then check the annotation appeared and either save as it or make some changes on it; most of the time the changes are occurring in POS,

English Gloss, and MSA. Output saved in files that looks like MADAMIRA_EGY file with two extra attributes which are: Source mod to indicate if the annotator updated it or not (No: annotator does not update it, Yes: annotator update it), ANNO which has value diwan_approved to state that annotation is verified By DIWAN. In This Thesis the number of words that double checked By DIWAN are 18743(501 unique words). Figure 4.4 shows sample from DIWAN output

word_id	word	word_bw	Surface	lex	POS Bw	MSA lemma	gloss	pos	S_M	anno
673	الله	Allh	Allh	All~h_1	+Allh/NOUN_PROP+	All~h_1	Allah;God	noun_prop	No	diwan_approved
30071	عماد	EmAd	EmAd	EimAd_1	+EmAd/NOUN_PROP+	EimAd_1	Imad	noun_prop	Yes	diwan_approved
10124	يما	ymA	ymA	ymA_1	+ymA/NOUN+	>um~y_1	my mother	noun	Yes	diwan_approved
4	أتجوز	>tjwz	Atjwz	Aitojaw~iz_1	A/IV1S+tjwz/IV+	Aitozwj_1	marry;get_married_[CALIMA]	verb	No	diwan_approved
6	هو	hw	hw	huwa_1	+hw/PRON_3MS+	huwa_1	it/he_[SAMA]	pron	No	diwan_approved
2843	?	?	?	?_0	+?/PUNC+	?_0	?	punc	No	diwan_approved
38402	هذ	hAd	hA*A	hA*A_1	+h*A/DEM_PRON_MS+	h`*A_1	this_[masc.sg.]	pron_dem	Yes	diwan_approved
31	الجامعة	AljAmEp	AljAmEp	jAmoEap_1	Al/DET+jAmE/NOUN+p/NSUFF_FEM_SG	jAmoEap_1	universities;university_[CALIMA]	noun	No	diwan_approved
29	المدرسة	Almdrsp	Almdrsp	madorasap_1	Al/DET+mdrs/NOUN+p/NSUFF_FEM_SG	madorasap_1	school;schools_[CALIMA]	noun	No	diwan_approved
492	الصفحات	AlSfHAt	AlSfHAt	SafoHap_1	Al/DET+SfH/NOUN+At/NSUFF_FEM_PL	SafoHap_1	pages;leaves_[SAMA]	noun	No	diwan_approved
491	تقلب	tqlb	tqlb	qalab_1	t/IV2MS+qlb/IV+	qalab_1	turn_around;reverse;overthrow;topple_[CALIMA]	verb	Yes	diwan_approved
2606	تقدر	tqdr	tqdr	qidir_1	t/IV2MS+qdr/IV+	AstTAE_1	able_to;be_able;be_able_to;be_capable_of;mak	verb	Yes	diwan_approved
26584	جهاز	jhAz	jhAz	jihAz_1	+jhAz/NOUN+	jihAz_1	machine;apparatus;device;appliance;system;equ	noun	No	diwan_approved
486	بالمكعبات	bAlmkEbAt	bAlmkEbAt	mukaE~ab_2	b/PREP+Al/DET+mkEb/NOUN+At/NSUFF_FEM_PL	mukaE~ab_2	cube;cubiform_[SAMA]	noun	No	diwan_approved
485	يلعب	blEb	bylEb	liEib_1	b/PROG_PART+y/IV3MS+lEb/IV+	liEib_1	play_[CALIMA]	verb	No	diwan_approved
32369	حالو	HAlw	HAlh	HAl_1	+HAl/NOUN+h/POSS_PRON_3MS	nafos_1	self	noun	Yes	diwan_approved
37803	اسبوع	AsbwE	AsbwE	AusobuwE_1	+AsbwE/NOUN+	AusobuwE_1	week;weeks_[CALIMA]	noun	No	diwan_approved
480	صرلو	Srlw	Sar lh	SAr_1	+SAr/PV+(null)/PVSUFF_SUBJ:3MS+l/PREP+h/PRON_3MS	SAr_1	become;begin_to	verb	Yes	diwan_approved
479	الصيدلي	AlSydly	AlSydly	Sayodaliy~_2	Al/DET+Sydly/ADJ+	Sayodaliy~_2	pharmaceutical_[SAMA]	adj	No	diwan_approved
478	طبعا	TbEAA	TbEA	TaboE_1	+TbE/NOUN+A/CASE_INDEF_ACC	TaboE_1	naturally;of_course	noun	No	diwan_approved
19318	نسمع	nsmE	nsmE	simiE_1	n/IV1P+smE/IV+	simiE_1	listen;be_heard;hear_[CALIMA]	verb	No	diwan_approved
2404	خلينا	xlynA	xlynA	xal~aY_1	+xly/CV+nA/CVSUFF_DO:1P	xal~aY_1	;allow;be_allowed_[CALIMA]	verb	Yes	diwan_approved
525	طب	Tb	Tyb	Tayb_1	+Tyb/INTERJ+	Tayb_1	TBA_[CALIMA]	interj	Yes	diwan_approved
472	الهمشش	AlmHSS	AlmHSS	mHSS_1	Al/DET+mHSS/NOUN+	muHoSaS_1	hashish_addict_[SAMA]	noun	Yes	diwan_approved

Figure 4.4: DIWAN Output Sample

4.3 Discussion

As we noticed previously, using DIWAN to do a double check on MADAMIRA annotations will increase the productivity in building annotated corpus, also using DIWAN will preserve the quality of the

annotated corpus because it minimizes the entries that users will enter it manually. So if we suppose that there is no tool that helps us in annotating our corpus then the annotation process will take too long time and maybe also cause a lot of errors in annotation. From our experience in manual annotation step, annotating of NO_ANALYSIS words (1244 words) takes about two months, which is too long time and also many errors appear in first iteration then in the second iteration. We minimize these errors and finally we make a double check on it using DIWAN. So if we don't use DIWAN then annotator will be ask to enter all things manually and be very careful when entering tags in pB part, because any error will cause a problem, and as we know there are a lot of tags, all of these tags DIWAN presents it as auto-complete list, also gender, number, aspect and person are presented as list which will also minimize error because if annotator enters it then sometimes will enter gender as female, male, another time as f, and m, and also may enter person sometimes as one, second, third, another time as 1,2, and 3, and may enter number as singular, plural, Dual or as s, p, and d, so all of these errors will cause a real problem because we need it in standard format to make our corpus

applicable in another applications, So using DIWAN will save time and quality.

till now we annotate about 500 unique words which equivalent to 18743 non unique words; which means that we annotated all duplicates of 500 unique words; as an example the word "بابا" is repeated about 26 times in the corpus in a different context, also these 500 unique words are the most frequent words in our corpus which means that some of them are duplicated more than 500 times in our corpus. Finally, every time we use DIWAN we accelerate process more and more.

4.4 Evaluation

To evaluate the quality of annotations, we consider the inter-annotator agreement. We chose a sample of 59 words from our corpus that we annotated, and asked a colleague (Faeq Rimawi, from Sina Institute) to annotate this sample separately, then we compared the annotations.

As we notice in Table 4.6, the percentage of differences in annotations regardless of cause of difference is 20.3% which comes from differences in MSA lemma, lemma, and from Buckwalter POS with POS, also

sometimes the two annotator gives two different values for many attributes in the same word such as different values for both POS and Buckwalter POS, or for MSA lemma and gloss.

	Number of words	**Percentage**
Difference in total	12 word	20.30%
Difference in Lemma	2 word	3.30%
Difference in MSA Lemma	4 word	6.70%
Difference in gloss	4 word	6.70%
Difference in BW, POS	6 word	10.10%

Table 4.6: Inter-annotator agreement result

Figure 4.5 shows a sample for differences between two annotators; this sample explained the number of words in Table 4.6. For example, the two annotators gave the word " دشر" different values for both MSA lemma, and gloss, so we considered it in the differences of both.

word	word bw	surface	lemma	bw	MSA lemma	gloss	pos	Anno	Status
الشيطان	Al$yTAn	Al$yTAn	lex:$iyTAn_1	Al/DET+$yTAn/NOUN+	$iyTAn_1	devil;she-devil;shrew;devils;demons;satan;devilish_[CALIMA]	noun	F	Same
الشيطان	Al$yTAn	Al$yTAn	lex:$iyTAn_1	Al/DET+$yTAn/NOUN+	$iyTAn_1	devil;she-devil;shrew;devils;demons;satan;devilish_[CALIMA]	noun	F	Same
الشيطان	Al$yTAn	Al$yTAn	lex:$iyTAn_1	Al/DET+$yTAn/NOUN+	$iyTAn_1	devil;she-devil;shrew;devils;demons;satan;devilish_[CALIMA]	noun	D	Same
الشيطان	Al$yTAn	Al$yTAn	lex:$iyTAn_1	Al/DET+$yTAn/NOUN+	$iyTAn_1	devil;she-devil;shrew;devils;demons;satan;devilish_[CALIMA]	noun	D	Same
وجاي	wjAy	wjAy	lex:jAy_1	w/CONJ+jAy/ADJ+	jAy_1	coming;comming_[CALIMA]	adj	F	Differ in MSA Lemma
وجاي	wjAy	wjAy	lex:jAy_1	w/CONJ+jAy/ADJ+	jAy_1	coming;comming_[CALIMA]	adj	F	Differ in MSA Lemma
وجاي	wjAy	wjAy	lex:jAy_1	w/CONJ+jAy/ADJ+	qAdm_1	coming;comming_[CALIMA]	adj	D	Differ in MSA Lemma
وجاي	wjAy	wjAy	lex:jAy_1	w/CONJ+jAy/ADJ+	qAdm_1	coming;comming_[CALIMA]	adj	D	Differ in MSA Lemma
ياكل	yAkl	yAkl	lex:>akal-u_1	y/IV3MS+Akl/IV	>akal-u_1	eat;consume_[SAMA]	verb	F	Differ in Surface
ياكل	yAkl	yAkl	lex:>akal-u_1	y/IV3MS+Akl/IV	>akal-u_1	eat;consume_[SAMA]	verb	F	Differ in Surface
ياكل	yAkl	y>kl	lex:>akal-u_1	y/IV3MS+>kl/IV+	>akal-u_1	eat;consume	verb	D	Differ in Surface
ياكل	yAkl	y>kl	lex:>akal-u_1	y/IV3MS+>kl/IV+	>akal-u_1	eat;consume	verb	D	Differ in Surface
بالعربي	bAlErby	bAlErby	lex:Earabiy_1	b/PREP+Al/DET+Erby/ADJ+	Earabiy_1	Arabic	adj	F	Differ in POS, BW
بالعربي	bAlErby	bAlErby	lex:Earabiy_1	b/PREP+Al/DET+Erby/ADJ+	Earabiy_1	Arabic	adj	F	Differ in POS, BW
بالعربي	bAlErby	bAlErby	lex:Earabiy_1	b/PREP+Al/DET+Erby/ADJ+	Earabiy_1	Arabic	adj	F	Differ in POS, BW
بالعربي	bAlErby	bAlErby	lex:Earabiy_1	b/PREP+Al/DET+Erby/NOUN+	Erbyp_1	arabs;arab;Arabic_[CALIMA]	noun	D	Differ in POS, BW
بالعربي	bAlErby	bAlErby	lex:Earabiy_1	b/PREP+Al/DET+Erby/NOUN+	Erbyp_1	arabs;arab;Arabic_[CALIMA]	noun	D	Differ in POS, BW
بالعربي	bAlErby	bAlErby	lex:Earabiy_1	b/PREP+Al/DET+Erby/NOUN+	Erbyp_1	arabs;arab;Arabic_[CALIMA]	noun	D	Differ in POS, BW
بشر	d$r	d$r	lex:d$r_0	+d$r/CV+	>utrk_0	leave;unhand	verb	F	Differ in Gloss , MSA lemma
بشر	d$r	d$r	lex:da$ar_1	+d$r/CV+(null)/CVSUFF_SUBJ:2MS	tarak_1	leave_[CALIMA]	verb	D	Differ in Gloss , MSA lemma

Figure 4.5 : annotators differences Sample

Chapter 5

Conclusion and Future Work

This Thesis presented a sample from our annotated corpus; it presented 500 words that annotated very well according to our annotation methodology. It discussed the linguistic variations between Palestinian dialect and Modern Standard Arabic especially in terms of morphology, orthography, and lexicon. It also discussed our annotation methodology, the benefits of using MADAMIRA-EGY, and annotation tool DIWAN, to semi-automate and speed up the annotation process.

This Thesis raised many issues that need to be addressed in the future work and researches such as; Complete using DIWAN tool to approve MADAMIRA_EGY annotations by annotator; the result of this step is fully annotated PAL corpus ; this corpus contains 16334 unique words annotated with relevant meta data such as (Surface, Lemma, POS, Buckwalter POS, MSA lemma, English gloss), Complete the development of Palestinian-specific morphological annotation tags and

CODA guidelines, build A Palestinian lexicon, which will be extracted from surface, lemma, MSA lemma, and English gloss attributes in corpus, extend MADAMIRA to analyze Palestinian text by learning MADAMIRA_EGY with DIWAN output data, Corpus will be extended to include more text; these texts will be collected from different resources such as the new parts(2014) from Palestinian TV show "وطن ع وتر", and from TV Shows "فنجان البلد", "حروف وطن", and the famous Palestinian series "التغريبة الفلسطينية", also from others Palestinian Facebook and Twitter pages and also collecting more text from Palestinian forums and Finally by collecting poems that written in Palestinian dialect such as the collection of poems"ميجنا" for the Palestinian poet "تميم البرغوثي".

Finally, all lexical annotations for our corpus such as Lemma will be linked with Arabic ontology resources. The corpus will finally be published on public for researchers.

References

[1] Rambow, O., Chiang, D., Diab, M., Habash, N., Hwa, R., Sima'an, K., Lacey, V., Levy, R., Nichols, C., and Shareef, S. 2006. Parsing Arabic Dialects. Final Report-Version 1, January 18, 2006.

[2] Maamouri, M., Bies, A., Buckwalter, T., Diab, M., Habash, N., Rambow, O., and Tabessi, D. 2006. Developing and using a Pilot Dialectal Arabic Treebank. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06

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

[4] Benajiba, Y., and Diab, M. 2010.A Web Application for Dialectal Arabic Text Annotation. In Proceedings of the LREC Workshop for Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages: Status, Updates, and Prospects. 2010

[5] Diab, M., Habash, N., Rambow, O., Altantawy, M., and Benajiba, Y. 2010. COLABA: Arabic dialect annotation and processing. In Proceedings of LREC Workshop on Semitic Language Processing. 2010. p. 66-74

[6] Diab, M., and Habash, N.: Presentation about Arabic Dialect Processing. MEDAR 2009, Cairo, Egypt, April 21, 2009

[7] Lubany, A.: Dictionary of Palestinian Vocabularies and Loan Words (Arabic-Arabic) معجم العامي و الدخيل في فلسطين عربي –عربي.2006.Lebanon Library.

[8] Habash, N., Diab, M., and Rambow, O. Conventional Orthography for Dialectal Arabic: Principles and Guidelines –Egyptian Arabic.2012. Columbia university.

[9] Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Belguith, L., and Habash, N. Conventional Orthography for Tunisian Arabic.2014.ANLP Research group, MIRACL Lab, University of Sfax, Tunisia.

[10] Habash, N., Rambow, O., and Kiraz, G. 2005. Morohological analysis and generation for Arabic Dialects. Semitic '05 Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages. Pages 17-24

[11] Habash, N., and Rambow, O. 2006. MAGEAD: Morphological Analyzer and Generator for Arabic Dialects. In proceedings of the 21st International conference on Computational linguistics and 44th Annual Meeting of ACL, pages 681-688, Sydney, July 2006.

[12] Altantawy, M., Habash, N., Rambow, O., and Saleh, I. 2010. Morphological Analysis and Generation of Arabic Nouns: A Morphemic Functional Approach. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta; 01/2010.

[13] Almeman, K., and Lee, M. 2012.Towards Developing a Multi-Dialect Morphological Analyzer for Arabic. In Proceedings of 4th International Conference on Arabic Language Processing, May 2-3, 2012, Rabat, Morocco.

[14] Habash, N., Rambow, O., Diab, M., and Kanjawi-Faraj, R. 2008.Guidelines for Annotation of Arabic Dialectness. In Proceedings of workshop on Arabic and its local Languages,Marrakech,Mococco,2008.

[15] Habash, N., Rambow, O., and Roth, R. .MADA+TOKEN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming, and Lemmatization. Center for Computational Learning Systems, Columbia University, New York, NY, USA.

[16] Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskandar, R., Habash, N., Pooleery, M., Rambow, O., and Roth, R.2014.MADAMIRA: A Fast, Comprehensive Tool Morphological Analysis and Disambiguation of Arabic. Center for Computational Learning Systems, Columbia University, New York, NY, USA., Department of Computer Science, The George Washington University, Washington, DC., and Google, Inc.

[17] Habash, N., Jarrar, M., Akra, D., Rimawi, F., and Arrar, M. Conventional Orthography for Dialectal Arabic: Palestinian Arabic.2014. Birzeit university.

[18] Habash, N., Roth, R., Rambow, O., Eskander, R., and Tomeh, N. Morphological Analysis and Disambiguation for Dialectal Arabic.2013.In Proceedings of NAACL-HT, pages 426-432, Atlanta,Georgia,9-14 June 2013, Association for Computational Linguistics.

[19] Habash, N., Eskander, R., and Hawwari, A. Morphological Analyzer for Egyptian Arabic.2012.In Proceedings of the Twelfth Meeting of Special Interest Group on Computational Morphology and Phonology(SIGMORPHON2012), pages 1-9, Montreal, Canada, June7, 2012, Association for Computational Linguistics.

[20] Jarrar, M., Habash, N., Akra, D., and Zalmout, N.: Building A corpus for Palestinian Arabic: Preliminary Study. 2014. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing, pages 18-27, Doha, Qatar, October, 2014, Association for Computational Linguistics.

[21] Maamouri, M., Krouna, S., Slimane, D., Hamrouni, N, and Habash, N. Egyptian Arabic (ARZ) Morphological Analysis & POS Annotation.2012.

[22] Buckwalter, T.: Buckwalter Arabic Morphological analyzer version 2.0.2004. LDC catalog Number LDC2004L02, ISBN1-58563-342-0.

Appendices

Appendix #1

Building a Corpus for Palestinian Arabic: a Preliminary Study

Mustafa Jarrar, *Nizar Habash, Diyam Akra, Nasser Zalmout

Birzeit University, West Bank, Palestine
{mjarrar,nzalmout}@birzeit.edu, diyam@student.birzeit.edu

*New York University Abu Dhabi, United Arab Emirates
nizar.habash@nyu.edu

Abstract

This paper presents preliminary results in building an annotated corpus of the Palestinian Arabic dialect. The corpus consists of about 43K words, stemming from diverse resources. The paper discusses some linguistic facts about the Palestinian dialect, compared with the Modern Standard Arabic, especially in terms of morphological, orthographic, and lexical variations, and suggests some directions to resolve the challenges these differences pose to the annotation goal. Furthermore, we present two pilot studies that investigate whether existing tools for processing Modern Standard Arabic and Egyptian Arabic can be used to speed up the annotation process of our Palestinian Arabic corpus.

1. Introduction and Motivation

This paper presents preliminary results towards building a high-coverage well-annotated corpus of the Palestinian Arabic dialect (henceforth PAL), which is part of an ongoing project called *Curras*. Building such a PAL corpus is a first important step towards developing natural language processing (NLP) applications, for searching, retrieving, machine-translating, spell-checking PAL text, etc. The importance of processing and understanding such text is increasing due to the exponential growth of socially generated dialectal content at recent Social Media and Web 2.0 breakthroughs.

Most Arabic NLP tools and resources were developed to serve Modern Standard Arabic (MSA), which is the official written language in the Arab World. Using such tools to understand and process Arabic dialects (DAs) is a challenging task because of the phonological and morphological differences between DAs and MSA. In addition, there is no standard orthography for DAs. Moreover, DAs have limited standardized written resources, since most of the written dialectal content is the result of ad hoc and unstructured social conversations or commentary, in comparison to MSA's vast body of literary works.

The rest of this paper is structured as follows: We present important linguistic background in Section 2, followed by a survey of related work in Section 3. We then present the process of collecting the Curras Corpus (Section 4) and the challenges of annotating it (Section 5).

2. Linguistic Background

In this section we summarize some important linguistic facts about PAL that influence the decisions we made in this project. For more information on PAL and Levantine Arabic in general, see (Rice and Sa'id, 1960; Cowell, 1964; Bateson, 1967; Brustad, 2000; Halloun, 2000; Holes, 2004; Elihai, 2004). For a discussion of differences between Levantine and Egyptian Arabic (EGY), see Omar (1976).

2.1 Arabic and its dialects

The Arabic language is a collection of variants among which a standard variety (MSA) has a special status, while the rest are considered colloquial dialects (Bateson, 1967, Holes, 2004; Habash, 2010). MSA is the official written language of government, media and education in the Arab World, but it is not anyone's native language; the spoken dialects vary widely across the Arab World and are the true native varieties

Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), pages 18–27,
October 25, 2014, Doha, Qatar.

of Arabic, yet they have no standard orthography and are not taught in schools (Habash et al., 2012, Zribi et al., 2014).

PAL is the dialect spoken by Arabic speakers who live in or originate from the area of Historical Palestine. PAL is part of the South Levantine Arabic dialect subgroup (of which Jordanian Arabic is another dialect). PAL is historically the result of interaction between Syriac and Arabic and has been influenced by many other regional language such as Turkish, Persian, English and most recently Hebrew. The Palestinian refugee problem has led to additional mixing among different PAL sub-dialects as well as borrowing from other Arabic dialects. We discuss next some of the important distinguishing features of PAL in comparison to MSA as well as other Arabic dialects. We consider the following dimensions: phonology, morphology, and lexicon. Like other Arabic dialects, PAL has no standard orthography.

2.2 Phonology

PAL consists of several sub-dialects that generally vary in terms of phonology and lexicon preferences. Commonly identified sub-dialects include urban (which itself varies mostly phonologically among the major cities such as Jerusalem, Jaffa, Gaza, Nazareth, Nablus and Hebron), rural, and Bedouin. The Druze community has also some distinctive phonological features that set it apart. The variations are a miniature version of the variations in Levantine Arabic in general. Perhaps the most salient variation is the pronunciation of the /q/ phoneme (corresponding to MSA ق *q*[1]), which realizes as /'/ in most urban dialects, /k/ in rural dialects, and /g/ in Bedouin dialects. The Druze dialect retains the /q/ pronunciation. Another example is the /k/ phoneme (corresponding to MSA ك k), which realizes as /tš/ in rural dialects. These difference cause the word for قلب *qlb* 'heart' to be pronounced as /qalb/, /'alb/, /kalb/ and /galb/ and to be ambiguous out of context with the word كلب *klb* 'dog' /kalb/ and /tšalb/. And similarly to EGY (but unlike Tunisian Arabic), the MSA phoneme /θ/ (ث *θ*) becomes /s/ or /t/, and the MSA phoneme /ð/ (ذ *ð*) becomes /z/ or /d/ in different lexical contexts, e.g., MSA كذب *kðb* /kaðib/ 'lying' is pronounced /kizib/ in PAL and /kidb/ in EGY.

Similar to many other dialects, e.g. EGY and Tunisian (Habash et al., 2012; Zribi et al., 2014), the glottal stop phoneme that appears in many MSA words has disappeared in PAL: compare MSA رأس *rÂs* /ra's/ 'head' and بئر *bŷr* /bi'r/ 'well' with their Palestinian urban versions: /rās/ and /bīr/. Also, the MSA diphthongs /ay/ and /aw/ generally become /ē/ and /ō/; this transformation happens in EGY but not in other Levantine dialects such as Lebanese, e.g., MSA بيت *byt* /bayt/ 'house' becomes PAL /bēt/.

PAL also elides many short vowels that appear in the MSA cognates leading to heavier syllabic structure, e.g. MSA جبال /jibāl/ 'mountains' (and EGY /gibāl/) becomes PAL /jbāl/. Additionally long vowels in unstressed positions in some PAL sub-dialects shorten, a phenomenon shared with EGY but not MSA: e.g., compare /zāru/ (زاروا zAr+uwA) 'they visited' with /zarū/ (زاروه zAr+uw+h) 'they visited him'. Finally, PAL has commonly inserted epenthetic vowels (Herzallah, 1990), which are optional in some cases leading to multiple pronunciations of the same word, e.g., /kalb/ and /kalib/ (كلب *klb* 'dog'). This multiplicity is not shared with MSA, which has a simpler syllabic structure and more limited epenthesis than PAL.

2.3 Morphology

PAL, like MSA and its dialects and other Semitic languages, makes extensive use of templatic morphology in addition to a large set of affixations and clitics. There are however some important differences between MSA and PAL in terms of morphology. First, like many other dialects, PAL lost nominal case and verbal mood, which remain in MSA. Additionally, PAL in most of its sub-dialects collapses the feminine and masculine plurals and duals in verbs and

[1]Arabic orthographic transliterations are provided in the Habash-Soudi-Buckwalter (HSB) scheme (Habash et al., 2007), *except where indicated*. HSB extends Buckwalter's transliteration scheme (Buckwalter, 2004) to increase its readability while maintaining the 1-to-1 correspondence with Arabic orthography as represented in standard encodings of Arabic, i.e., Unicode, etc. The following are the only differences from Buckwalter's scheme (indicated in parentheses): Ā آ (|), Â أ (>), ŵ ؤ (&), Ǎ إ (<), ŷ ئ (}), ħ ة (p), θ ث (v), ð ذ (*), š ش ($), Ď ظ (Z), ς ع (E), γ غ (g), ý ى (Y), ã ً (F), ũ ٌ (N), ĩ ٍ (K). Orthographic transliterations are presented in italics. For phonological transcriptions, we follow the common practice of using '/.../' to represent phonological sequences and we use HSB choices with some extensions instead of the International Phonetic Alphabet (IPA) to minimize the number of representations used, as was done by Habash (2010).

most nouns. Some specific inflections are ambiguous in PAL but not MSA, e.g., حبيت *Hbyt* /Habbēt/ 'I (or you [m.s.]) loved'.

Second, some specific morphemes are slightly or quite different in PAL from their MSA forms, e.g., the future marker is /sa/ in MSA but /Ha/ or /raH/ in PAL. Another prominent example is the feminine singular suffix morpheme (Ta Marbuta), which in MSA is pronounced as /at/ except at utterance final positions (where it is /a/). In some PAL urban sub dialects, it has multiple allomorphs that are phonologically and syntactically conditioned: /a/ (after non-front and emphatic consonants), /e/ (after front non-emphatic consonants), /it/ (nouns in construct state such as before possessive pronouns) and /ā/ (in deverbals before direct objects): e.g. بطة *bTħ* /baTT+a/ 'duck', حبة *Hbħ* /Habb+e/ 'pill', بطتنا *bTnA* /baTT+it+na/ 'our duck' and /mdars+ā +hum/ 'she taught them'.

Third, PAL has many clitics that do not exist in MSA, e.g., the progressive particle /b+/ (as in /b+tuktub/ 'she writes'), the demonstrative particle /ha+/ (as in /ha+l+bēt/ 'this house'), the negation cirmcumclitic /ma+ +š/ (as in /ma+katab+š/ 'he did not write') and the indirect object clitic (as in /ma+katab+l+ō+š/ 'he did not write to him'). All of these examples except for the demonstrative particle are used in EGY.

2.4 Lexicon

The PAL lexicon is primarily Arabic with numerous borrowings from many different languages. MSA cognates generally appear with some minor phonological changes as discussed above; a few cases include more complex changes, e.g. /biddi/ 'I want' is from MSA /bi+widd+i/ 'in my desire' or /illi/ 'relative pronoun which/who/that' which corresponds to a set of MSA forms that inflect for gender and number (الذي *Alðy*, التي *Alty*, etc.). Some common PAL words are portmanteaus of MSA words, e.g., /lēš / 'why?' corresponds to MSA /li+'ayy+i šay'/ 'for what thing?'. Examples of common words that are borrowed from other languages include the following:

- روزنامه /roznama/ 'calendar' (Persian)
- كندرة /kundara/ 'shoe' (Turkish)
- بندورة /banadora/ 'tomato' (Italian)
- بريك /brēk/ 'brake (car)' (English)
- تليفيزيون /talifizyon/ 'television' (French)
- محسوم /maHsūm/ 'checkpoint' (Hebrew)

3. Related Work

3.1 Corpus Collection and Annotation

There have been many contributions aiming to develop annotated Arabic language corpora, with the main objective of facilitating Arabic NLP applications. Notable contributions targeting MSA include the work of Maamouri and Cieri, (2002), Maamouri et al. (2004), Smrž and Hajič (2006), and Habash and Roth (2009). These efforts developed annotation guidelines for written MSA content producing large-scale Arabic Treebanks.

Contributions that are specific to DA include the development of a pilot Levantine Arabic Treebank (LATB) of Jordanian Arabic, which contained morphological and syntactic annotations of about 26,000 words (Maamouri et al., 2006). To speed up the process of creating the LATB, Maamouri et al. (2006) adapted MSA Treebank guidelines to DA and experimented with extensions to the Buckwalter Arabic Morphological Analyzers (Buckwalter, 2004). The LATB was used in the Johns Hopkins workshop on Parsing Arabic Dialect (Rambow et al., 2005; Chiang et al., 2006), which supplemented the LATB effort with an experimental Levantine-MSA dictionary. The LATB effort differs from the work presented here in two respects. First, the LATB corpus consists of conversational telephone speech transcripts, which eliminated the orthographic variations issues that we face in this paper. Secondly, when the LATB was created, there were no robust tools for morphological analysis of any dialects; this is not the case any more. We plan to exploit existing tools for EGY to help the annotation effort.

Other DA contributions include the Egyptian Colloquial Arabic Lexicon (ECAL) (Kilany, et al., 2002), which was developed as part of the CALLHOME Egyptian Arabic (CHE) corpus (Gadalla, et al., 1997). In addition to YADAC (Al-Sabbagh and Girju, 2012), which was based on dialectal content identification and web harvesting of blogs, micro blogs, and forums of EGY content. Similarly, the COLABA project (Diab et al., 2010) developed annotated dialectal content resources for Egyptian, Iraqi, Levantine, and Moroccan dialects, from online weblogs.

3.2 Dialectal Orthography

Due to the lack of standardized orthography guidelines for DA, along with the phonological differences in comparison to MSA, and dialectal variations within the dialects themselves, there are many orthographic variations for written DA content. Writers in DA, regardless of the context, are often inconsistent with others and even with themselves when it comes to the written form of a dialect; writing with MSA driven orthography, or writing words phonologically sometimes. These orthography variations make it difficult for computational models to properly identify and reason about the words of a given dialect (Habash et al, 2012a), hence, a conventional form for the orthographic notations is important. Within this scope, we can view this problem for Levantine dialects as an extension of the work of Habash et al. (2012a) who proposed the so-called CODA (Conventional Orthography for Dialectal Arabic). CODA is designed for the purpose of developing conventional computational models of Arabic dialects in general. Habash et al. (2012a) provides a detailed description of CODA guidelines as applied to EGY. Eskander et al. (2013) identify five goals for CODA: (i) CODA is an internally consistent and coherent convention for writing DA; (ii) CODA is created for computational purposes; (iii) CODA uses the Arabic script; (iv) CODA is intended as a unified framework for writing all DAs; and (v) CODA aims to strike an optimal balance between maintaining a level of dialectal uniqueness and establishing conventions based on MSA-DA similarities. CODA guidelines will be extended to cover PAL in this paper, as discussed in Section 5.3.

3.3 Dialectal Morphological Annotation

Most of the work that explored morphology in Arabic focused on MSA (Al-Sughaiyer and Al-Kharashi, 2004; Buckwalter, 2004; Habash and Rambow, 2005; Graff et al., 2009; Habash, 2010). The contributions for DA morphology analysis, however, are relatively scarce and are usually based on either extending available MSA tools to tackle DA specificities, as in the work of (Abo Bakr et al., 2008; Salloum and Habash, 2011), or modeling DAs directly, without relying on existing MSA contributions (Habash and Rambow, 2006). Due to the variations between MSA and DAs, available MSA tools and resources cannot be easily extended or transferred to work properly for DA (Maamouri, et al., 2006; Habash, et al., 2012b). Therefore, it is important to develop annotated and morpheme-segmented resources, along with morphological analysis tools, that are specific and tailored for DAs. One of the notable recent contributions for EGY morphological analysis was CALIMA (Habash et al., 2012b). The CALIMA analyzer for EGY and the commonly used SAMA analyzer for MSA (Graff et al., 2009) are central in the functioning of the EGY morphological tagger MADA-ARZ (Habash et al., 2013), and its successor MADAMIRA (Pasha et al., 2014), which supports both MSA and EGY.

The work we present in this paper builds on the shoulders of these previous efforts from the development of guidelines for orthography and morphology (in MSA and EGY) to the use of existing tools (specifically MADAMIRA MSA and EGY) to speed up the annotation process.

4. Corpus Collection

Written dialects in general tend to have scarce resources in terms of written literature; written materials usually involve informal conversations or traditional folk literature (stories, songs, etc.). It is therefore often difficult to find resources for written dialectal content. In addition, resources of dialectal content are prone to significant noise and inconsistency because they tend to lack standard orthographies and rely on ad hoc transcriptions and orthographic borrowing from the standard variety. In the case of Arabic, unlike MSA that dominates the formal and written content outlets, as in the press, scientific articles, books, and historical narration, DAs are more naturally used in traditional and informal contexts, such as conversations in TV series, movies, or on social media platforms, providing socially powered commentary on different domains and topics. And given the lack of standard orthography, there is common mixing of phonetic spelling and MSA-cognate-based spelling in addition to the so-called Arabizi spelling – writing DAs in Roman script, rather than Arabic script (Darwish, 2014 and Al-Badrashiny et al., 2014). Such noise imposes many challenges regarding the collection of high-coverage high-accuracy DA corpora. It is therefore important to remark that although *bigger is better* when it comes to corpus size, we focus more in this first iteration of our PAL corpus on precision and variety rather than mere

size. That is, we tried not only to manually select and review the content of the corpus, but also to assure that we covered a variety of topics and contexts, localities and sub-dialects, including the social class and gender of the speakers and writers. This is because such aspects help us discover new language phenomena in the dialect as will be discussed in the next section.

Table 1 presents the resources that we manually collected to build the PAL Curras corpus. There are 133 social media threads (about 16k words) from blogs (e.g., مدونة عبد الحميد العاطي Abdelhameed Alaaty's blog), forums (e.g., شبكة الحوار الفلسطيني The Palestinian dialogue network), Twitter, and Facebook. The collection was done by reading many discussion threads and selecting the relevant ones to assure diversity and PAL representative content. Content that is heavily written in a mix of languages, or a mix of other dialects was excluded. In the same way, we also manually collected some PAL stories, and a list of PAL terms and their meanings, which reflect additional diversity of topics, contexts, and social classes. About half of our corpus comes from 41 episode scripts from the Palestinian TV show وطن ع وتر "Watan Aa Watar". Each episode discusses and provides satirical critiques regarding different topics of relevance to the Palestinian viewers about daily life issues. The show's importance stems from the fact that the actors use a variety of Palestinian local dialects, hence enriching the coverage of the corpus.

Table 1. The Curras Corpus Statistics

Document Type	Word Tokens	Word Types	Documents
Facebook	3,120	1,985	35 threads
Twitter	3,541	2,133	38 threads
Blogs	8,748	4,454	37 threads
Forums	1,092	798	33 threads
Palestinian Stories	2,407	1,422	6 stories
Palestinian Terms	759	556	1 doc
TV Show: وطن ع وتر *Watan Aa Watar*	23,423	8,459	41 episodes
Curras Total	**43,090**	**19,807**	**191**

5. Corpus Annotation Challenges

This section presents our approach to annotating the Curras corpus. We start with a specification of our annotation goals, followed by a discussion of our general approach. We then discuss in more details two important challenges that need to be addressed for annotation of a new dialectal corpus: orthography and morphology.

5.1 Annotation Specification

The words are annotated in context. As such, the same word may receive different annotations in different contexts. We define the annotation of a word as a tuple $<w, w_B, c, c_B, l, p_B, g, i>$ described as follow. (Examples of such annotations are illustrated in Table 5.):

- ***w*: Raw (Unicode)** The raw input word defined as a string of letters delimited by white space and punctuation. The word is represented in Arabic script (Unicode).
- ***w_B*: Raw (Buckwalter)** The same raw input word in the commonly used Buckwalter transliteration (Buckwalter, 2004).
- ***c*: CODA (Unicode)** The Conventional Orthography (Habash et al., 2012) version of the input word.
- ***c_B*: CODA (Buckwalter)** The Buckwalter transliteration of the CODA form.
- ***l*: Lemma** The lemma of the word in Buckwalter transliteration. The lemma is the citation form or dictionary entry that abstracts over all inflectional morphology (but not derivational morphology). The lemma is fully diacritized. We follow the definition of lemma used in BAMA (Buckwalter, 2004) and CALIMA-ARZ (Habash et al., 2012b).
- ***p_B*: Buckwalter POS** The Buckwalter full POS tag, which identifies all clitics and affixes and the stem and assigns each a sub-tag. This representation treats clitics as separate tokens and abstracts the orthographic rewrites they undergo when cliticized. See the handling of the l/PREP+Al/DET in word #6 in Table 5. This representation is used by the LDC in the Penn Arabic Treebank (PATB) (Maamouri et al., 2004) and tools such as MADAMIRA (Pasha et al., 2014). It is a high granularity representation that allows researchers to easily go to coarser granularity POS (Diab 2007; Habash, 2010; Alkuhlani et al., 2013). The Buckwalter POS tag can be fully diacritized or undiacritized. Given the added complexity of producing diacritized text manually by annotators, we opted at this stage to only use undiacritized forms.

- ***g*: Gloss** The English gloss, an informal semantic denotation of the lemma. In Tables 3-5, we only use one English word for space limitations.
- ***i*: Analysis** A specification of the source of the annotation, e.g., ANNO is a human annotator, and MADA is the MADAMIRA system with some minor or no automatic post-processing. In Tables 3 and 4, which are produced automatically, the Analysis field is replaced with a status indicating how usable the automatic annotation is.

5.2 General Approach

To speed up the process of annotating our corpus, we made the following decisions. First, and quite obviously from the previous section, we made a conscious decision to follow on the footsteps of previous efforts for MSA and EGY annotation done at the Linguistic Data Consortium and Columbia's Arabic Modeling group in terms of guidelines for orthography conventionalization and morphological annotation. This allows us to exploit existing guidelines with only essential modification to accommodate PAL and produce annotations that are comparable to those done for MSA and EGY. This, we hope, will encourage research in dialectal adaptation techniques and will make our annotations more familiar and thus usable by the community.

Second, and closely related to the first point, we exploit existing tools to speed up the annotation process. In this paper, we specifically use the MADAMIRA tool (Pasha et al., 2014) for morphological analysis and disambiguation of MSA and EGY. Our choice of using this tool is motivated by the assumption that EGY/MSA and PAL share many orthographic and morphological features. This assumption was validated by pilot experiments, presented below, and which show most of the PAL annotations can be generated automatically. However, a manual step is then needed to verify every annotation, to correct errors and fill in gaps. The manual annotation has not been completed yet as of the writing of this paper submission.

Finally, we made one major simplification to the annotations to minimize the load on the human annotator: we do not produce diacritized morphological analyses in the Buckwalter POS tag. The reasons for this decision are the following: (i) full diacritization is a complex task that most Arabic speakers do not do and thus it requires a lot of training and precious attention to detail; (ii) MSA and EGY produce many morphemes and lexical items that are quite similar to PAL except in terms of the short vowels (compare the lemmas for word #5 in Tables 3, 4 and 5); (iii) PAL has many cases of multiple valid diacritizations as mentioned above. While we think a convention should be defined to explain the variation and model it, it is perhaps the topic of a future effort that is more focused on PAL phonology. We make an exception for the lemmas and diacritize them since lemmas are important in indicating the core meaning of the word. In case of different pronunciations of the lemma, we choose the shortest.

5.3 A Conventional Orthography for PAL

As explained in Section 2, PAL, like other Arabic dialects, does not have a standard orthography. Furthermore, there are numerous phonological, morphological and lexical differences between PAL and MSA that make the use of MSA spelling as is undesirable. PAL speakers who write in the dialect produce spontaneous inconsistent spellings that sometimes reflect the phonology of PAL, and other times the word's cognate relationship with MSA. For example, the word for 'heart' (MSA قلب qalb) has four spellings that correspond to four sub-dialectal pronunciations: قلب *qlb* /qalb/, ألب Âlb /'alb/, كلب *klb* /kalb/, and جلب *jlb* /galb/. Similarly, the common shortening of some long vowels (from MSA to PAL) leads to different orthographies as in قانون *qAnwn* 'law' (MSA /qānūn/), which can also be written with a shortened first vowel قنون *qnwn* /'anūn/ reflecting the PAL pronunciation. PAL also has some clitics that do not exist in MSA, which leads to different spellings, e.g. the PAL future particle ح *H* /Ha/ can be written attached to or separate from the verb that follows it. Even when a morpheme exists in MSA and PAL, it may have additional forms or pronunciations. One example is the definite article morpheme ال *Al* /il/ which has a non-MSA/non-EGY allomorph /li/ when attached to nominals with initial consonant clusters. As a result, a word like /li+blād/ 'the homeland/countries' can be spelled to reflect the morphology as البلاد *AlblAd* or the phonology لبلاد *lblAd*, with the latter being ambiguous with 'for countries' (in PAL /la+blād/). Finally, there are words in PAL that have no cognate in MSA and as such have no

clear obvious spelling to go with, e.g., the word /barDo/ 'additionally' is spontaneously written as برضو *brDw*, برضه *brDh* and برضة *brDẖ*.

This, of course, is not a unique PAL problem. Researchers working on NLP for EGY and Tunisian dialects developed CODA guidelines for them (Habash et al., 2012a; Zribi et al., 2014). These guidelines were by design intended to apply (or be easily extended) to all Arabic dialects, but were only demonstrated for two. Our challenge was to take these guidelines (specifically the EGY version) and extend them. There were three types of extensions. First, in terms of phonology-orthography, we added the letter ك k to the list of root letters to be spelled in the MSA cognate to cover the PAL rural sub-dialects that pronounces it as /tš/. Second, in terms of morphology, we added the non-EGY demonstrative proclitic ه *h+* and the conjunction proclitic ت *t+* 'so as to' to the list of clitics, e.g., بهالبيت *bhAlbyt* 'in this house' and تيشوف *tyšwf* 'so that he can see'. Finally, we extended the list of exceptional words to cover problematic PAL words. All of the basic CODA rules for EGY (and Tunisian) are kept the same.

Pilot Study (I): We conducted a small pilot study in annotating the CODA for PAL words. We considered 1,000 words from 77 tweets in Curras. The CODA version of each word was created in context. 15.9% of all words had a different CODA form from the input raw word form. 42% of these changes involve consonants (two-fifths of the cases), vowels (one-fifth of the cases) and the hamzated/bare forms of the letter Alif ا *A*. Examples of consonant change can be seen in Table 5 (words #4 and #10). An additional 29% word changes involve the spelling of specific morpheme. The most common change (over half of the time) was for the first person imperfect verbal prefix ا *A* when following the progressive particle ب *b*: بكتب *bktb* as opposed to باكتب *bAktb*. About 18% of the changed words experience a split or a merge (with splits happening five time more than merges). An example of a CODA split is seen in Table 5 (word #9). Finally, only about 8% of the changed words were PAL specific terms; and less than 7% involved a typo or speech effect elongation. These results are quite encouraging as they suggest the differences between CODA and spontaneously written PAL are not extensive. Further analysis is still needed of course.

In Tables 3 and 4 (column CODA), we show the results of using the MADAMIRA-MSA and MADAMIRA-EGY systems on a set of ten words, while Table 5 shows the manually selected or corrected CODA. MADAMIRA generates a CODA version (contextually) by default. We expect the EGY version to be more successful than the MSA version in producing the CODA for PAL given the shared presence of many morphemes in EGY and PAL. However, when we ran the same set of words through MADAMIRA-EGY, we encountered many errors in words, morphemes and spelling choices in PAL that are different from EGY, e.g., the raw word منحب *mnHb* 'we love' (CODA بنحب *bnHb*) is analyzed as the EGY ما نحب *mA nHb* 'we do not love'!

5.4 Morphological Annotation Process and Challenges

To study the value of using an existing morphological analyzer for MSA or EGY in creating PAL annotations, we conducted the following pilot study.

Pilot Study (II): We ran the words from a randomly selected episode of the PAL TV show "Watan Aa Watar" (460 words) through both MADAMIRA-MSA and MADAMIRA-EGY. We analyzed the output from both systems to determine its usability for PAL annotations. We consider all analyses that are correct for PAL annotation or usable via simple post processing (such as removing CASE endings on MSA words) to be correct (as in word #2 in Tables 3-5). Words that receive incorrect analyses or no analyses require manual modifications.

The results of this experiment are summarized in Table 2. Table 3 and 4 illustrate sample results for ten words and Table 5 includes the manually created results.[2]

Table 2. Accuracy of automatic annotation of PAL text

Statistics	MADAMIRA MSA	MADAMIRA EGY
No Analysis	17.78%	7.24%
Wrongly Analyzed	18.43%	14.75%
Correctly Analyzed	63.79%	78.01%

The No Analysis (NA) words in Tables 2, 3 and 4 refer to the words that the morphological analyzer couldn't recognize. This failure may be

[2] The examples in Tables 3-5 are presented in the Buckwalter transliteration (Buckwalter, 2004) to match the forms as they appear in the annotated corpus.

a result of missing lexical entry, specific PAL morphology or typos. As expected, MADAMIRA-MSA had 2.5 times the number of NA cases compared to MADAMIRA-EGY. Examples include dialectal lexical terms (word #7) or dialectal morphology (words # 1 and #9).

The wrongly analyzed words are words that were assigned incorrect POS tag *in context*. For example, word #3 in Tables 3 and 4 is the result of mis-analyzing the proclitic l- as the preposition 'for/to' as opposed to the non-CODA spelling of the definite article in PAL. The analysis provided by MADAMIRA-EGY is correct for other contexts than the one illustrated here. Another example is word #8, which is a Levantine specific term hardly used in EGY and not used at all in MSA. MADAMIRA-MSA has a higher proportion of wrongly analyzed words than MADAMIRA-EGY.

Overall MADAMIRA-EGY produced analyses that were either correct and ready to use for PAL or requiring some minor modifications such as adjusting the vowels on the lemmas (e.g., word #5) in one of every five words.

Table 3 Automatic annotations by the MADAMIRA-MSA system. Entries with Status NA had no analysis.

	Raw		CODA		Lemma	Buckwalter POS (Diacritized)	Gloss	Status
1	ابوكوا	AbwkwA						NA
2	الاكل	AlAkl	الأكل	Al>kl	>akol	Al/DET+>akol/NOUN+a/CASE_DEF_ACC	eating	Usable
3	لبنوك	lbnwk	لبنوك	lbnwk	banok	li/PREP+bunuwk/NOUN+K/CASE_INDEF_GEN	bank	Wrong
4	التاني	AltAny	التأني	Alt>ny	ta>an~iy	Al/DET+ta>an~iy/NOUN	prudence	Wrong
5	الحمار	AlHmAr	الحمار	AlHmAr	HimAr	Al/DET+HimAr/NOUN+u/CASE_DEF_NOM	donkey	Usable
6	للراتب	llrAtb	للراتب	llrAtb	rAtib	li/PREP+Al/DET+rAtib/NOUN+i/CASE_DEF_GEN	salary	Usable
7	ايوة	Aywp						NA
8	بدها	bdhA	بدها	bdhA	bud~	bud~/NOUN+i/CASE_DEF_GEN+hA/POSS_PRON_3FS	escape	Wrong
9	بنردلك	bnrdlk						NA
10	هدول	hdwl						NA

Table 4 Automatic annotations by the MADAMIRA-EGY system. Entries with Status NA had no analysis.

	Raw		CODA		Lemma	Buckwalter POS (Diacritized)	Gloss	Status
1	ابوكوا	AbwkwA	ابوكو	Abwkw	Abuw	Abuw/NOUN+kuw/POSS_PRON_3MS	father	Correct
2	الاكل	AlAkl	الأكل	Al>kl	>akl	Al/DET+>akol/NOUN	eating	Correct
3	لبنوك	lbnwk	لبنوك	lbnwk	bank	li/PREP+bunuwk/NOUN	bank	Wrong
4	التاني	AltAny	التاني	AltAny	tAniy	Al/DET+tAniy/ADJ_NUM	second	Usable
5	الحمار	AlHmAr	الحمار	AlHmAr	HumAr	Al/DET+HumAr/NOUN	donkey	Usable
6	للراتب	llrAtb	للراتب	llrAtb	rAtib	li/PREP+Al/DET+rAtib/NOUN	salary	Correct
7	ايوة	Aywp	أيوه	>ywh	>ayowah	>ayowah/INTERJ	yes	Correct
8	بدها	bdhA	بدها	bdhA	bud~	bud~/NOUN+hA/POSS_PRON_3FS	escape	Wrong
9	بنردلك	bnrdlk	بنرد_لك	bnrd_lk	rad~	bi/PROG_PART+nu/IV1P+rud~/IV+li/PREP+ak/PRON_2MS	answer	Usable
10	هدول	hdwl						NA

Table 5 Manual Annotations in Curras. Entries with Analysis MADA were automatically converted and validated by the annotator. Entries with Analysis ANNO required some modification of the MADAMIRA output or were created from scratch.

	Raw		CODA		Lemma	Buckwalter POS (Undiacritized)	Gloss	Analysis
1	ابوكوا	AbwkwA	ابوكو	Abwkw	Abuw	Abw/NOUN+kw/POSS_PRON_3MS	father	MADA
2	الاكل	AlAkl	الأكل	Al>kl	>akl	Al/DET+>kl/NOUN	eating	MADA
3	لبنوك	lbnwk	البنوك	Albnwk	bank	Al/DET+bnwk/NOUN	bank	ANNO
4	التاني	AltAny	الثاني	AlvAny	vAniy	Al/DET+vAny/ADJ_NUM	second	ANNO
5	الحمار	AlHmAr	الحمار	AlHmAr	HmAr	Al/DET+HmAr/NOUN	donkey	MADA
6	للراتب	llrAtb	للراتب	llrAtb	rAtib	l/PREP+Al/DET+rAtb/NOUN	salary	MADA
7	ايوة	Aywp	أيوه	>ywh	>ayowah	>ywh/INTERJ	yes	MADA
8	بدها	bdhA	بدها	bdhA	bid~	bd/NOUN+hA/POSS_PRON_3FS	want	ANNO
9	بنردلك	bnrdlk	بنرد_لك	bnrd_lk	rad~	b/PROG_PART+n/IV1P+rd/IV+l/PREP+k/PRON_2MS	answer	MADA
10	هدول	hdwl	هذول	h*wl	ha*A	h*wl/DEM_PRON	these	ANNO

5 Conclusion and Future Work

We presented our preliminary results towards building an annotated corpus of the Palestinian Arabic dialect. The challenges and linguistic variations of the Palestinian dialect, compared with Modern Standard Arabic, were discussed especially in terms of morphology, orthography, and lexicon. We also discussed and showed the potential, and limitations, of using existing resources, especially MADAMIRA-EGY, to semi-automate and speed up the annotation process.

The paper has also pointed out several issues that need to be considered and researched further, especially the development of Palestinian-specific morphological annotation and CODA guidelines, a Palestinian lexicon, and the extension of MADAMIRA to analyze Palestinian text. Our corpus will be further extended to include more text, and all lexical annotations (i.e., Lemmas) will be linked with existing Arabic ontology resources such as the Arabic WordNet (Black et al., 2006). The corpus will be publicly available for research purposes.

Acknowledgement

This work is part of the ongoing project *Curras*, funded by the Palestinian Ministry of Higher Education, Scientific Research Council. Nizar Habash performed most of his work on this paper while he was in the Center for Computational Learning Systems at Columbia University.

References

H. Abo Bakr, K. Shaalan, and I. Ziedan. A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic. In The 6th International Conference on Informatics and Systems, INFOS2008. Cairo University, 2008.

M. Al-Badrashiny, R. Eskander, N. Habash, and O. Rambow. Automatic Transliteration of Romanized Dialectal Arabic. CoNLL, 2014.

S. Alkuhlani, N. Habash and R. M. Roth. Automatic Morphological Enrichment of a Morphologically Underspecified Treebank. In Proc. of Conference of the North American Association for Computational Linguistics (NAACL), Atlanta, Georgia, 2013.

R. Al-Sabbagh and R. Girju. YADAC: Yet another dialectal Arabic corpus. In Proc. of the Language Resources and Evaluation Conference (LREC), pages 2882–2889, Istanbul, 2012.

M. C. Bateson. Arabic Language Handbook. Center for Applied Linguistics, Washington D.C., USA, 1967.

W. Black, Elkateb, S., & Vossen, P. (2006). Introducing the Arabic wordnet project. In In Proceedings of the third International WordNet Conference (GWC-06).

K. Brustad. The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects. Georgetown University Press, 2000.

T. Buckwalter. Buckwalter Arabic morphological analyzer version 2.0. LDC catalog number LDC2004L02, ISBN 1-58563-324-0, 2004.

D. Chiang, M. Diab, N. Habash, O. Rambow, and S. Shareef. Parsing Arabic Dialects. In Proceedings of the European Chapter of ACL (EACL), 2006.

M. W. Cowell. A Reference Grammar of Syrian Arabic. Georgetown University Press, 1964.

Kareem Darwish. Arabizi Detection and Conversion to Arabic. In the Arabic Natural Language Processing Workshop, EMNLP, Doha, Qatar, 2014.

M. Diab. Towards an Optimal POS tag set for Modern Standard Arabic Processing. In Proc. of Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria, 2007.

M. Diab, N. Habash, O. Rambow, M. Altantawy, and Y. Benajiba. COLABA: Arabic Dialect Annotation and Processing. LREC Workshop on Semitic Language Processing, Malta, 2010.

Y. Elihai. The olive tree dictionary: a transliterated dictionary of conversational Eastern Arabic (Palestinian). Washington DC: Kidron Pub, 2004.

R. Eskander, N. Habash, O. Rambow, and N. Tomeh. Processing Spontaneous Orthography. In Proceedings NAACL-HLT, Atlanta, GA, 2013.

H. Gadalla, H. Kilany, H. Arram, A. Yacoub, A. El-Habashi, A. Shalaby, K. Karins, E. Rowson, R. MacIntyre, P. Kingsbury, D. Graff, and C. McLemore. CALLHOME Egyptian Arabic Transcripts. Linguistic Data Consortium, Catalog No.: LDC97T19, 1997.

D. Graff, M. Maamouri, B. Bouziri, S. Krouna, S. Kulick, and T. Buckwalter. Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73, 2009.

N. Habash and O. Rambow. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In ACL, Ann Arbor, Michigan, 2005.

N. Habash, A. Soudi, and T. Buckwalter. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, Arabic Computational Morphology: Knowledge-based and Empirical Methods. Springer, 2007.

N. Habash and R. Roth. CATiB: The Columbia Arabic Treebank. In ACL, 2009.

N. Habash. Introduction to Arabic natural language processing, volume 3. Morgan & Claypool Publishers, 2010.

N. Habash, M. Diab, and O. Rabmow. (2012a) Conventional Orthography for Dialectal Arabic. In Proc. of LREC, Istanbul, Turkey, 2012.

N. Habash, R. Eskander, and A. Hawwari. (2012b) A Morphological Analyzer for Egyptian Arabic. In Proc. of the Special Interest Group on Computational Morphology and Phonology, Montréal, Canada, 2012.

N. Habash, R. Roth, O. Rambow, R. Eskander, and N. Tomeh. Morphological Analysis and Disambiguation for Dialectal Arabic. In Proc. of NAACL, Atlanta, Georgia, 2013.

M. Halloun. A Practical Dictionary of the Standard Dialect Spoken in Palestine. Bethlehem University, 2000.

R. Herzallah. Aspects of Palestinian Arabic Phonology: A Nonlinear Approach. Ph.D. thesis, Cornell University. Distributed as Working Papers of the Cornell Phonetics Laboratory No. 4, 1990.

C. Holes. Modern Arabic: Structures, Functions, and Varieties. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press, 2004.

H. Kilany, H. Gadalla, H. Arram, A. Yacoub, A. El-Habashi, and C. McLemore. Egyptian Colloquial Arabic Lexicon. Linguistic Data Consortium, Catalog No.: LDC99L22, 1999.

M. Maamouri, A. Bies, T. Buckwalter, M. Diab, N. Habash, O. Rambow, and D. Tabessi. Developing and using a pilot dialectal Arabic treebank. In Proc. of LREC, Genoa, Italy, 2006.

M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt, 2004.

M. Maamouri, A. Bies, S. Kulick, M. Ciul, N. Habash and R. Eskander. Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In Proc. of LREC, Reykjavik, Iceland, 2014.

M. Maamouri, and C. Cieri. Resources for Arabic Natural Language Processing at the Linguistic Data Consortium. In Proc. of the International Symposium on Processing of Arabic. Faculté des Lettres, University of Manouba, Tunisia, 2002.

M. Omar. Levantine and Egyptian Arabic: Comparative Study. Foreign Service Institute. Basic Course Series, 1976.

A. Pasha, M. Al-Badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow and R. M. Roth. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In Proc. of LREC, Reykjavik, Iceland, 2014.

O. Rambow, D. Chiang, M. Diab, N. Habash, R. Hwa, K. Sima'an, V. Lacey, R. Levy, C. Nichols, and S. Shareef. 2005. Parsing Arabic Dialects. Final Report, 2005 JHU Summer Workshop.

F. Rice and M. Sa'id. Eastern Arabic: an introduction to the spoken Arabic of Palestine, Syria and Lebanon. Beirut: Khayat's 1960.

W. Salloum and N. Habash. Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation. In Proc. of the First Workshop on Algorithms and Resources for Modeling of Dialects and Language Varieties, Edinburgh, Scotland, 2011.

O. Smrž and J. Hajič. The Other Arabic Treebank: Prague Dependencies and Functions. In Ali Farghaly, editor, Arabic Computational Linguistics. CSLI Publications, 2006.

I. Zribi, R. Boujelbane, A. Masmoudi, M. Ellouze Khmekhem, L. Hadrich Belguith, and N. Habash. A Conventional Orthography for Tunisian Arabic. In Proc. of LREC, Reykjavik, Iceland, 2014.