الجمهورية الجزائرية الديمقراطية الشعبية

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
وزارة التعليم العالي و البحث العلمي
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
جامعة عمّار ثليجي بالأغواط
UNIVERSITY AMAR TELIDJI OF LAGHOUAT

كلية العلوم

FACULTY OF SCIENCES

DEPARTMENT OF COMPUTER SCIENCE

Automatic Classification of Audio Sequences: Application to Algerian Dialects

Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science (LMD)

Thesis defended on June 15, 2019

Bougrine Soumia

Jury members

N. Lagraa	Professeur	University of Laghouat	President
M. Abbas	Director of Research	CRSTDLA Alger	Examiner
B. Ziani	MCA	University of Laghouat	Examiner
H. Cherroun	Professeur	University of Laghouat	Advisor
D. Ziadi	Professeur	University of Rouen	Co-Advisor
M. Dadoune	MCA	University of Laghouat	Guest

2018/2019

To my parents, brother, sisters, and husband.

Acknowledgements

"GOD Thank you for giving me the strength and encouragement especially during all the challenging moments in completing this thesis. I am truly grateful for your exceptional love and grace during this entire journey"

This thesis would note have been possible without the support and guidance of many people.

First and foremost, my heartiest gratitude to my supervisors, Professor Hadda Cherroun and Professor Djelloul Ziadi, for their patience, continuous supervision, guidance, advice, and support in completing this thesis.

Special thanks to Prof. Nasreddine Lagraa, Dr. Mourad Abbas, Dr. Benameur Ziani, and Dr. Messaoud Dadoune for accepting the evaluation of my thesis despite their occupations, it was an honor for me to work with them.

I am specially grateful for the support of all co-authors who made contributions to my articles and their stimulating role in addressing the issues in this thesis: Dr. Ahmed Abdelali, Aicha Chorana, and Abdallah Lakhdari.

I am grateful to all my colleagues at Laboratoire d'Informatique et de Mathématiques (LIM) and specially MoDoS team.

Most of all, I am indebted to my family and my future husband for their comprehension, support, love and encouragement, without them this thesis would not have been possible. Thanks for always being so supportive, encouraging, and patient.

Finally, my thanks go to all the people who have supported me to complete the research work directly or indirectly.

مُلخّص

إن التعرف الآلي على لهجة المتحدث مهمة صعبة و تصبح أكثر تعقيدًا عند التعامل مع اللهجات ضعيفة الموارد و من نفس البلد. تعتبر المدونات الصوتية عاملا أساسيا لتطوير أنظمة التعرف الآلي للهجة المتحدث وتقييمها بدقة.

إن هدفنا في هاته الرسالة، يتمكن من ناحية في بناء مدونات صوتية للهجات الجزائرية و التي تعتبر من أكثر اللهجات العربية تعقيدًا و دراستها لم تلقى إهتمامًا. و من ناحية أخرى، تطوير نظام لتميز لهجات العربية خصوصا إذا كانت تنتمي إلى نفس البلد.

اهتمامنا في هذا البحث خص ثلاثة محاور متعلقة بمعالجة اللغات الطبيعية. المحور الأول يتعلق ببناء الموارد اللغوية (مدونات أو متون). قمنا بإنشاء مدونتين صوتيتين جزائريتين. أولا: Alg-Daridjah مدونة صوتية مخصصة للدراسات اللغوية. و قد تم جمعها بإستخدام طريقة التسجيل المباشر التي أتاحت لنا السيطرة على نوعية البيانات المجمعة. و هي مدونة متوسطة الحجم تشمل بعض اللهجات المتداولة في 17 ولاية، عدد المتحدثين هو 109 ، و تشمل أكثر من 6000 محادثة. المدونة الثانية: Kalam'DZ تحوي على عدد كبير من التسجيلات (المحادثات) المستخرجة من الواب و هي مخصصة لتدريب أنظمة التعلم الآلي. إعتمدنا في جمعها على تطبيقات مفتوحة المصدر. تحوي هذه المدونة على 8 لهجات الأكثر تداولاً في الجزائر، تشمل 4881 متحدث و حجمها أكثر من 104 ساعة.

المحور الثاني متعلق بتهميش مدونات. إستخدمنا تقنية التعهيد الجماعي للتحقق من صحة تصنيف لهجات مدونة Kalam'DZ التي تمت بصفة شبه أوتماتيكية. من هذه التجربة، قمنا باقتراح قائمة لأفضل الممارسات التي يجب إتباعها في عملية تهميش المدونات بناءًا على هذه التقنية.

و أخيرا، إقترحنا نظامين لتعرف على لهجات العربية المتداولة في نفس البلد. في كلا النظامين إعتمدنا على معلومات حول لحن اللغة و التقنيات الحديثة لتعلم الآلي. ركزنا على قياس القدرة التمييزية للحن اللغة و تقنيات التعليم المعمق. النظام الأول يستند على التصنيف الكلاسيكي أما الثاني على التصنيف الهرمي الذي يستخدم معلومات لغوية هرمية للهجات الجزائرية. اثبتا كل من النظامين نجاعته مقارنة بالانظمة الأعمال ذات صلة على الرغم من أنها تتعامل مع لهجات من بلدان مختلفة او لهجات المناطق الجزائرية. النتائج المتحصل عليها تبين أن المعلومات حول اللحن أقل تميزًا من الصوتية و لكن تفوقها في الأداء عندما يكون طول التسجيلات قصير. بينما بينت النتائج ان المزج بين المعلومات اللحنية و

الصوتية يحسن في النظام التعرف على اللهجات. بالإضافة، النتائج بينت أن نظام التصنيف الهرمي قد حسن بشكل ملحوظ نظام التعرف على اللهجات.

الكلمات المفتاحيّة: اللهجات الجزائرية، التعليم المعمق، تحديد اللهجة،لحن اللغة، مدونات الصوتية.

Abstract

Dialect IDentification (DID) is a challenging task, becomes more complicated when dealing with under-resourced and an inter-country dialects. Speech database or corpora are crucial for both developing and evaluating accurate DID systems.

The aims of our work are, on the one hand, to build Spoken Arabic Algerian corpora knowing that Algerian dialects are among the most complex Arabic dialects and their study received very title attention. On the other hand, to develop an DID system to identify Arabic intra-country dialects.

Our three contributions are dedicated to Arabic Algerian Natural Language Processing fields. First, concerning language resource, we have building two spoken Arabic Algerian corpora. First, a parallel corpus ALG-DARIDJAH dedicated especially to linguistics studies. It has been collected using a direct recording method which allow more control on the collected data quality. It is a medium size corpus encompasses some sub-dialects spoken in 17 departments with 109 speakers and more than 6K utterances. The second corpus is KALAM'DZ which is a large web-based corpus dedicated to train machine learning systems. Its designed collection and processing recipe is a generic one and it relies on many open source tools. KALAM'DZ corpus encompasses the 8 major dialects with 4881 speakers and more than 104.4 hours.

Our second contribution concerns corpus annotation. In fact, we investigate an altruistic crowdsourcing to validate our semi-automatic dialect annotation of KALAM'DZ corpus. From this experiments, we have determined a list of best practices for altruistic crowdsourcing for corpus annotations.

Finally, we have investigated two solutions to the Algerian Arabic DID (AADID), on intra-country context, relying on the prosodic features and new machine learning techniques. For both systems, we have focused on measuring the discriminative power of the prosody information and deep learning modeling, especially Feed-Forward Neural Network. The first system is based on a flat classification while the second uses the hierarchical one that employs the hierarchy structure of Algerian dialects. The performances of our both systems are comparative to those of literature despite they deal with inter-country Arabic dialects or Algerian dialect areas. The obtained results show that the prosodic features are less discriminative than acoustic one but its showed their superiority in performance when the test utterance sizes are short which is a DID requirements. While, the ADID system based on a combination of prosodic and acoustic information has improved the dialect detection. However, the hierarchical system has significant improvement the ADID.

Keywords: Algerian dialects, Deep Neural Networks, Dialect Identification, Prosody, Speech Corpus.

Résumé

l'Identification Automatique des Dialects (IAD) est une tâche difficile, qui devient plus compliquée lorsqu'il s'agit des dialectes parlés à l'intérieur d'un même pays (intra-pays) et avec des ressources linguistiques insuffisantes. Les corpus orales sont essentiels pour élaborer et évaluer des systèmes performants d'IAD.

L'objectif de notre travail a été d'une part, de construire des corpus parlés de l'arabe dialectal algérien en sachant que les dialectes algériens sont parmi les dialectes arabes les plus complexes et que leur étude n'a pas vraiment reçue l'attention. D'autre part, nous nous sommes intéressés au développement d'IAD systèmes pour identifier les dialectes arabes parlés à l'intérieur d'un même pays.

Nos trois contributions sont dédiées aux domaines du traitement automatique du langage naturel, spécialement l'arabe dialectal algérien. Premièrement, concernant la construction de ressources linguistiques, nous avons construit deux corpus de dialecte arabe algérien parlés. Le premier est, ALG-DARIDJAH, consacré spécialement aux études linguistiques. Il a été collecté à l'aide de la méthode d'enregistrement direct qui permet de mieux contrôler la qualité des données collectées. C'est un corpus de taille moyenne qui englobe certains sous-dialectes parlées dans 17 départements avec un total 109 locuteurs et de plus de 6 K de discours. Le deuxième corpus est KALAM'DZ, un grand corpus collecté depuis le Web. Il est dédié aux systèmes d'apprentissage automatique. La recette de collecte et de traitement est générique et repose sur de nombreux outils open source. Il englobe les 8 majeurs dialectes avec 4881 locuteurs et plus de $104,4$ heures.

Notre deuxième contribution concerne l'annotation du corpus. En fait, nous avons investi le "*crowdsourcing*" altruiste pour valider notre annotation semi-automatique des dialectes du corpus KALAM'DZ. À partir de ces expériences, nous avons établi une liste des meilleures pratiques pour le crowdsourcing altruiste pour les annotations de corpus.

Enfin, nous avons proposés deux solutions à l' IAD Arabe (IADA) algérien, dans le contexte intra-pays. Nous appuyons sur les caractéristiques prosodiques et les nouvelles techniques d'apprentissage automatique. Pour les deux systèmes, nous nous sommes concentrés sur la mesure du pouvoir discriminatoire de la prosodie et de la modélisation de l'apprentissage profond, en particulier les réseaux de neurones à propagation avant. Le premier système est basé sur une classification classique tandis que le second sur classification hiérarchique qui utilise les informations linguistiques hiérarchiques des dialectes algériens. Les performances de nos deux systèmes sont comparable à ceux de la littérature alors qu'ils traitent des dialectes arabes régionales. Les résultats obtenus montrent que les caractéristiques prosodiques

sont moins discriminantes que les caractéristiques acoustiques, mais elles ont démontré leur supériorité lorsque les tailles de discours sont courtes. Cependant, le système ADID basé sur une combinaison d'informations prosodiques et acoustiques a amélioré la détection de dialecte. De plus, les résultats ont prouvé que le système basé sur la classification hiérarchique a nettement améliore la détection des dialectes.

Mots clés: Arabe Dialectal Algérien, Apprentissage Profond, Identification Automatique des Dialects, Prosodie, Corpus Orale.

Publications

Bougrine, S., Cherroun, H., Ziadi, D., Lakhdari, A. & Chorana, A. (2016), Toward a Rich Arabic Speech Parallel Corpus for Algerian sub-Dialects, *in* 'LREC'16 Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT)', pp. 2–10.

Bougrine, S., Chorana, A., Lakhdari, A. & Cherroun, H. (2017), Toward a Web-based Speech Corpus for Algerian Dialectal Arabic Varieties, *in* 'Workshop on Arabic Natural Language Processing (WANLP), Association for Computational Linguistics', pp. 138–146.

Bougrine, S., Cherroun, H. & Abdelali, A. (2017), 'Altruistic Crowdsourcing for Arabic Speech Corpus Annotation', *Procedia Computer Science* **117**, 137 – 144.

Bougrine, S., Cherroun, H. & Ziadi, D. (2017), 'Hierarchical Classification for Spoken Arabic Dialect Identification using Prosody: Case of Algerian Dialects', *arXiv preprint arXiv:1703.10065*.

Bougrine, S., Cherroun, H. & Ziadi, D. (2018), 'Prosody-based Spoken Algerian Arabic Dialect Identification', *Procedia Computer Science* **128**, 9 – 17.

Bougrine, S., Cherroun, H. & Abdelali., A. (2018), Spoken Arabic Algerian Dialect Identification, *in* 'IEEE the International Conference on Natural Language and Speech Processing (ICNLSP)', Algiers, Algeria.

Contents

List of Figures

List of Tables

Abbreviations

AADID	Algerian Arabic DID
AA	Ancient Arabic
ACL	Association for Computational Linguistics
Acous/Phone	Acoustic/Phonetic level
AIP	Association International Phonetics
ALB	ALgiers-Blanks
ALGASD	ALGerian Arabic Speech Database
AMCASC	Algerian Modern Colloquial Arabic Speech Corpus
ANOVA	ANalyses Of VAriance
CA	Classical Arabic
CF	CrowdFlower
CNNs	Convolutional Neural Networks
DA	Dialectal Arabic
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DID	Dialect IDentification
DL	Deep Learning
DNNs	Deep Neural Networks
EGY	EGYptian dialect
ELRA	European Language Resource Agency
FFNN	Feed-Forward Neural Networks
GA	Genetic Algorithms
GLF	GuLF dialect
GMM	Gaussian Mixture Model
GS	Gold Standard
GWAP	Games With A Purpose
HADID	Hierarchical LID
HC	Hierarchical Classification

HIL	HILālī
HMM	Hidden Markov Models
HSA	Hierarchy Structure for Algerian
IAA	Inter-Annotator Agreement
IM	Interval Measures
IPA	International Phonetic Alphabet
IRQ	IRaQi dialect
KACST	King Abdulaziz City of Science and Technology corpus
KSU	King Saud University corpus
LCPN	Local Classifier per Parent Node
LDC	Linguistic Data Consortium
LEV	LEVantine
LID	Language IDentification
LPC	Linear Predictive Coding
MAG	MAGhrebi dialect
MAQ	MA'Qilian
MDP	Multi-Dialect Parallel corpus
MFCC	Mel Frequency Cepstral Coefficient
MSA	Modern Standard Arabic
MTurk	Amazon Mechanical Turk
NLP	Natural Language Processing
OLAC	Open Language Archives Community
Phono	Phonotactic
POS	Part-Of-Speech
PRH	PRe-Hilālī
Proso	Prosodic
PVI	Pairwise Variability Index
RM	Rhythm Metric
RNN	Recursive Neural Network
SAAVB	Saudi Accented Arabic Voice Bank corpus
SAH	Hilālī-Saharan
SAT	SAhel-Tell dialect
SDC	Shifted Delta Cepstrum
SDID	Spoken Dialect IDentification
SLID	Signal-based LID
ST	SemiTones
SUL	SULaymite

SVM	Support Vector Machine
TuDiCoI	TUnisian DIalect COrpus Interlocutor corpus
Tukey HSD	Tukey Honest Significant Differences
UBM	Universal Background Model

Introduction

The field of study that focuses on the interaction between human language and computers is called Natural Language Processing, or NLP for short (Kumar 2011). NLP is a way that uses computers to analyze, understand, and derive meaning from human language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge to perform tasks.

The goals of NLP are to design and build applications that facilitate human interaction with machines and other devices through the use of natural language such as *Machine Translation* used by Google Translate, *Sentiment Analysis* which is oriented to identify subjective information in texts. It can be a judgment, an opinion or an emotional state. *Spam filters* such as those used by Gmail to determine which emails are good and which ones are spam. *Spoken Language Identification* is used to identify the language being spoken by some unknown speaker using a given speech sample.

For most of the front-end NLP researches, language resources are necessary to develop and evaluate NLP systems. Language resources are general term used to refer to corpus, plural corpora. Corpus is a large collection of data, such as written texts or recorded speeches (Crystal 2008). These corpora can be written or spoken only, or both (Lindquist 2009). Moreover, such corpora have to be large to achieve NLP communities expectations especially those working on machine learning based application. There is few corpora for Arabic, which is considered as under resourced language, compared to English corpora, which have started to build since the 1960*s* (Mansour 2013).

Geographically, Arabic is one of the most widespread languages of the world (Behnstedt & Woidich 2013). Arabic has 290 million native speakers (ranked as the 4^{th} language among all) and the total speakers in the word is about 442 million (Paul et al. 2009). Actually, it has two major variants: Modern Standard Arabic (MSA), and Dialectal Arabic (DA). MSA is the official language of all Arab countries. It is used in administrations, schools, official radios, and press. However, DA is the language of informal daily communication. There is

a large number of Arabic dialects and they are grouped in five variants: Gulf, Levantine, Iraqi, Egyptian, and Maghrebi (Versteegh 1997). Recently, DA became also the medium of communication on the Web, in chat rooms, social media etc. This fact, amplifies the need for language resources and language related NLP systems for dialects.

Many NLP fields have been done to study the Maghrebi Arabic dialect(Harrat et al. 2017). In our thesis, we are interested to spoken dialect identification field. Dialect IDentification (DID) is the process in which a dialect is identified automatically. As a task of natural language processing, it uses two types of mediums, speech or text on which the process is carried out. The complexity of dealing with speech is more complex than text, as speech lacks the standardization of text (Chen et al. 2010).

Problematic

For many languages, the state of the art of designing and developing speech corpora has achieved a mature situation. A lot of Arabic corpora focused on MSA variant. This is due to the fact that MSA is spoken by all Arabs, it is more easy to collect data and it has a more standard form than dialects. However, dialect speech corpora have only recently received quiet attention. These corpora are mainly fee-based and the free ones are really rare. Zaghouani (2014), in his recent critical survey on freely available Arabic corpora, mentions that there is only one free corpus, namely Multi-Dialect Parallel (MDP) (Almeman et al. 2013), which gathers MSA and three Arabic dialects (Gulf, Egypt, and Levantine).

Moreover, Harrat et al. (2017) presents an overview of Maghrebi Arabic NLP. They note that the research in this area is at the early stage, which is mainly due to the fact that the dialects of Maghreb region are under-resourced. Let us mention that Algerian dialects are under-resourced: i) We can observe that there is no free corpus dedicated to Algerian dialect variety. ii) There is very few attempts have considered Algerian Arabic dialects. In fact, just Algerian Modern Colloquial Arabic Speech Corpus (AMCASC) Djellab et al. (2017), which have considered 5 accents/dialectal Algerian areas by recording 735 telephone conversations of Algerian speakers from 20 Algerian cities.

Most researches on DID have only been carried out for non-Semitic languages. Whereas, little attention has been paid to spoken Arabic Dialect IDentification (ADID), especially when dialects belong to the same geographical area or country. Djellab et al. (2017) and Hanani et al. (2015) were the only relevant work that proposed ADID systems for Intra-country context. They designed an acoustic/phonetic approaches. They experimented the

approach on Arabic Palestinian accents from four regions, namely: Jerusalem, Hebron, Nablus and Ramallah. Djellab et al. (2017) proposed an approach on a set of data spoken in three Algerian areas: the east, center, and west of Algeria. We highlight the fact that a little attention has been paid to Algerian ADID. In addition, we confirm that there is a lack of ADID system that exploit the prosodic information. In fact, only Rouas et al. (2006) and Biadsy & Hirschberg (2009) have considered using such information. Rouas et al. (2006) investigated dialects within the same area, while Biadsy & Hirschberg (2009) focused on inter-country dialects.

In addition, all of the proposed ADID systems use conventional classification method in occurrence SVM, HMM-GMM, BayesNet. However, Deep Learning is still a novelty in the area. Especially as result of their provided efficiency for Language/Dialect identification (Ferrer et al. 2016, Lopez-Moreno et al. 2016), they are not widely investigated for Arabic NLP tasks.

Objectives and Contribution

In this thesis, our main objective is to perform Algerian dialect identification by investing prosodic information and harnessing deep learning technique. Due to the fact there is no free corpus dedicated to Algerian dialect variety. We have investigated to build language resources. To achieve those objectives, we have tried to response to:

- How to build high quality language resources?
- Have prosodic feature information a discriminative power between dialects?
- What about performance of Hierarchical classification compared to flat classification?

In this thesis, our contributions are dedicated to Arabic Algerian NLP fields. Figure 1 illustrates our contributions, we summarize them in the following points:

- **Building Corpora:** In the first contribution, we have built two spoken Arabic Algerian corpora. First, ALG-DARIDJAH corpus encompasses some sub-dialects spoken in 17 departments with 109 speakers and more than 6 K utterances. Second, KALAM'DZ is web-based corpus, in which the resources are taken from YouTube, Online Radio and TV. It encompasses the 8 major Algerian Arabic sub-dialects with 4881 speakers and more than 104.4 hours segmented in utterances of at least 6 s. We have devised a recipe in order to facilitate building large-scale Speech corpus which harnesses Web resources.

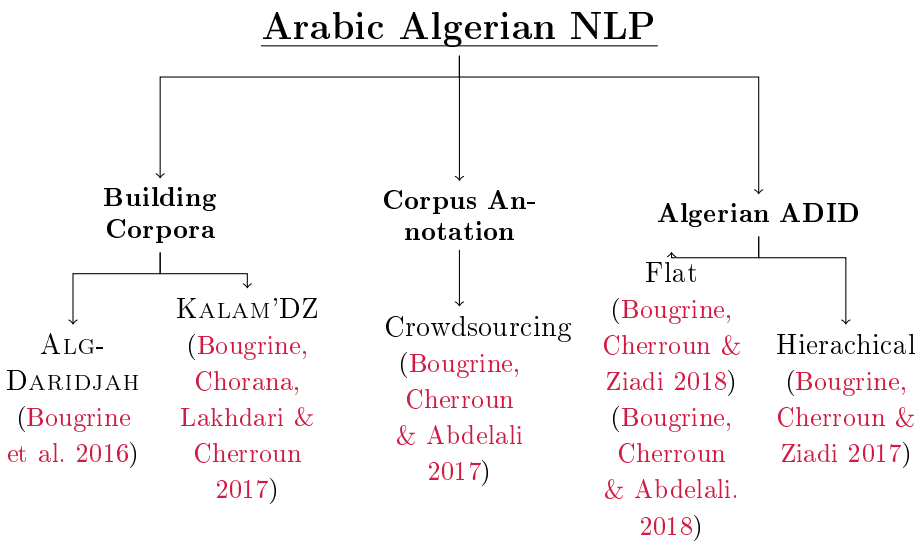


Figure 1: Our Contributions.

- **Corpus Annotation:** In this contribution, we have used the altruistic crowdsourcing technique to annotated the speech corpora. We narrate our experience on the validation of a semi-automatic task for dialect annotation of Kalam'DZ corpus. Furthermore, we have determined a list of best practices for altruistic crowdsourcing for corpus annotations.
- **Algerian ADID:** Concerning the ADID investigation, first we have proposed a flat classification then a hierarchical one to identify spoken Algerian dialects, which is the exploitation of ADID on intra-country context. In the both systems, we focus on measuring the discriminative power of the prosody in Algerian Arabic dialects. Indeed, existing ADID systems rely mainly on knowledge extracted from acoustic/phonetic and phonotactic cues while dismissing prosodic ones in spite of their advantages. In fact, it has been proved that dialect variations are notably pursued in prosodic features (Wells 1982). In addition, Deep Learning is investigated because it improve their efficiency for Language/Dialect identification (Lopez-Moreno et al. 2016).

Dissertation Structure

This thesis is organized in the following way: an introduction to Arabic natural language processing chapter and two main parts.

In Chapter 1, we briefly show the specifications of Arabic language, its dialects, and some related NLP applications. Then, we give a brief overview of Algerian Arabic dialects and their linguistic characteristics.

In Part 'I', through three chapter, we present our contributions related to building and annotating language resources. In Chapter 2, we briefly recall their classification, who collects and distributes it, the key concepts to build it, and the main existing Arabic corpora. Chapter 3 is dedicated to describe the building methodology and the main features of our both spoken Arabic Algerian corpora: ALG-DARIDJAH and KALAM'DZ. In Chapter 4, we review some related work that have employed crowdsourcing for Arabic NLP tasks. Then, we explain the designed altruistic crowdsourcing project and its deployed quality control strategy. Then, from the altruistic crowdsourcing experiments, we extract the list of best practices.

Part 'II' focuses on our contributions linked to spoken dialect identification and our both contribution composed of three chapters. Chapter 5 shows some preliminaries on speech processing and language identification fields. In addition, we review some related work that have dealt with Arabic dialect identification. Prosodic statistic analysis of Algerian dialects are presented in Chapter 6. Our both contributions to solve the problem of Algerian Arabic dialect identification are described and evaluated in Chapter 7.

We recall, in the conclusion, the various works carried out and results obtained. We offer perspectives of research at the end of this work.

Chapter 1

Arabic Natural Language and Its Dialects

In this chapter, we briefly show the specification of Arabic language, its dialects, and some related NLP applications. Then, we give an overview of Algerian Arabic dialects and their linguistic characteristics.

1.1 Arabic Language

According to Ethnologue [1], there are 147 language families in the world (Lewis 2009). The large language family is Afro-Asiatic (Hamito-Semitic) [2], which includes several language sub-families: Semitic, Berber, Chadic, Cushitic, Omotic, and Ancient Egyptian (Coptic). In global, it encompasses about 377 languages and dialects. Figure 1.1 gives the map of Afro-Asiatic languages. The Semitic family consists of 70 languages spoken by more than 467 million people across the Middle East, North Africa, and the Horn of Africa (Pereltsvaig 2012).

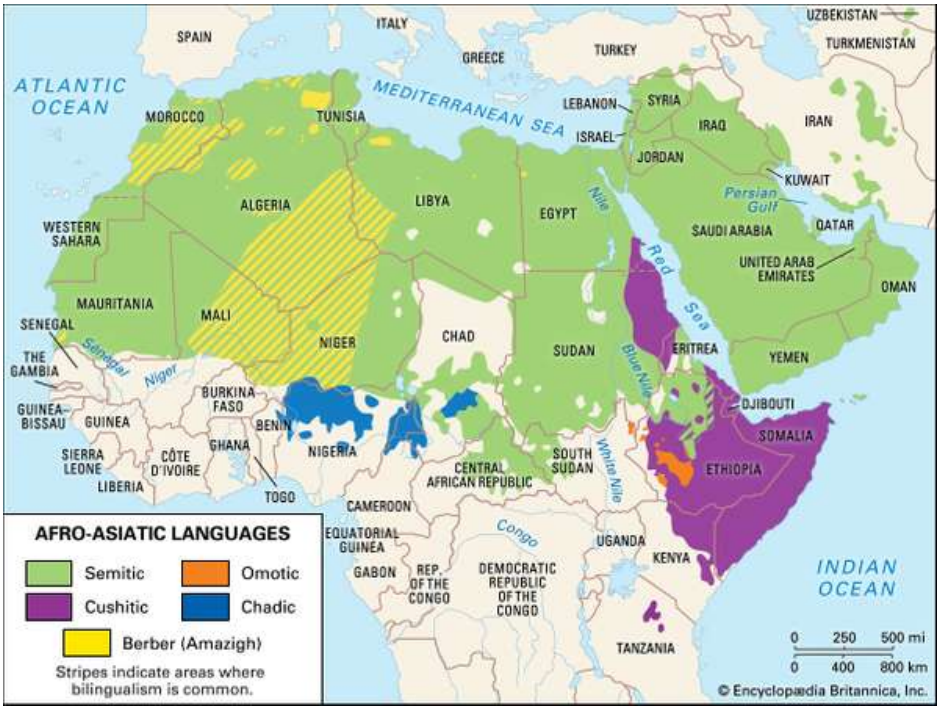


FIGURE 1.1: Afro-Asiatic Languages Map (Ekkehard Wolff 2018).

Arabic (al-'arabiyyah) is one of the classic well-established and the most widely spoken *Semitic* languages (Al-Sharkawi 2016). Arabic language and its varieties are included in *South Central Semitic* group (Pereltsvaig 2012). According to the 2017 edition of Ethnologue

[1] An annual reference of languages of the world that provides statistics and other information on the living languages of the world.

[2] https://www.ethnologue.com/subgroups/afro-asiatic

(Paul et al. 2009), Arabic has 290 million native speakers, which is ranked as the fourth most widely spoken first language among all languages and ranked as the fifth most spoken language in the word with 442 million total speakers. It is spoken in the Middle East and North Africa.

The Arabic language defines the identity of its own people, there is no other language that defining the identity (Suleiman 1994). In the Middle East, Arabic speakers do not originate from the same ethnic background. The Syrians and Lebanese are from *Phoenician* origin while some Egyptians are descendants of the *Pharaohs*. In Iraqi, the most are of *Kurdish* origin, but there is a minority *Assyrian*. The majority of Saudi Arabia and Gulf states population are a *Bedouins* origin. In North Africa, many Moroccans, Algerians, and Tunisians are *Berber* (Farghaly 2010).

Modern Standard Arabic (MSA) is the official language in most of the Arab countries, whereas Dialectal Arabic (DA) varies from one country to another. MSA is sometimes also called العربية الفصحى /al-'arabiya al-fuṣḥā/ (the eloquent Arabic) in order to distinguish it from the dialects. It is used in administrations, schools, official radios, press, and some TV programs. DA is often referred to colloquial Arabic -vernaculars-, which is described collectively with the adjective العامية /al-'āmiya/ (the common) (Al-Sharkawi 2016). It is used in public places, situations of informal communications, and social media. It is borrowed words from European languages, especially French, English, and Italian (Pereltsvaig 2012).

In the rest of thesis, we will write words/phrases in Arabic script. In order to make accessible for who do not read Arabic, we represent the Arabic alphabet by using the transliteration, which means the characters are represented by single Latin letters, often with diacritics. Appendix A reports the characters used in transliteration and their equivalent in International Phonetic Alphabet (IPA) characters has been added. The goal of IPA is to provide an accurate presentation of the speech sounds of all human languages (Harper & Maxwell 2008).

1.2 Arabic Dialects

Islam and the Arab expansions are the most important factors in the development of the Arabic language and its varieties (dialects), and the structures within these varieties (Al-Sharkawi 2016).

There is a large number of Arabic dialects that are geographically classified by researches into five major dialectal areas: Arabian Peninsula, Levantine, Mesopotamian, Egyptian,

and Maghrebi (Versteegh 1997). Their geographical classification is relatively recent compared to the sociological division classifications: nomad Bedouins, sedentary Bedouins, and city-dwellers (Embarki 2008). Diab & Habash (2008) proposed geographical classification illustrated in Figure 1.2.

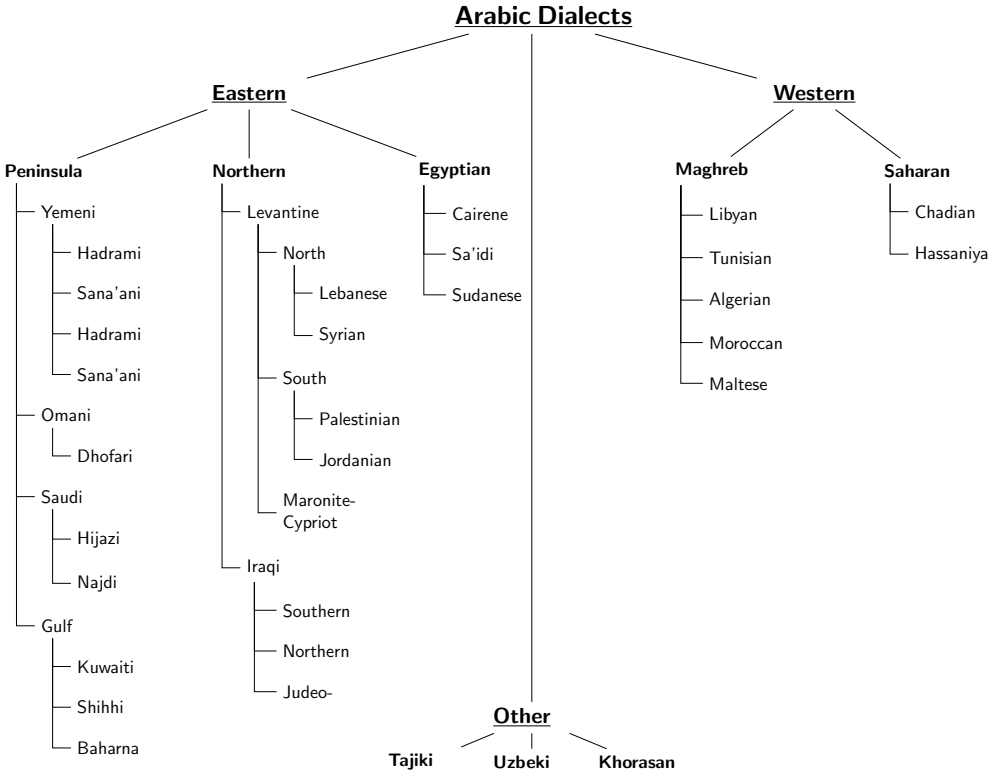


FIGURE 1.2: Geographical Arabic Dialect Classification.

More recently, Habash (2010) and Zaidan & Callison-Burch (2014) also give a very similar geographical classification of spoken Arabic into five dialect groups: Maghrebi (MAG), Egyptian (EGY), Levantine (LEV), Gulf (GLF), and Iraqi (IRQ). Habash (2010) notes that the members of any dialect group are not completely homogenous linguistically. Figure 1.3 illustrates the map of major Arabic dialect groups. Appendix B gives a brief detail of each Arabic dialect group and some of their sub-dialects.

1.3 Natural Language Processing for Dialectal Arabic

Most research in recent years has focused on Natural Language Processing (NLP) for dialectal Arabic due to the wide usage in social media. Shoufan & Alameri (2015) presents a survey

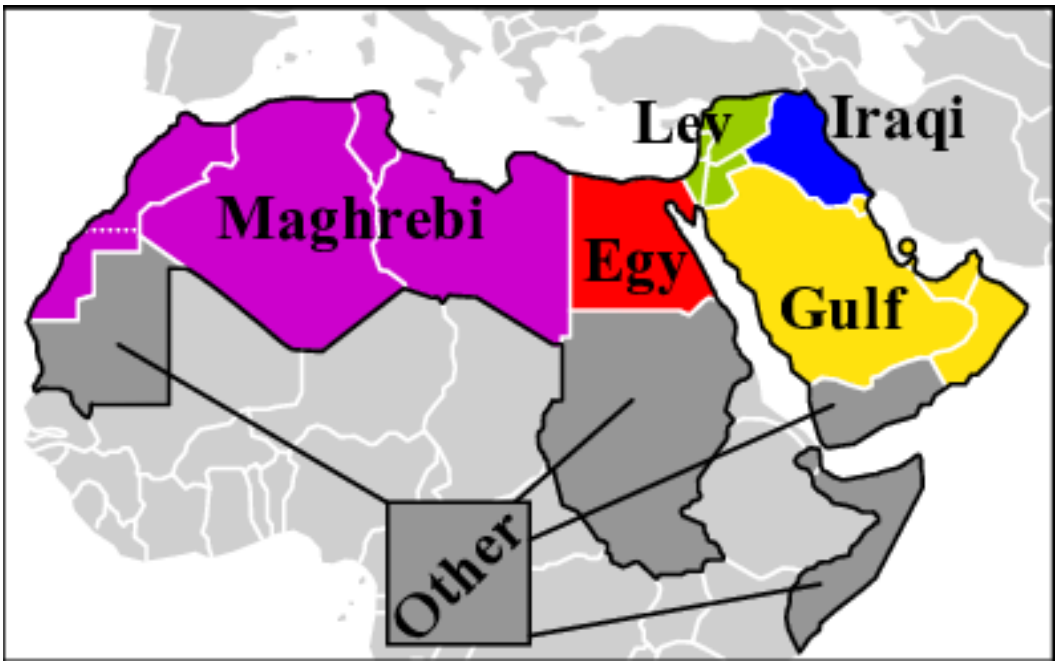


FIGURE 1.3: Arabic Dialect Map, from Zaidan & Callison-Burch (2014); *Other refers to Arabic dialects of Mauritania, Sudan, Yemen, and Somalia.*

on researches that dealing the NLP with Arabic dialects, which can be used to identify the contributions that address to specific NLP for specific dialect. They have categorized the researches that have done in this area in four main categories: basic language analysis, building language resources, semantic-level analysis, and identifying Arabic dialects. As results for their literature review, we can highlight some points:

- Few researchers have been addressed to this area compared to the number of speakers of Arabic dialects, which are spoken by more than 390 million people in total.
- Little attention have been paid to dialects of some large population countries, such as the dialects of Algeria, Morocco, and Sudan. While the Egyptian Arabic is the most studied dialect in Arabic NLP due to the fact that it is spoken in the largest population of the Arabic world.
- A numerous investigations are done for building language resources category due to the fact that Arabic dialects is still under-resourced language.

Recently, Harrat et al. (2017) presents an overview of Maghrebi Arabic NLP. They note that the research in this area is at the early stage, which is mainly due to the fact that the dialects of Maghreb region are under-resourced. In addition, the text dialect identification of Maghrebi Arabic dialects is more researched compared to morphological analysis. While the syntactic analysis is totally ignored.

1.4 Characteristics of Arabic Dialects

In this thesis, we focus on two features selected from the language/dialect information levels: acoustic/phonetic and prosodic information for that in what follows we study the distinct variation of Arabic dialects for these two features.

1.4.1 Phonological Variations of Arabic Dialects

Phonology is a branch of linguistics which studies the sound systems of languages. One of the aims of phonology is to found the patterns of distinctive sound in a language (Crystal 2008). A phoneme is the smallest meaning-distinguishing sound unit in the abstract representation of the sounds of a language (Habash 2010).

For Arabic dialects, there are five major dialectal phonological variations in the pronunciation of vowels and four for consonants: the three dental fricative ث /t̲/, ذ/d̲/, ظ /ḏ̲/ and a voiceless uvular stop ق /q/. Embarki (2008) gives the reflection of these phonological features according to the geo-sociological classification of Arabic dialects. The realization of these features is reported in Table 1.1.

	GLF	IRQ	LEV	EGY	MAG
Nomadic Bedouin	ǵ-ǧ, t̲, d̲, d̲,ī, ū, ē, ō, ā̇, i, u, a	ǧ, t̲, d̲, d̲, d̲, ī, ū, ē, ō, ā̇, i, u, a	k, t̲, d̲, ī, ū, ē, ō, ā, i, u, a	g, s, z, ẓ, ī, ū, ē, ō, ā, i, u, e, o, a	g, t̲, d̲, ḏ̲, ī, ū, ā, i, u, a, ə
Nomadic Sedentary	ǧ-g, t̲, d̲, d̲, ī, ū, ē, ō, ā̇, i, u, a	g, t̲, d̲, d̲, ī, ū, ē, ō, ā̇, i, u, a	ḳ-g, t̲, d̲, d̲, ī, ū, ē, ō, ā̇, i, u, a	k, s, z, ẓ, ī, ū, ē, ō, ā, i, u, e, o, a	g, t, d, ḍ, ī, ū, ā, i, u, a, ə
City-dwellers	ǧ-g, t̲, d̲, d̲, ī, ū, ē, ō, ā̇, i, u, a	q, t, d, ḍ, ī, ū, ē, ō, ā, i, u, a	ʔ, t-s, d-z, ḍ-ẓ, ī, ū, ē, ō, ā, i, u, a	’, s, z, ẓ, ī, ū, ē, ō, ā, i, u, e, o, a	q, t, d, ḍ, ī, ū, ā, i, u, a, ə

TABLE 1.1: The Realization of the Major Phonological Features According Geo-sociological Classification.

In addition to these phenomena, Habash (2010) mentions the following phonological variations:

- The MSA alveolar affricate ج /j/: for instance, جميل ’handsome’ is /jamīl/ in MSA, /gamīl/ in EGY, /žamīl/ in LEV, and /yamīl/ in GLF. For that, the /j/ in MSA is realized as: /g/ in EGY, as /ž/ in LEV and as /y/ in GLF.

- The MSA consonant (ك /k/): is realized as /š/ for GLF, IRQ and the Palestinian rural sub-dialect of LEV. For example, سمك 'fish' is /samak/ in MSA, EGY and most of LEV but /simaš/ in IRQ and GLF.

1.4.2 Prosodic Variants of Arabic Dialects

Prosody is the musical expression of speech that a speaker provides when he speaks. It refers to the variations of physical prosodic parameters, such as the fundamental frequency of the speech signal, the height, intensity and/or duration of the sounds. These variations are perceived by the listener as changes in height (or melody), length, and loudness. The formal representation of the prosodic parameters in oral expression is the intonation, accentuation, rhythm, speech rate, and pause (Crystal 2008, Rouas 2005).

In summary, the linguistic knowledges today which we have about the discriminative power of the prosody in Arabic dialects can be summarized in what follows:

- *Arabic dialects differ in their prosodic structure:* Barkat et al. (1999) evaluated the discriminating power of prosodic pattern in Arabic dialects in a linguistic study. They have shown that the prosodic information can be sufficient to identify Western and Eastern Arabic dialects.
- *Intonation represents a salient discriminative feature in Arabic dialects:* Ghazali et al. (2005) studied the nature of intonation of five Arabic dialects. They observed that the intonation patterns are different between Eastern and Western Arabic dialects. Furthermore, Yeou et al. (2007) confirmed this result using another sample of Eastern and Western Arabic dialects.
- *Arabic dialects present significant differences at the syllable structure:* Syllable structure is closely related to prosodic phenomena, such as rhythm and speech rate. Hamdi (2007) shown that the different types of syllabic structure (or syllable weight) observed in Arabic dialects can be used as a discriminative element. As results, they find that the Moroccan characterized by heavy syllables which short vowels, Lebanon by open syllables with long vowels; Tunisia presents an intermediary tendency. Then, they demonstrated that rhythm variation of Arabic dialects is correlated with syllable structure. Moreover, Bouziri et al. (1991) have early confirmed this fact by studying the stress measured through the syllable structure.
- *Rhythm and intonation are discriminant parameters for some Algerian dialects:* Benali (2004) studied the role of the rhythm and the intonation in a human identification by

means of two Algerian dialects, which are spoken in Algiers and Oran. He noted that rhythm and intonation are very discriminant parameters, particularly the speech rate and the variation of fundamental frequency.

1.5 Typology of Algerian Arabic Dialects

As mentioned previously, we investigate to Arabic Algerian NLP fields for that we focus on Algerian dialects and their characteristics.

Algeria is a large country administratively divided into 48 departments. Its first official language is MSA. However, Algerian sub-dialects are widely the predominant means of communication. Algerian Arabic is used to refer to the dialect spoken in Algeria, known as Daridjah to its speakers. Algerian dialect presents complex linguistic features mainly due to both Arabization processes that led to the appropriation of the Arabic language by population from Berber origin, in addition to the deep colonization. In fact, Arabic Algerian dialect is affected by other languages such as Turkish, French, Italian, and Spanish (Leclerc 2012).

According to the Arabization process, dialectologists show that Algerian Arabic dialects can be divided into two major groups: Pre-Hilālī and Bedouin dialect. Both dialects are different by many linguistic features (Caubet 2000, Marçais 1986).

Pre-Hilālī dialect is called sedentary dialect. It is spoken in areas that are affected by the expansion of Islam in the 7^{th} century. At this time, the partially affected cities are Tlemcen, Constantine, and their rural surroundings. The other cities have preserved their mother tongue language (Berber). Marçais (1977) has divided Pre-Hilālī dialect into two dialects: village (mountain), and urban dialect.

- Village dialect is located between Trara mountains and Mediterranean sea. The central town of this area is Nedroma, which is located in the northwestern corner of Algeria. There is also a village dialect located in the northeastern corner of Algeria: it is between Collo, Djidjelli, and Mila.
- Urban dialect is located in the northern cities: Tlemcen, Cherchell, Dellys, Djidjelli, and Collo.

Bedouin dialect is spoken in areas which are influenced by the Arab immigration in the 11^{th} century (Palva 2006, Pereira 2011). In Marçais (1986), the authors has divided Bedouin dialect into five distinct basic dialects:

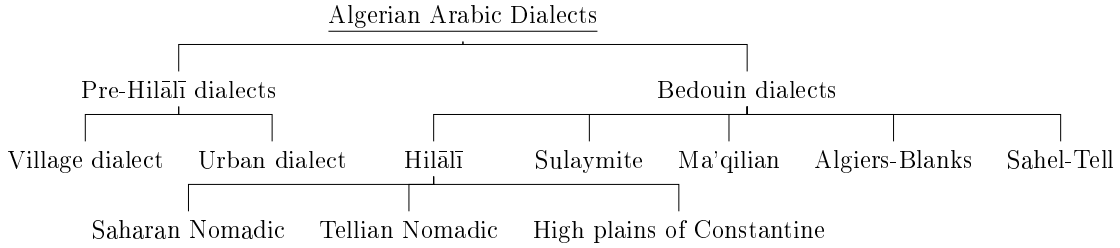


FIGURE 1.4: Hierarchy Structure for Algerian Dialects.

1. Bedouin dialect of eastern Constantine, which is located in the region of El Kala and Souf. It is called 'Sulaymite' dialect because it is connected with Tunisian Bedouin dialects.

2. Bedouin dialect of the central and western side of Oran. It is called Ma'qilian dialect because it is connected with Moroccan Bedouin dialects. It covers a part of the arrondissement of Tlemcen, Oran, Sidi Bel Abbès, and Saïda.

3. Bedouin dialect of the center of Algeria and Sahara. It is called Saharan Nomadic, it covers almost the totality of the Sahara of Algeria, towards the east to Oued Righ, towards the south to the Tademaït plateau, and until the west (its limit has not been clarified).

4. Bedouin dialect of the Tell and of the Algerian-Oran Sahel. It is called Tellian Nomadic, that occupies a large part of the Tell of Algeria: Bordj Bou Arréridj, Sétif, El-Eulma, and El Kantara.

5. The dialect of the high plains of Constantine, which covers the north of Hodna region, and extends to the rough area from Bordj Bou Arréridj to Seybouse river.

In (Marçais 1986), the author gathered the three dialects (3, 4, 5) under the name of Hilālī dialect, which takes its name from Banū Hilāl tribe. In addition to that, there are other dialects which has are originally urban dialect and have been completely influenced by Bedouin dialect. For that, we have classified these dialects, so-called Urban Completely Bedouin (UCB). Regarding some linguistic differences, we have divided this last dialect into two sub-dialects, namely Algiers-Blanks and Sahel-Tell.

In spite of the sparse and the specific linguistic studies that have dealt with Arabic dialects, there is no efficient and complete dialect hierarchy dedicated to Algerian dialects. We have compiled a preliminary version of such hierarchy from the above historical knowledge. We summarize this Hierarchy Structure for Algerian (HSA) dialects in Figure 1.4. Parts of this

HSA is also confirmed by some linguistic studies essentially phonological, lexical and morphological (Marçais 1986, Palva 2006). However, this hierarchy structure has not benefited from the deep prosodic analysis. In what follows, we review the phonetic and phonological data of Algerian Arabic dialects.

Embarki (2008) resumes the descriptions proposed in different bibliographic sources, to identify the 'major' discriminating phonological units for different types of Arabic dialects. He is based mainly on the presence or absence of some phonemes of MSA in Arabic dialects, these phonemes are four consonants: the dentals fricative /$\underline{t}$/ ('ث'), /$\underline{d}$/ ('ذ'), /$\underline{\dot{d}}$/ ('ظ') and a voiceless uvular stop /q/ ('ق'). For Algerian dialects classification, Versteegh et al. (2006) have previously used these four consonants to discriminate the two major groups: Pre-Hilālī and Bedouin dialect, where Pre-Hilālī dialect characterized by: /q/ is pronounced /k/ ('ك'), except in Tlemcen where it is a glottal stop /'/ ('أ'), and the loss of inter-dentals and pass into the dentals /t/ ('ت') , /d/ ('د'), /ḍ/. For Algerian Bedouin dialect, the four discriminative consonants are characterized by: /q/ ('ق') is pronounced /g/ ('ڤ') and the inter-dentals are fairly preserved.

Early, Marçais (1986) presents other discriminative phonemes for two major dialect groups and adds four discriminative features. For Pre-Hilālī dialect, the phonemes /k/ ('ك'), for instance كلب /keleb/ 'dog', is pronounced as: a palatal [k^{y}][3], or an affricate [k^{ʃ}, t^{ʃ}][4], or as a fricative [ʃ][5]; /t/ is pronounced as an affricate [t^{s}]; /j/ becomes [ʒ] when it is single, and [dʒ][6] when it is doubled. In addition, the diphthongs /ay/ and /aw/ usually become /ī/ and /ū/. For example, let us take the words: حيط /ḥayṭ/ 'wall' becomes /ḥīṭ/, لون /lawn/ 'color' becomes /lūn/.

The Bedouin dialect is characterized by the preservation of diphthongs /ey/ and /ow/ or their contraction /ē/ and /ō/. Each group have other dialect features added to the four discriminative consonants, which are presented in Table 1.2.

1.6 Conclusion

This chapter began by exposing the position of Arabic language and its varieties according to the structure of the language family in the world. Second, we reported the geographical

[3] كيلب /kyelb/

[4] تشلب /tshelb/

[5] شلب /shelb/

[6] As in the word نجاح /ndjaḥ/ 'success'

Dialect	'ج' /j/	'غ' /ġ/	Diphthongs /ey/ and /ow/	The Final Vowel
Sulaymite	/j/	/ġ/	Reduced to /ē/ and /ō/	/y/ becomes /i/
Ma'qilian	/j/	/ġ/	Presented/Reduced to /ē/ and /ō/	
Saharan Nomadic	/j/	'ق' /q/	Presented/Reduced to /ē/ and /ō/	
Tellian Nomadic	/dj/	/ġ/	Presented/Reduced to /ī/ and /ū/	/u/ becomes /o/
High plains of Constantine	/dj/	/ġ/	Reduced to /ī/ and /ū/	

TABLE 1.2: Discriminative Consonants of Bedouin Dialects.

classification of Arabic dialects as there is a large number of Arabic varieties. For each group, we described the main features and its sub-dialects. Third, we have reported some remarks on NLP applications dealing with Arabic dialects and special Maghrebi group. Then, we showed some characteristic of Arabic dialects, such as phonological and prosodic varieties. Finally, we described the Algerian Arabic dialects and its phonological varieties.

Part I

Language Resources

Chapter 2

Preliminaries

Contents

This chapter gives an overview of concepts and generalities on language resources. In fact, we will speak about the main characteristics that feature the spoken corpora.For the purpose of this thesis, we will review the main existing Arabic corpora.

2.1 Language Resources

A general term used to refer to language resources is corpus, plural corpora, which is a large collection of data, such as written texts or recorded speech (Crystal 2008). These corpora can be written or spoken only, or both (Lindquist 2009).

Corpus data may be used by both linguistics and Natural Language Processing (NLP) researchers. In the field of NLP, the use of corpora is necessary for most of the front-end NLP researches. In fact, they are important to build and evaluate NLP systems specially those based on machine learning where they will use for training and evaluating models. Among them, we cite machine translation, speech recognition, speech synthesis, speaker recognition/verification, spoken language, etc.

In the field of linguistics, corpora are also very useful to study the specific phenomena of linguistic such as lexical ambiguities, which is the study of the presence of two or more possible meanings within a single word, fixed expressions which refers to the study of the standard form of expression that has taken on a more specific meaning than the expression itself, etc. Moreover, corpora are used more in other disciplines such as cognitive science, which is the study of the mind and its processes, or foreign language teaching that refers to the methods of teaching to acquire a non-native language (Kurdi 2016).

2.2 Corpus Classification

In order to establish a corpus classification, many criteria can be used, such as the distinction between spoken and written corpora, the number of languages or dialects in the corpus, thematic representativity of data in corpora ...

Written versus Spoken Corpus

Corpus can be categorized according to the type of the collected data, which is written texts, spoken records, or both. Written texts in corpora might be drawn from books, web pages,

or newspaper archives, such as The Guardian (for English), Le Monde (for French) and Al-Hayat (for Arabic). Often, the social media websites, such as twitter and facebook, and the personal website (blogs) are recently a very popular source for written texts. These texts are scanned or downloaded electronically. The size of written corpus depends on the specific reasons of its building. For example, the preparation of a language dictionary required a very large written corpus.

The spoken corpus (or speech corpus) could in principle include a set of audio recordings only. For example, the CALLHOME Egyptian Arabic corpus contains only the telephony conversations between native speakers of Egyptian Arabic (Canavan et al. 1997). At the other hand, the spoken corpus can include the audio transcription such as orthographic or phonetic ones. For example, Arabic Speech Corpus[1] contains the audio recording, the orthographic, and the phonetic transcription of the south Levantine Arabic dialect (Damascian accent) (Halabi 2016).

Corpus Language

A corpus can be expressed in one or more languages. For that, we can distinguish between: monolingual corpora, multilingual corpora, or parallel corpora. Monolingual corpora are corpora whose content a single language. For example, The CALLHOME Egyptian Arabic corpus contains only Egyptian Arabic (Canavan et al. 1997). Multilingual corpora content a different languages. For example, the KSU Rich Arabic corpus is collected for nine Arabic dialects and MSA (Alsulaiman et al. 2013). Parallel corpora is gathering one language and their translations into one or more other languages. For example, Multi-Dialect spoken parallel corpus encompass MSA and their translations into three Arabic dialects: Gulf, Egypt and Levantine (Almeman et al. 2013).

Corpus Data Representativity

The corpus can also be classified according to the data representativity that form them. There are three types of representativity: balanced, pyramidal, or opportunistic corpus. Balanced corpus gathered the equally data for each language or topic to guarantee the representativeness. Pyramidal corpora include data distributed by levels, which built for central topics large collections and for less important topics small collections. Opportunistic corpora are

[1] Is free corpus which it can be downloaded at http://en.arabicspeechcorpus.com/

used for under-resourced language. Therefore, it gathers all available resources and not take on consideration data representativeness.

2.3 Spoken Corpus Usage

In what follows, we focus only on Spoken Corpus (SC). Let as mention that, each SC creation is designed for specific purposes, which determine the content and the design of a corpus. In fact, it can be used for research purposes or for real life technological applications (Gibbon et al. 1997). Let us mention that some corpora created for specific field they can be used for others fields.

Concerning SC for research purposes on linguistics, such the study of specific hypothesis about speech production, the researchers are more exigent. In fact, they need that word are well pronounced. These type of corpora are used for the study of *phonetic*; which is the study of all aspects of speech; *sociolinguistic*; which is the study of the relationship between language and society; *psycholinguistic* research; is the study how we develop, perceive, and produce language.

Technological applications that use SC can be divided into four classes: speech synthesis, speech recognition, spoken language systems, and speaker recognition/verification. Each application needs its specific corpus design. For example, speech synthesis requires large speech data from only one or two speakers, whereas speech recognition requires small speech data from many speakers.

2.4 Spoken Corpus Specification

This section gives an overview about the speech corpus specification that characterize a corpus from another, such as: the spoken content (topics), speaker profiles, number of speakers, speaking style, recording materials and environment ...

The Spoken Content

The spoken content is what is spoken in a speech corpus. It is the first major feature which determines the usage of the resource. To define the content of a corpus, there are four

methods, which can be applied alone or mixed. It can be characterized by vocabulary, by domain, by task, or by phonological distribution.

The simplest way to specify the spoken content is by *vocabulary*, which is derived automatically from the usage of the resource. For instance, Saudi Accented Arabic Voice Bank (SAAVB) corpus (Alghamdi et al. 2008) is characterized by the vocabulary richness, which is the presence of high percentage of unique words in the transcription files. Indeed, it is designed for speech recognition systems. SAAVB corpus vocabulary includes: numbers, city names, currencies, yes/no answer, spelling of a phonetically rich words.

Another method for controlling the contents of a speech corpus collection is by *domain*. It is method that achieved a closed vocabulary without restricting the speakers. The domain could be for instance: TV program, weather, and restaurants. For example, TuDiCoI (Graja et al. 2010) corpus contains recorded dialogs between staff and clients in the railway services of Sfax town, Tunisia. Mainly usage is for automatic answer/response systems.

Solving a task/question can be another method to collect the spoken corpus. For example, SAAVB corpus (Alghamdi et al. 2008) includes the spoken answers of the question "Where is the nearest hospital?".

In some cases the contents of a speech corpus have to be specified in term of *phonological* units, like phonemes, syllables, and morphemes. For instance, a general purpose dialect recognition system will require the repetitions of discriminative phonemes. For example, SAAVB corpus (Alghamdi et al. 2008) includes the record of two accent variation sentences, in which the discriminative phonemes are presented.

Number of Speakers and How They are Selected

The number of speakers is one of the most important characteristics of a spoken language corpus. Speakers can be men, women, or children, dependent on the application. Speech corpora can be divided into three classes (Gibbon et al. 1997):

- Speech corpora, with few speakers (1 to 5), are often used to prepare phonetic elements dictionaries (allophones, diphones, etc.), to design prosodic models, for developing synthesis systems, or for basic speech research. Note that a few number of speakers does not necessarily mean a small corpus.
- Speech corpora, with many speakers (about 5 to 50), are often used in experimental factorial research. In general, the number of speakers and the number of repetitions of

the speech phenomena that are investigated to allow planned statistical tests to reach a pre-specified power.

- Speech corpora with large number of speakers (more than 50) are necessary to train and test speech recognition, speaker verification, and dialect identification systems.

In addition, one of the major feature that fixes the corpus usage is *How the speakers are selected* ? *Which means that are-they arbitrary selected*? or *with respect to some characteristics (speaker profiles)*? It is importance to specify the characteristics and distributions of speakers. The profile of speaker includes such information: distribution of sex and age, mother tongue, social factors (education/ proficiency/ profession), dialectal distribution. For instance, Modern Standard Arabic speech corpus (Abushariah et al. 2012) is recorded by 40 speakers (20 male and 20 female), which are native speakers differing in gender, age, country, geographical region (three major regions: Levant, Gulf, and Africa), profession, educational background, and mastery of Arabic language.

Speaking Style

Speaking style is another feature key that defines the possible uses of the speech corpus. It can be *read*, *answering*, *description*, or *spontaneous speech*. For a corpus production, it is strongly recommended at least to have two different speaking styles. To obtain a specific vocabulary, often answering[2] and descriptive[3] styles are used. Descriptive speech is more spontaneous than both read and answering style. In contrary to read, answering, and description style, in spontaneous speech, the speaker has no restrictions of the supervisor on his speech aside from a topic or a task. Although, most speech corpora contain read style.

Environmental Conditions and Recording Materials

The speech can be recorded in different types of acoustic environments, such as: sound proof room, quiet office, office with 1/2/5/10 employees, corridor, restaurant, street, inside vehicle, quiet living room (furniture, open/closed windows), etc. In addition, data can be collected with different recording types of microphones and transmission channels, such as: single microphone, set of microphones with different quality, mobile phone, land phone, etc.

[2]Answering speech style: covers all recordings that are prompted by a question.
[3]Descriptive speech: is asking the speaker to describe the showing picture or graph.

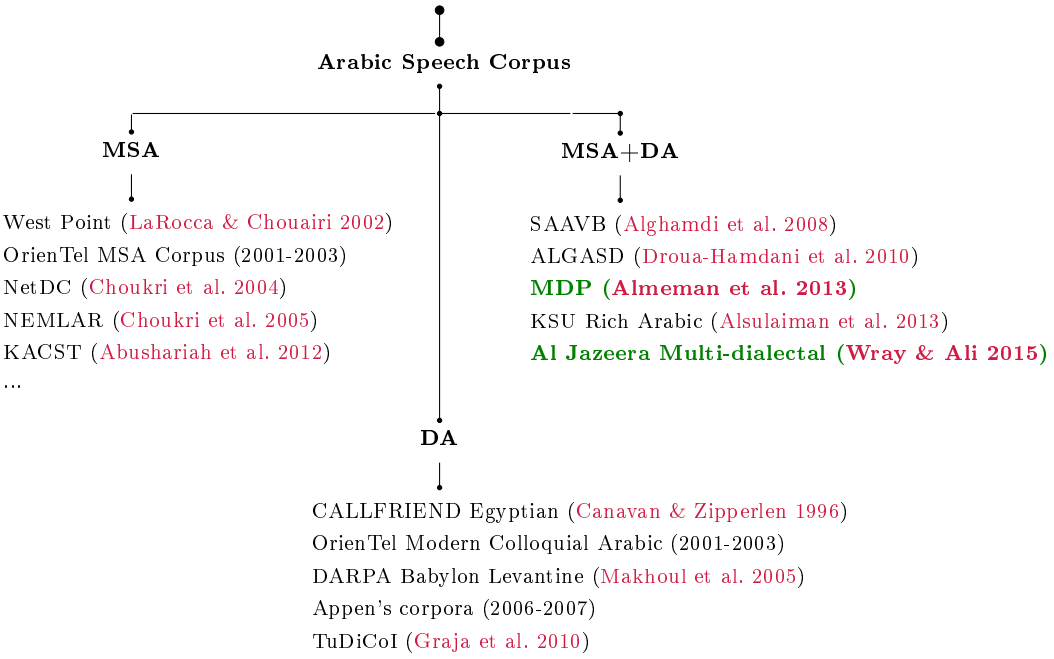


FIGURE 2.1: Main Spoken Arabic Corpora.

2.5 Arabic Spoken Corpus

For many languages such as English, speech corpora has achieved a mature situation compared to Arabic, which can still be considered a relatively resource-poor language. In this section, we review the speech corpora dealing with Arabic variants. We have classified most important Arabic Spoken Corpora according to the fact that they are built for Modern Standard Arabic (MSA) or Dialectal Arabic (DA), or both of them. Figure 2.1 illustrates the studied existing corpora according to this taxonomy where the green color is used to indicate free corpus. Following another classification way, we have classified them according to two criteria: *collecting method* and *Intra/Inter country* dialect collection context. They can be classified into five categories according to the collecting method. Indeed, it can be done by recording broadcast, spontaneous telephone conversations, telephone responses of questionnaires, direct recording, and web-based resourcing. The second criterion distinguishes the origin of targeted dialects in either Intra-country/region or Inter-country, which means that the targeted dialects are from the same country/region or dialects from different countries.

First, let us briefly review some MSA corpora. It is important to mention that most existing corpora are collected early in 2000s. For instance, LaRocca & Chouairi (2002) built the West Point Arabic Speech corpus, which is collected by direct recording method of *read speaking* style using microphone. It is available via Linguistic Data Consortium (LDC) catalogue[4].

[4]Code product is LDC2002S02.

The speech data are collected from 110 speakers, which are 75 speakers native (41 males) and 35 speakers non-native (25 males). The total size of this corpus is about 11.42 hours.

In addition any other, speech corpora for MSA are collected by using telephone response of questionnaire through the *OrienTel* [5] project funded by European Commission. It focuses on the development of language resources for speech-based telephony applications across the area between Morocco and Gulf States. These corpora are uttered by speakers from Egypt, Jordan, Morocco, Tunisia, and United Arab Emirates countries, which are available via the European Languages Resources Association (ELRA) catalogue[6].

Concerning Standard Arabic corpus collected by recording Broadcast News from radio by using receiver, we can cite *NetDC Arabic* (Choukri et al. 2004) and *NEMLAR*[7] (Choukri et al. 2005) corpus, with total sizes are about 21 and 40 hours respectively. Both corpora are collected by ELRA organization.

More recently, King Abdulaziz City of Science and Technology (KACST) (Abushariah et al. 2012), an Arabic MSA phonetic database, is collected by using direct recording. It encompasses speeches from eleven Arab countries. They are gathered to represent three different Arab regions: Levant, Gulf, and Africa. The speakers read 415 sentences, which are phonetically rich and balanced.

In contrast to this relative abundance of speech corpora for MSA, very few attempts have tried to collect Arabic Speech corpora for dialects. As far as we know, the pioneer Dialectal Arabic (DA) corpus has been collected in the middle of the nineties and it is *CALLFRIEND Egyptian* (Canavan & Zipperlen 1996). It contains 60 conversations between 5-30 minutes.

Another part of *OrienTel* project, cited below, has been dedicated to collect speech corpora for Arabic dialects of Egypt, Jordan, Morocco, Tunisia, and United Arab Emirates countries. In these corpora, the same telephone response of questionnaire method is used. These corpora are available via ELRA catalogue[8].

The *DARPA Babylon Levantine* Arabic speech corpus gathers some Levantine dialects collected by using direct recording of spontaneous speech method (Makhoul et al. 2005). It is spoken by speakers from four Arab countries: Jordan, Syria, Lebanon, and Palestine. The size of that corpus is about 45 hours, and it is available via LDC catalogue[9].

[5] http://www.speechdat.org/ORIENTEL/

[6] Respective code products are ELRA-S0222, ELRA-S0290, ELRA-S0184, ELRA-S0187 and ELRA-S0259.

[7] Code product: ELRA catalog ELRA-S0219.

[8] Respective code products are ELRA-S0221, ELRA-S0289, ELRA-S0183, ELRA-S0186, and ELRA-S0258.

[9] Code product is LDC2005S08.

Appen company[10] has collected three Arabic dialects corpora by means of spontaneous telephone conversations method. These corpora uttered by speakers from Gulf (975 conversations about 93 hours), Iraqi (474 conversation about 24 hours), and Levantine (982 conversation about 90 hours). They are available via LDC catalogue[11].

With a more guided telephone conversation recording protocol, *Fisher Levantine Arabic* corpus is collected, it is available via LDC catalogue[12]. The speakers are selected from Jordan, Lebanon, Palestine, Lebanon, Syria, and other Levantine countries.

TUnisian DIalect COrpus Interlocutor (TuDiCoI) (Graja et al. 2010) is a spontaneous dialogue speech corpus dedicated to Tunisian dialect. It contains recorded speeches of 127 dialogues which represent 893 utterances. The speech data is dedicated to the travel domain. In fact, all speeches are conversations recoded between train staff and clients in the railway of Sfax town, Tunisia.

Concerning corpora that gather MSA and Arabic dialect, we have reviewed some of them. Saudi Accented Arabic Voice Bank (SAAVB) corpus is dedicated to speakers from all the cities of Saudi Arabia country using telephone response of questionnaire method (Alghamdi et al. 2008). The main characteristic of this corpus is that before recording, a speaker and environment selections are performed. The selection aims to control speaker age and gender in addition to the telephone type.

In an intra country context, ALGerian Arabic Speech Database (ALGASD) corpus is collected by direct recording of MSA sentences and dialect sentences read by Algerian speakers (Droua-Hamdani et al. 2010). The speakers are selected from 11 regions from Algeria. This corpus contains 300 speakers and the total number of utterances is 1080. They have 600 recordings of dialect sentences (repetition of only both dialect sentences) and 480 MSA sentences (with 136 of the individual sentences).

Multi-Dialect Parallel (MDP) (Almeman et al. 2013) is among the rare free corpora, which is confirmed by Zaghouani (2014) in his recent critical survey on freely available corpora. It gathers MSA and three Arabic dialects. Namely, the dialect are from Gulf, Egypt, and Levantine. The speech data are collected by domain (travel and tourism) and recording is performed by direct method. The number of speakers is about 52 and the percent of dialect utterances is about 77% of total size corpus (32 hours).

[10] Appen is a global company and leader in high-quality, human-annotated datasets for machine learning and artificial intelligence.

[11] Respective code products are LDC2006S43, LDC2006S45, and LDC2007S01.

[12] Code product is LDC2007S02.

King Saud University (KSU) Rich Arabic (Alsulaiman et al. 2013) corpus encompasses speakers by different ethnic groups, Arabs and Non-Arabs (Africa and Asia). Concerning Arab speakers in this corpus, they are selected from nine Arab countries: Saudi, Yemen, Egypt, Syria, Tunisia, Algeria, Sudan, Lebanon, and Palestine. This corpus is rich in many aspects, among them, richness of the spoken content by vocabulary, by solving task, by phonological distribution, and speaking style (read, answering, description, and spontaneous). In addition, different recording sessions, environments, and systems are taken into account.

Al Jazeera multi-dialectal speech is a free corpus with total size about 57 hours, based on Broadcast News of Al Jazeera (Wray & Ali 2015). Its annotation is performed by crowd sourcing technology. It encompasses MSA and four major Arabic dialectal categories: Egyptian, Levantine, North African, and Gulf.

In an intra country context, there are only Algerian Modern Colloquial Arabic Speech Corpus (AMCASC), is a corpus dedicated to Algerian Arabic dialect varieties (Djellab et al. 2017), which is based on telephone conversations collecting method, is a large corpus about 735 speakers and more than 72 hours that takes three regional dialectal varieties.

Table 2.1 summaries our review of these major spoken Arabic corpora specially those dedicated to dialects. This study leads us to make some remarks:

- We should mention the fact that these corpora are mainly fee-based and the free ones are extremely rare. We can enumerate only two free corpora: *MDP* (Almeman et al. 2013) and *Al Jazeera Mutli-dialectal* (Wray & Ali 2015).
- Let us also observe that the direct recording method combined with a chosen text, is less used in spite of that it exhibits better the language/dialect features.
- For some dialects, especially Egyptian and Levantine, there are some investigations in terms of building corpora and designing NLP tools. While, very few attempts have considered Algerian Arabic dialect. Which, make us affirm that the Algerian dialect and its varieties are considered as under-resourced language.
- In an intra country context, only AMCASC (Djellab et al. 2017) corpus gathered Algerian Arabic dialect varieties.
- The best of our knowledge, there is no Web-based speech dataset/corpus that deals with Arabic speech data neither for MSA nor for dialects. While for other languages, there

are some investigations. We can cite the large recent collection Kalaka-3 (Rodríguez-Fuentes et al. 2016). This is a speech database specifically designed for Spoken Language Recognition. The dataset provides TV broadcast speech for training, and audio data extracted from YouTube videos for testing. It deals with European languages.

2.6 Conclusion

In this chapter, we have first given some preliminaries on language resources such their classification aspects: the type of data (spoken or read), the number of treated languages(monolingual, multilingual, or parallel), and the data representativity (balanced, pyramidal, or opportunistic). Second, we present who collects and distributes corpora. Third, we show both usage of spoken corpora, which are used for research purpose and for many technological application. Then, we give quasi-exhaustive overview about the specification of speech corpus. Finally, we attempt to provide a brief summary of the literature relating to Arabic spoken corpus. We note from this review that no web-based spoken Arabic corpus and the Algerian dialects are under-resourced languages. For that, the next chapter describes our first contribution: two Algerian dialects corpus, one of them is Web-based corpus.

Corpus	Speakers	Dialect	Content	Channel	Sampling rate	From
CALLFRIEND	-	EGY	Spontaneous	Telephone	8 KHz	LDC
OrienTel Morocco	772	MOR	Response of questionnaire	Telephone	8 KHz	ELRA
OrienTel Tunisia	792	TUN	Response of questionnaire	Telephone	8 KHz	ELRA
OrienTel Egypt	750	EGY	Response of questionnaire	Telephone	8 KHz	ELRA
OrienTel UAE	880	Emirates	Response of questionnaire	Telephone	8 KHz	ELRA
OrienTel Jordan	757	JOR	Response of questionnaire	Telephone	8 KHz	ELRA
DARPA Levantine	164	JOR, PAL, SYR, LEB	Spontaneous	Microphone	16 KHz	LDC
Gulf Appen	975	GLF	Spontaneous	Telephone	8KHz	LDC
Iraqi Appen	474	IRQ	Spontaneous	Telephone	8KHz	LDC
Levantine Appen	982	LEV	Spontaneous	Telephone	8KHz	LDC
Fisher Levantine	-	JOR, LEB, PAL, SYR	Spontaneous	Telephone	8KHz	LDC
TuDiCoI	127	TUN	Spontaneous	-	-	ANLP-RG
SAAVB	1033	Saudi	read, spontaneous, answering	Telephone	8KHz	KACST
MDP	52	GLF, EGY LEV	Translating	Microphone	8KHz	MISC
AMCASC	735	3 ALG	Spontaneous	Telephone		MISC
KSU Rich Arabic	288	Africa & Middle East	Spontaneous, answering, read	Mobile, microphone	16 KHz	KSU
Al Jazeera multi-dialectal	-	EGY, LEV, GLF	Spontaneous	Microphone	16KHz	QCRI

Table 2.1: Speech Corpora for Arabic Dialects. (ALG= Algerian, EGY= Egyptian, GLF= Gulf, IRQ= Iraqi, JOR= Jordanian, LEB= Lebanese, LEV= Levantine, MISC= Miscellaneous, MOR= Moroccan, PAL= Palestinian, SYR= Syrian, TUN= Tunisian)

Chapter 3

Our Arabic Algerian Dialects Corpora

Contents

For many languages, the state of the art of designing and developing speech corpora has achieved a mature situation. On the other extreme, there are few corpora for Arabic. Moreover, very few attempts have considered Algerian Arabic dialect. This chapter is dedicated to describe the building methodology and the main features of our both spoken Arabic Algerian corpora: Alg-Daridjah and Kalam'DZ.

3.1 ALG-DARIDJAH Corpus

First, we have created **Alg-Daridjah** corpus dedicated to Algerian dialect variety. We have baptised our corpus **Alg-Daridjah** for **ALG**erian **Darijah**. In the Maghreb, the term Daridjah 'الدارِجَة' means dialectal Arabic. In the Est, they use the vocable 'العامية' /al-ʿamiya/.

3.1.1 Methodology

The crucial points to be taken into consideration when designing and developing relevant speech corpus are numerous. We can mention some of them: scope and the size of the corpus, richness of speech topics and content, number of speakers, gender, regional dialects, recording environment and materials. We have attempted to cover a maximum of these considerations. We will underline each considered point in what follows.

In order to build our dialect corpus, we have to resolve and fix some choices, namely: *Which text (words, sentences, and paragraphs) we have to put in the corpus that foster its richness? Which method to use in order to collect speech data?* and finally *Which speaker profile we have to adopt?*

First, let us recall that collecting dialect speech data can be done by means of four ways: straightforward recording some reports and telecasts from regional radios and TV, by telephone response of questionnaire, by spontaneous telephone conversations, or by direct recording. Obviously, it can be performed also by combining at least two ways.

In our case, we have selected *Direct Recording* method to collect speech data in order to provide a rich corpus that can be used to display differences and similarities between dialects. Furthermore, a direct recording compared with broadcast news-based method, allows us to control speaker's profile.

3.1.1.1 Corpus Text Content

Concerning the speech material, we have adopted a handmade text. This latter with selected and guided content allows the corpus to be parallel. So, it allows highlighting fairly dialect features and catching speaker's finger, ...

We prompt the speaker to utter utterance through a question sheet. Our question sheet is composed of many words, sentences and paragraphs. We have categorized them into four parts according to their purpose: Spontaneous speech, read sentences in MSA, translate text, and sub-spontaneous sentences.

Spontaneous Speech –SS part–

In this part, the speaker answers 05 questions, which are: *Where are you from? What is the time? Tell us about what do you like in your city? Tell us about the weather today? Describe to us the last meal you have eaten?* The aim of these data is to record conversations.

Read Sentences in MSA –RS part–

In this category, we have selected ten sentences, which are phonetically rich and balanced. We have taken the tenth list from the research of Boudraa et al. (2000). This part has twofold purposes. First, MSA speech are represented in the corpus. Second, it is shown by Ammar et al. (2014) that the accent in MSA utterance is widely influenced by the dialect of the speaker.

Translate Text –TT part–

The speaker has to translate 18 common words, 16 sentences from MSA to his dialect:

- *Digits:* from 0 through 10 and *days of the week.*
- *Accent sentences/paragraphs:* this part includes four sentences, which contain the pronunciation of the main five discriminative sounds in Arabic dialects. This part gathers between the three forms of sentences, which are normal, negative and interrogative.

  The aim of adding this part is justified by the studies of Taine-Cheikh (2000) and Holes (2004). Taine-Cheikh proves that the pronunciation of [q] and [ð], which are

in Arabic the letter 'ق' /qaf/ and 'ذ' /ḏal/ respectively, are discriminative sounds in Arabic dialects, especially they are quite discriminative when we deal with Algerian dialects. In addition, Holes adds to this list three other discriminative sounds. Two interdentals sounds [θ], and [ðˤ], which are in Arabic the letter 'ث' /ṯa/ and 'ظ' /ḍa / letters respectively, and an alveolar affricate sound [ʤ] which is 'ج' /jim/ in Arabic letter.

The texts of this part are selected from two different corpora. From SAAVB corpus (Alghamdi et al. 2008), we have selected two accent variation sentences, this is due to two reasons:

1. The first sentence includes the five discriminative sounds in MSA words that rarely change their nucleus in the dialect of a speaker, it is:
   "جَاءَ الضُيُوف الثَلاثَة بالذَهَب قَبلَ الظُهر"
   /ja:' alḍuyu:f alṯala:ṯah bialḏahab qabla alḍuhr/
   "The three guests came with the gold before noon".
2. The second sentence is an interrogative sentence. It is an indirect question introduced by "why" which expresses the interrogative style:
   "لماذا سَافرتَ الى الخَارج في العيد ؟"
   /lima:ḏa: sa:fart 'ila: alḫa:rij fi: alʿi:d/?
   "Why did you travel abroad during the festival?"

From another linguistic corpus (Ammar et al. 2014), we have selected two sentences. They include interdental fricative sounds, with a short sentence in negative style.

- *A short text story:* this part includes 12 sentences, which are selected from the well known story *The North Wind and the Sun.* We have intentionally added this part as this story represents a reference text used by the Association International Phonetics (AIP) for phonetic description of world languages[1]. AIP provides the record of this story in many languages including MSA. It represents de facto standard in most corpora that are used in language processing for phonetic research. Furthermore, it is widely used for Arabic dialect processing. It makes performing comparative studies more easier. Specifically for Arabic dialects, this story is presented in the corpus of Hamdi et al. (2004) and also in Araber Corpus for Arabic dialects collected by Barkat-Defradas available through *Corpus Pluriels* of praxiling laboratory [2].

[1] https://www.internationalphoneticassociation.org
[2] http://www.praxiling-corpuspluriels.fr

Sub-Spontaneous Sentences –SSS part–

In this category, the speaker has to narrate, in his dialect, an image story by interpreting a set of 24 ordered pictures selected from kid story *Frog, where are you?* which is called in linguistic literature *The Frog Story* (Mayer 1969). This part leads to 24 utterances. It is added to the corpus text content for many purposes. The first, is that, this story is near to daily speech in its grammatical and lexical varieties. In addition, it contains different levels and types of linguistic expression. Like *The North Wind and the Sun* story, it is frequently used in main corpora. Furthermore, it can be used to perform prosodic measurements due to the call of the narration style as proven by Himmelmann & Ladd (2008). In fact, for highlighting prosodic features, a semi-guided narration style allows two antagonist criteria. First, like in a spontaneous speech, it gives freedom to the speaker to narrate in his style. Second, it guides the narrator to follow story events and to use a specific vocabulary, essentially the motion verbs.

In Arabic, this story is used by Barkat et al. (1999) to sketch up the fact that the prosody features can be discriminative for Arabic dialects. In addition, it is presented in Araber corpus, cited bellow, which has been used in prosodic research of Rouas (2007).

Table 3.1. gives a summary of the corpus's speech material described below. *Ratio* column expresses the ratio of each part in the whole text content in term of sentences.

Part	#Utterance	Ratio (%)
Spontaneous Speech	05	8.8
Read Sentences in MSA	10	17.5
Translate Text	18	31.6
Sub-Spontaneous Sentences	24	42.1
Total	57	100

TABLE 3.1: Corpus Speech Material.

3.1.1.2 Speaker Profile

The speakers are chosen from adult population with an age in the range 18-50 years. We have made sure that speakers and their parents are native from the department of the corresponding dialect.

Let us mention that when the department is a great metropolis, we have made sure that the speakers are from the same area. In fact, we have noticed that for Algiers, Oran and Constantine departments, we can have more than one sub-dialect.

The speaker's personal information are provided by the speaker him-self. These information are namely: first and last name (optional), date and place of birth, education level, origin of his/her parents and their education level. The applicant has to inform about the different places/dates where/when he lived since he/she has 2 years old.

3.1.1.3 Material and Environment Recording

Concerning the recording environment, some conditions are respected. In fact, all recordings are performed in a quiet environment and when it is possible, we have performed recording in a sound proof room. In order to avoid noise in the recorded speeches, all the recording are done in nice days without wind, thunder, or rains.

Concerning the hardware configuration of the recording material, we have use a dictaphone *TASCAM DR-05* sets. The sample rate of recording is 48K sample/sec with 16 bits resolution.

In order to explain how we have got the speech data, let us sketch the procedure in these two steps:

- Before Recording Step: First, we explain to each applicant the purpose and objectives of the targeted speech corpus in order to reassure and give him confidence. In addition, the applicant fills in an information form that will provides useful information related to the speaker profile. Second, each applicant has to take notes in advance about the whole question sheet because of the translation of some sentences in his dialect.
- During Recording Step: Many verbal trials are allowed until we and the speaker are satisfied about the quality of dialect utterance, specially RS, TT and SSS parts. Among the satisfaction criteria that we have looked after that the speaker avoids using intensively French terminology as possible as he can. However, he can use lexical borrowing, where the speaker use a foreign word, which is adapted to his syntactic/morphology. Generally, each recording is performed twice. However, when the speaker feels uncomfortable or shy, we repeat the recording until we get an acceptable quality one.

3.1.2 Corpus Description

Due to the lack of efficient linguistic atlas for Algeria, we have chosen to collect spoken data from chef-lieu of departments where there are prominent population density. This fairy fine-grained division favors efficient dialect feature captures.

Within this preliminary version, ALG-DARIDJAH covers more than one third of the whole sub-dialects of Algeria. The covered sub-dialects are from departments of Adrar, Algiers, Annaba, Bordj Bou Arréridj, M'sila, Constantine, Djelfa, El-Bayadh, El-Oued, Ghardaïa, Laghouat, Média, Mostaganem, M'sila, Tiaret, Tissemsilt, Tlemcen and Oran. Let us note that dialects spoken in these departments are quite close to each other. They differ mainly in pronunciation and some local words. For this reason, we talk about sub-dialects.

Number of targeted departments	17
Number of speakers	109
Number of utterances	6213
Average recording by speaker	2.50 mn
Total recording	∽ 4 h 30 mn
Size of corpus	1.8 Gb

TABLE 3.2: Corpus Statistics and Details.

Table 3.2 gives a summary of some statistics of the built corpus. A sample of ALG-DARIDJAH corpus is available online[3]. ALG-DARIDJAH corpus encompassed 17 sub-dialects, an average of more than 6 speakers by sub-dialect and 109 in total. The male speaker ratio is about 44%. Table 3.3. gives more details on the distribution of speakers for each sub-dialect.

Concerning provided annotations, each utterance in ALG-DARIDJAH has time-aligned orthographic word transcription. In addition to this annotation, we have performed automatic syllable segmentation of all utterances by using Praat tool enhanced by the script Prosogram V 2.9 (Mertens 2004). Figure 3.1 reports a sample of an orthographic transcription of one utterance among Translate Text part (TT) sentences.

As a first corpus analysis, there is a lot of lexical differences between sub-dialects. We have captured some of them which are illustrated in Table 3.2.

[3] http://perso.lagh-univ.dz/~hcherroun/Alg-Daridja.html

Departement	**Gender**		Total
	#Male	#Female	
Adrar	04	02	06
Algiers	04	04	08
Annaba	05	05	10
Bordj Bou Arréridj	-	01	01
Constantine	01	01	02
Djelfa	05	05	10
El-Bayadh	01	03	04
El-Oued	05	05	10
Ghardaïa	05	05	10
Laghouat	05	05	10
Médéa	04	05	09
Mostaganem	-	01	01
M'sila	01	09	10
Tiaret	02	04	06
Tissemsilt	02	02	04
Tlemcen	02	02	04
Oran	02	02	04
Total	48	61	109

TABLE 3.3: ALG-DARIDJAH Speakers Distribution.

3.1.3 Corpus Package Organization

In order to facilitate the deployment of ALG-DARIDJAH corpus, we have adopted the same packaging method as TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al.

MSA	جَاءَ الضُيُوف الثَلاثَة بالذَهَب قَبلَ الظُهر /ja:' alḍuyu:f alṯala:ṯah bialḏahab qabla alḍuhr/
Laghouat dialect	جَاونَا ثْلاَث ضْيَافْ وْ جَابُو مْعَاهُمْ ذْهَبْ قُدَامْ الظُهُرْ /jawna tla:ṯ ḍya:f w ja:bu: m'a:hum ḏhab gudam alḍuhr/
Oran dialect	جَاو تِلْتْ ضْيَافْ بذْهَبْ قْبَلْ الدُهُرْ /jaw tilt dia:f bidhab qbal alduhur/
El-Bayadh dialect	ضْيَافْ ثْلاَتَ جَابُو ذْهَبْ قُدَامْ الظُهُرْ /ḍya:f tla:ta ja:bu: ḏhab guda:m alḍhur/

FIGURE 3.1: A Sample of Orthographic Transcription.

1993). In fact, TIMIT is considered as de facto standard of speech corpora. In the root directory of Alg-Daridjah package, we provide some textual files with .txt extension:

```
- Prompt.txt: Table of sentence prompts.
- Speakerinfo.txt: Table of speaker description.
- Orthocode.txt: Table of symbols used in orthographic transcriptions.
```

For each speaker, we have four folders, each one contains a part of text corpus. In each folder, there are a wave file and two transcription files for each utterance.

English	MSA	Sub-dialects	
Look for	يبحثا	Adrar: يْحَوطُو	/yḥawṭu:/
		Ghardaïa: يْحَوسُو	/yḥawsu:/
		Laghouat: يْدَورُو	/ydawru:/
		Tiaret: يْرَڤْبُو	/yragbu:/
Child	الطفل	El-Bayadh: بَزْ	/baz/
		Tlemcen: لِولِدْ	/wild/
		Oran: غُرِيَانْ	/ġuryaːn/
Dispute	تنازعتا	Algiers: دَارْبُو	/daːrbuː/
		Annaba: تْعَارڪُو	/tʿaːrkuː/
		Constantine: تْفَاتْنُو	/tfaːtnuː/
		Laghouat: دَوسُو	/dawsuː/
		Médéa: ضَارْبُو	/ḍaːrbuː/
		Tiaret: دَاڤُو	/daːguː/
		Tlemcen: دَابْزُو	/daːbzuː/

Figure 3.2: Illustration of Some Lexical Differences in Alg-Daridjah.

Alg-Daridjah corpus is a parallel corpus that encompasses Algerian Arabic sub-dialects. It is based on direct recording method. Thus, many considerations are controlled while building this corpus. The size of Alg-Daridjah corpus is restricted. In next section, we tend to fill this gap by giving a complete recipe to build a large-size speech corpus by using Web-based resourcing. This recipe can be adopted for any under-resourced language. It eases the challenging task of building large datasets by means of traditional direct recording. Which is known as time and cost consuming.

3.2 Kalam'DZ Corpus

Turning now to the second created corpus, *KALAM'DZ* is Web-based dedicated also to Algerian dialect variety. Our idea relies on Web resources, an essential milestone of our era. In fact, the Web 2.0, becomes a global platform for information access and sharing that allows collecting any type of data at scales hardly conceivable in near past.

As mentioned previously according to our review of the major Arabic dialects corpora, in Section 2.5, there is no Web-based speech corpus that deals with Arabic speech data neither for MSA nor for dialects. While for other languages, there are some investigations. We can cite the large recent Web-based collection *Kalaka*-3 (Rodríguez-Fuentes et al. 2016) for European languages. This is a speech database specifically designed for Spoken Language Recognition. The dataset provides TV broadcast speech for training, and audio data extracted from YouTube videos for testing.

3.2.1 Methodology

In this section, we first describe, in general way, the complete recipe, that we have designed, in order to collect and annotate a Web-based spoken dataset for an under-resourced language/dialect. Then, we illustrate this recipe to build our Algerian Arabic dialectal speech corpus mainly dedicated to dialect and speaker identification.

Global View of the Recipe

The recipe described in the following can be easily tailored according to potential uses of the corpus and on the specificities of the targeted language resources and its spoken community. Figure 3.3 shows the global view of the recipe.

1. *Inventorying Potential Web sources:* First, we have to identify sources that are the most targeted by the communities of the languages/dialects in concerns. Indeed, depending on their culture and preferences, some communities show preference for dealing with some Web media over others. For example, Algerian people are less used to *Instagram* or *Snapchat* compared with Middle Est and Gulf ones. Moreover, each country has its own most used communication media. For instance, some societies (Arab ones) are more productive on TVs and Radios, compared with west communities that are more present and productive on social media.

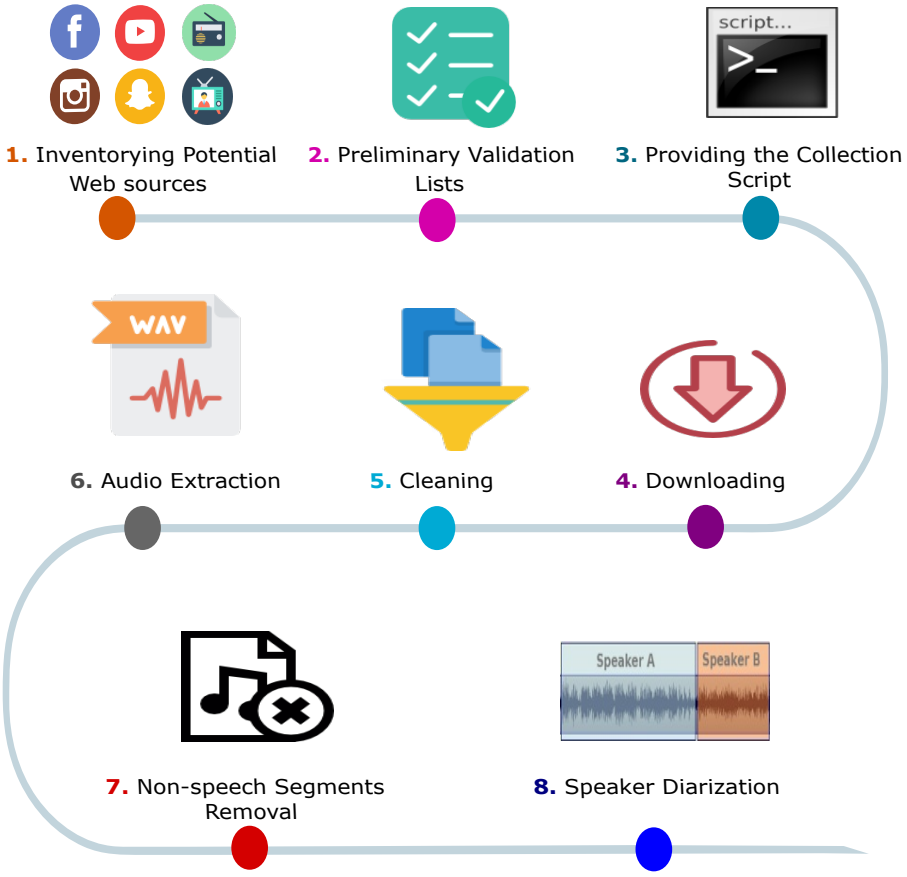


FIGURE 3.3: The Global View of the Recipe.

2. *Extraction Process:* In order to avoid crawling useless data, this step is achieved through three stages

   (a) *Preliminary Validation Lists*: For each chosen Web source, we define the main keywords that can help automatically search video/audio lists. When such lists are established, a first cleaning is performed keeping only the potential suitable data. Sizing such lists depends on the sought scale.

   (b) *Providing the collection script*: For each resource, we fix and implement the suitable way to collect data automatically. Open Source tools are the most suitable. In fact, downloading a speech from a streaming or from YouTube or even from online Tv needs different scripts. The same fact has to be taken into account concerning their related metadata[4] which are very useful for annotation.

   (c) *Downloading*: This is a time consuming task. Thus, it is important to consider many facts such as preparing storage and downloading the related metadata, ...

[4]YouTube video Metadata such as *published_date, duration, description, category...*

(d) *Cleaning*: Now, the videos/audios are locally available, a first scan is performed in order to keep the most appropriate data to the corpus concerns. This can be achieved by establishing a strategy depending on the corpus future use.

3. *Annotation and Pre-processing*: For a targeted NLP task, pre-processing the collected speech/video can include segmentation, White noise removing.... Some annotations can simply be provided from the related metadata of the Web-source when they exist. However, this task makes use of other annotation techniques like crowdsourcing where crowd are called to identify the targeted dialect/speaker or/and perform translations.

The method can be generalized to other languages/dialects without linguistic and cultural knowledge of the regional language or dialect by using video/audio search query based on the area (location) of targeted dialect/language. Then use the power of crowdsourcing to annotate corpus.

3.2.2 Corpus Building

For the context of the Algerian dialects, in order to build a speech corpus that is mainly dedicated to dialect/speaker identification using machine learning techniques, we have chosen several resources. This corpus building is sketched through the recipe' steps application.

3.2.2.1 Web Sources Inventory

The main aim is to allow the richness of the corpus. In fact, it is well known that modeling a spoken language needs a set of speech data counting the within-language/intersession variability, such as speaker, content, recording device, communication channel, and background noise (Li et al. 2013). It is desirable to have sufficient data that include the intended intersession effects.

Table 3.4 reports the main Web sources that feed our corpus. Let us observe that there are several speech Topics which allows capturing more linguistics varieties. In fact, this inventory contains *Local radio channels* resources. Fortunately, each Algerian province has at least one local radio (a governmental one). It deals with local community concerns and exhibits main local events. Some of their reports often use their own local dialect. It is the same case for amateur radios. Both, these radio channels and Tvs are Web-streamed live.

Source	Sample	Topics
Algerian Tv		
	Ennahar	News
	El chorouk	News, General
	Samira, Bena	Cook
Local Radios		
	48 departments	Social, local, General
On YouTube		
Algerian PodCast	Anes Tina	Politic, Culture, Social
Algerian	Khaled Fkir	Blogs, Cook
YouTubers	CCNA DZ	Tips, Fun
	Mesolyte	Advices, Beauty Technology, Vlog[5]
Algerian TAG		
	–	Advices, Tips, Social discussions

TABLE 3.4: Main Sources of Videos/Radio.

In addition, we have chosen some Algerian TVs for which most programs are addressed to large community. So, they use dialects. Finally, we have targeted some YouTube resources such as Algerian PodCasts, Algerian Tags, and channels of Algerian YouTubers.

3.2.2.2 Extraction Process

Now having these Web sources, and as they are numerous, we process in two steps in order to acquire video/audio speech data. First, we drawn up lists by crawling information mainly meta data about existing data related to potential videos/audios that potentially contain Algerian dialect speech. The deployed procedure relies on mainly two different scripts according to the Web resource type.

In order to collect speech data from local radio channels, we refer back manually to the programs of radio to select report and emission lists that are susceptible to contain dialectal speech.

Concerning data from YouTube, the lists are fed by using YouTube search engine through its Data API [6]. In addition, the extraction of related metadata are performed using Python package BeautifulSoup $V4$ dedicated to Web crawling (Nair 2014).

[6]YouTube Data API, https://developers.google.com/YouTube/v3/

In order to draw up search queries, we have used three lists. The first one *Dep*, contains the names of the 48 Algerian provinces, spelled in MSA, dialect and French. While the second list *Cat* contains the selected YouTube categories. Among the actual YouTube categories, we have targeted: *People & Blogs, Education, Entertainment, How-to & Style, News & Politics, Non-profits & Activism categories, and Humor*. The third list *Comm_Word* contains a set of common Algerian dialect words (called White Algerian terms) that are used by all Algerians. These chosen set is reported in Table 3.5.

واش	What	مشاكل	Problems	نقاش	Discussion
جرنان	Journal	الحقرة	Injure	بلاك	Maybe
تريسيتي	Electricity	يصرا	Happen	صاري	Happened
دارجة	Colloquial	دزيرية	Algerian	الدشرة	Village
النشرة	News	شوف	See	جزايري	Algerian
دزاير	Algeria	لولاية	Department	الباك	Baccalaureat
بروبليم	Problem	ماكلة	food	وشرايك	Your Opinion

TABLE 3.5: A Sample of Common Algerian Terms Used as Keywords.

Then, we iterate searching using a search query model that has four keywords. The first and the second ones are from *Dep* and *Cat* lists respectively. The remaining two keywords are selected arbitrary from *Comm_Word*. This query formulation can guarantee that speakers are from the fixed province and the content topics are varied thanks to YouTube topic classification.

Concerning Algerian TVs source, the search queries are drawn up using mainly two keywords. The first one is the name the channel and the second word refers to the report name. In fact, a prior list is filled manually with emission and report names that uses dialects. Easily, videos from Algerian YouTubers/Podcasts channels are searched using the name of the corresponding author.

For all YouTube search queries, we selected the first 100 videos (when available). When lists are drawn up, we start the cleaning process that discards the irrelevant video entries. In fact, we remove all list entries whose duration is less than 5s. The videos whose topic shows that it doesn't deal with dialects are also discarded. This is done by analyzing manually the video title then its description and keywords.

In addition, to be in compliance with YouTube Terms of Services, first, we take into account the existence of any Creative Commons license associated to the video.

The whole cleaning process leads us to keep 1182 Audios/videos among 4 000 retrieved ones. Video's length varies between 6s and 1 hour. The cleaning task is manually carried out by us.

Now, the download process is launched using the resulting cleaned lists. Concerning the data from radio channels, we script a planned recording of the reports from the stream of the radio. Here also the recording amount depends on the desired scale.

Concerning the data from Radios, we deploy, in the download script, mainly the *VLC* Media Player tool[7] with the *cron*[8] Linux command. In order to download videos from YouTube, we have deployed YouTube-dl[9] a command-line program.

3.2.2.3 Preprocessing and Annotation

A collection of Web speech data is not immediately suitable for exploration in the same way a traditional corpus is. It needs more cleaning and preprocessing. Let us recall, that our illustrative corpus will serve for dialect/speaker identification using machine learning techniques. For that purpose, we have applied the following processing for each downloaded video/audio:

1. *Audio extraction*: FFmpeg[10] tool is used to extract the audio layer. In addition, the SoX[11] tool, a sound processing program, is applied to get single-channel 16 kHz WAV audio files.

2. *Non-speech segments removal*: we have also remove all non speech segments such as music or white noise by running a VAD (Voice Activation Detection) tool to remove as many as possible.

3. *Speaker Diarization*: this step is performed to determine who speaks when, and to assign for each utterance a speaker ID. It is achieved using VoiceID Python library based on the LIUM[12] Speaker Diarization framework (Meignier & Merlin 2010). The output from VoiceID segmentation is a set of audio files with information about speaker ID, and utterance duration.

[7] VLC media player version2.2.4, https://www.videolan.org/vlc/
[8] cron: is a time-based job scheduler that specifies shell commands to run periodically on a given schedule.
[9] YouTube-dl version3, http://YouTube-dl.org/
[10] FFmpeg version3.2, http://www.ffmpeg.org/
[11] SoX, http://sox.sourceforge.net/
[12] LIUM laboratory of the University of Mans, France.

For this preliminary version, most manual annotations are made thanks to our laboratory team with the help of 8 volunteers. These In Lab annotations concern assigning for each utterance the spoken dialect, validation of speaker gender (previously detected automatically by VoiceID). During these manual annotations, we check that utterances deal with dialect. Otherwise, they are discarded.

3.2.3 Corpus Main Features

First of all, we note that this preliminary version of our corpus is collected and annotated in less than two months, and the building is still an ongoing process. A sample of our corpus is available online[13]. KALAM'DZ covers most major Arabic sub-dialects of Algeria.

Number of Dialects	8
Total Duration	104.4 hours
Clean Speech Ratio	39, 15 %
Number of speakers	4881
Speech duration by Speaker	6s – 9hours

TABLE 3.6: Corpus Global Statistics.

Table 3.6 reports some global statistics of the resulted corpus. *Clean Speech Ratio* row gives the ratio of Speeches that have good sound quality. The remaning portion of speeches present some noise background mainly music or they are recorded in outdoor. However, they can be used to build dialect models in order to represent the within-language variability. Table 3.7 shows the distribution of categories per dialect.

More details on KALAM'DZ corpus are reported in Table 3.8. Let us observe that some dialects are more represented. This is due to people distribution and Web culture. For instance, Algiers and Oran are metropolitan cities. So, their people productivity on the Web is more abundant.

In order to facilitate the deployment of KALAM'DZ corpus, as done for ALG-DARIDJAH corpus, we have adopted the same packaging method as TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al. 1993).

[13] https://github.com/LIM-MoDos/KalamDZ

Sub-Dialect	N&P	Edu.	Ent.	H&S	P&B	Hum.
Hilali-Saharan	29.4	-	-	-	-	-
Hilali-Tellian	3.6	-	-	-	-	-
High-plains	2.0	-	-	-	-	-
Ma'qilian	4.1	1.5	-	7.4	10.2	1.6
Sulaymite	6.7	-	-	-	6.9	-
Algiers Blanks	5.1	2.2	8.7	-	1.4	2.8
Sahel-Tell	9.1	-	-	-	-	-
Pre-Hilālī	0.7	-	-	-	-	-
Total	60.7	3.7	8.7	7.4	19.5	4.4

TABLE 3.7: Distribution of Categories per Dialect (in hours). (N&P= News & Politics, Edu.= Educations, Ent.= Entertainment, H&S= How-to & Style, P&B= People & Blogs, Hum.= Humor)

3.3 Potential Uses

ALG-DARIDJAH and KALAM'DZ corpora can be useful for many purposes both for NLP and computational linguistic communities. In fact, they can be used for building models for both speaker and dialect identification systems for the Algerian dialects. For linguistic and sociolinguistics communities, they can serve as base for capturing dialects characteristics.

In addition, ALG-DARIDJAH corpus is the first parallel corpus for Algerian Arabic dialects. This type of corpus will help researchers to study the dialects themselves, focusing on the differences between the dialects. Especially for those languages that show a marked difference between the dialects. In addition, linguists can use our fine-grained spoken corpus to create contemporary dialect atlases by gathering similar sub-dialects.

Furthermore, KALAM'DZ corpus can be deployed to build another corpus version to serve any image/video processing based applications because all videos related to are also available.

Dialect	Departments	Speakers	Web-sources (h)			Total (h)	GQ (%)
			TV	Radios	YouTube		
Hilali-Saharan	Laghouat, Djelfa, Ghardaia, Adrar, Bechar, Naâma' South	1338	12.7	16.5	-	29.4	38.5
Hilali-Tellian	Setif, Batna, Bordj-Bou-Arreridj	605	3.6	-	-	3.6	32.2
High-plains	Constantine, Mila, Skikda	297	2.0	-	-	2.0	42.7
Ma'qilian	Sidi-Bel Abbas, Saïda, Mascara, Relizane, Oran, Aïn Timouchent, Tiaret, Mostaganem, Naâma' North	421	4.1	-	20.7	24.8	38.3
Sulaymite	Annaba, El-Oued, Souk-Ahras, Tebessa, Biskra, Khanchela, Oum El Bouagui, Guelma, El Taref	914	6.7	-	6.9	13.6	39
Algiers Blanks	Algiers, Blida, Boumerdes, Tipaza	723	5.1	-	16.1	21.2	42.3
Sahel-Tell	Médea, Chlef, Tissemsilt, Ain Defla	447	3.1	6.0	-	9.1	45.5
Pre-Hilālī	Tlemcen Nadrouma, Dellys, Jijel, Collo, Cherchell	136	0.7	-	-	0.7	34.7
Global		4881	38.2	22.5	43.7	104.4	39.1

TABLE 3.8: Distribution of Speakers and Web-sources per Sub-dialect in KALAM'DZ Corpus. (GQ= Good Quality)

Chapter 4

Altruistic Crowdsourcing for the Validation of Dialect Annotation of Kalam'DZ

Contents

In this chapter, we review some related work that have employed crowdsourcing for Arabic NLP tasks. Then, we explain the designed of the altruistic crowdsourcing project and the deployed quality control strategy. Then, from the altruistic crowdsourcing experiments, we extract the list of best practices.

4.1 Introduction

Crowdsourcing is a form of collaboration that involves the public *crowd* in scientific research to address real world problems in the form of an open call (Quinn & Bederson 2011). Crowdsourcing takes many forms, which are typically classified along three dimensions: the motivation of human contributors (fun, altruism, or payment), the way in which individual results are aggregated, and how quality is controlled (Sabou et al. 2014). The major crowdsourcing genres most widely adopted by scientific communities are:

- Mechanised labour as a type of *paid-for* crowdsourcing, where volunteers choose to carry out small tasks and are paid a small amount of money in return.
- Games With A Purpose (*GWAP*) enable human contributors to carry out computation tasks as a side effect of playing online games.
- *Altruistic* crowdsourcing which refers to cases where a task is carried out by a large number of volunteer contributors. Altruistic crowdsourcing approaches leverage the intrinsic motivation of a community interested in a domain (Kaufmann & Schulze 2011).

The use of crowdsourcing techniques has impacted a broad range of science-related projects especially Natural Language Processing (NLP) research (Sabou et al. 2014). The advantages of crowdsourcing compared with expert-based annotation have already been discussed in (Wang et al. 2013) which show that crowdsourcing tends to be cheaper and faster solution for many NLP related tasks. Sabou et al. (2014) show that mechanised labour is the most popular genre among NLP researchers, while altruistic crowdsourcing is the least frequently used approach.

Speech corpora are crucial for both developing and evaluating NLP systems. Moreover, such corpora have to respond to NLP communities expectations and allow to be exploited in machine learning tasks and statistically motivated methods. For many languages, the state of the art results have been obtained and validated through using a large and well designed

corpora. On the other extreme, there are few corpora for Arabic (Mansour 2013). Moreover, very few attempts have been considered for Algerian Arabic dialect (Djellab et al. 2017).

In our first contribution, we have created KALAM'DZ corpus (Section 3.2), which is developed to cover the Arabic dialectal varieties of Algeria. This corpus is collected using web-based sources. Despite its important size, about more than 104 hours, the bulk of its annotations were harvested semi-automatically from the related web-source metadata. Thus, these annotations need validation. Using experts, this process can take more time and can be costly.

In our second contribution, we aim to enrich and validate KALAM'DZ annotations by exploring and harnessing altruistic crowdsourcing. Many reasons have guided our choice of such crowdsourcing genre for that task.

On one hand, altruistic crowdsourcing approaches can leverage the intrinsic motivation of a community interested in a domain that can reduce the incentive to cheat (for money or glory). In fact, Kaufmann & Schulze (2011) prove that, for many workers in crowdsourcing platforms, intrinsic motivation aspects are more important than extrinsic ones (such immediate payoffs), especially the different facets of enjoyment-based motivation like working on tasks to learn new or train existing skills.

On the other hand, micro-payment are not very popular in our targeted communities. Further, their presence in crowdsourcing platforms as client is very modest. In addition, in spite of their low cost, funding crowdsourcing projects is still not a common practice within the Algerian research community.

4.2 Crowdsourcing for Arabic NLP

Crowdsourcing usage for Arabic NLP tasks can be considered as new practice compared to others languages. In the followings, we focus on reviewing some relevant works on crowdsourcing usage for Arabic NLP. In fact, crowdsourcing has been used effectively for: both corpus annotation (text-based (Zaghouani & Dukes 2014), and speech-based (Hakouz et al. 2015, Wray & Ali 2015)), speech transcription (Wray et al. 2015), text translation (Kumar et al. 2014), and Arabic digitization with diacritics (Kassem et al. 2016).

Wray & Ali (2015) use the CrowdFlower (CF) platform to create a labeled multi-dialectal speech. Their task was restricted to users in the Arab world. Contributors were directed to listen to short speech segments and determine the dialect. They followed two ways to ensure quality control. First, Quiz Mode offered by CF where contributors have to answer five Gold

Standard (GS) tasks before entering the main portion of the task. Second, for every five tasks, contributors were presented with a GS task. The participation of contributors is ended if their task accuracy does not reach a given threshold.

Hakouz et al. (2015) present *Lahajet*, a GWAP for crowdsourcing classifications of different varieties of Dialectal Arabic in multi-dialectal audio. Lahajet consists of players that listen to short audio clips and select a character representing the dialect.

Wray et al. (2015) investigate different approaches using crowdsourcing for transcriptions of Dialectal Arabic speech with automatic quality control using CF. Concerning the validation, they modified the GS mechanism, in which at least 10% of the segments are transcribed by experts to distinguish between trusted and untrusted transcribers, where they compared the workers transcriptions against the automatic speech recognition system results.

Zaghouani & Dukes (2014) explore the use of Amazon Mechanical Turk (MTurk) to enrich the annotation of the Quranic[1] corpus. Two experiments were performed to manually annotated, which are the Part-Of-Speech (POS) tagging and grammatical case-ending. In order to measure the quality of the annotation, the GS annotation is provided by volunteer annotators who are typically Arabic linguists or Quranic experts. The whole experiments was carried over a period of one month. The total number of interested workers in both experiments combined was 137. To filter out potential non-qualified workers and to ensure the quality control by non-native speakers, an Arabic language qualification test was used. As main results, they show that annotating Arabic grammatical case is a harder task than POS tagging. For both tasks, they show that crowdsourcing is not effective.

Kumar et al. (2014) perform the CallHome Egyptian Arabic-English speech translation corpus by using crowdsourcing technique. They used Mturk crowdsourcing platform to obtain translations. For each Arabic transcript, there obtained four English translations. As result, 838 translators participated. They used many quality control mechanisms such as: select only Arabic speakers who have already worked on other translation tasks, GS is employed, each utterance is compared to Google Translate, and the utterance is presented as an image.

Kassem et al. (2016) present two GWAPs: *tashkeelWAP*[2], a two-player game and a single-player game, that allow the digitization of Arabic text by outsourcing it to native Arabic speaking players. Its aim is to offer an enjoyable gaming experience to players and as a bi-product obtain Arabic digitizations of documents with diacritics.

[1] Quranic corpus provides annotation which shows the Arabic grammar, syntax, and morphology for each word in the Quran, http://corpus.quran.com/

[2] tashkeelWAP: a web application, https://tashkeel.herokuapp.com

NLP Task			Platform	Corpus	# Workers	Cost	Duration (days)	Quality Control
Speech	Annotation	Wray & Ali (2015)	CF	Al Jazeera	2 053	0.03 USD (per 10 tasks)	21	Before staring: Quiz Mode, Within execution: for every five tasks have a GS task.
		Hakouz et al. (2015)	GWAP	Al Jazeera	-	-	-	-
	Transcription	Wray et al. (2015)	CF	Al Jazeera	149	0.05 USD (per 5 tasks)	-	Compared with automatic speech recognition.
Text	Annotation	Zaghouani & Dukes (2014)	Mturk	Quran	137	0.01$ (per task)	30	Arabic language qualification test.
	Translation	Kumar et al. (2014)	Mturk	Callhome	838	-	-	Select contributors, GS translations, compare to GT, utterance is an image.
	Digitization	Kassem et al. (2016)	GWAP	-	-	-	-	-

TABLE 4.1: Crowdsourcing-based Approaches for Arabic Natural Language Processing Tasks.

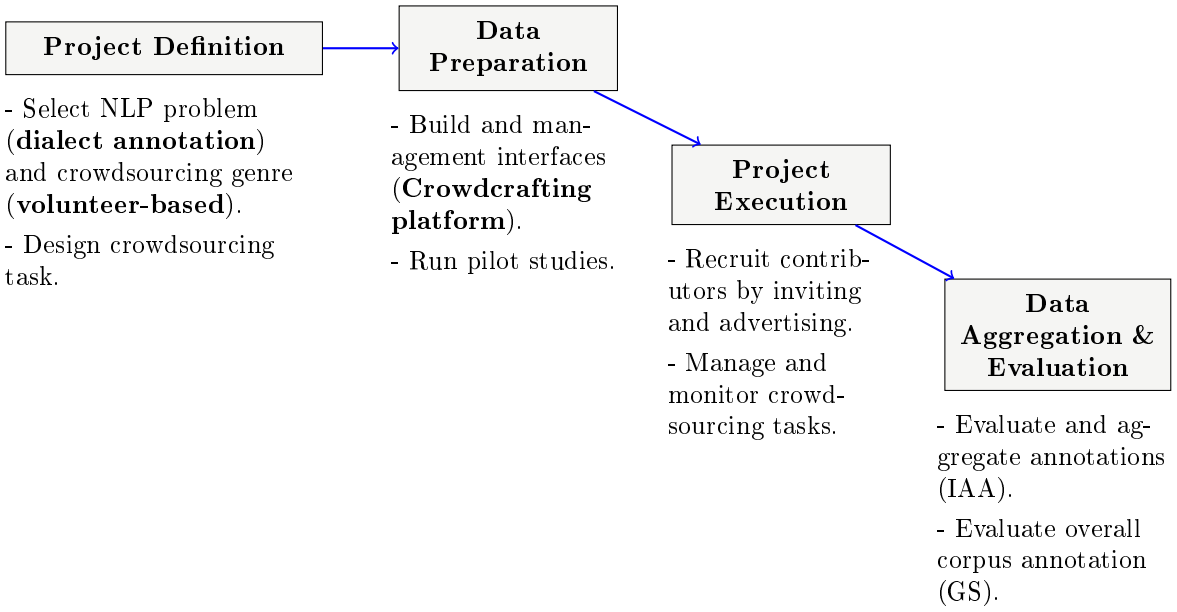


FIGURE 4.1: Crowdsourcing Corpus Annotation Process.

Table 4.1 gives more details on these crowdsourcing usage for Arabic NLP. We observed that the most reviewed Arabic NLP crowdsourcing tasks are done with fees by using Mturk or CrowdFlower. While altrustic crowdsourcing has not received any attention despite its advantages for NLP tasks (Borromeo & Toyama 2016).

4.3 Altruistic Crowdsourcing Project

As mentioned previously (Section 4.1), the dialect annotation of KALAM'DZ corpus is done semi-automatically from the related web-based sources. Using experts, the validation of this annotation takes more time and be costly. For that, altruistic crowdsourcing technique is deployed. In this section, we describe our crowdsourcing experiments in which we explain the designed of the altruistic crowdsourcing project and the deployed quality control strategy.

In order to make annotation scalable and of high quality, we have followed the crowdsourcing engineering process defined by Sabou et al. (2014) which we have adapted to our altruistic crowdsourcing Project. Figure 4.1 illustrates this process. It suggests designing the system in four stages: project definition, data preparation, project execution, and data aggregation & evaluation.

4.3.1 Project Definition

In this step, we defined the NLP task which is speech corpus dialect labeling and validation. The crowdsourcing genre selected is the altruistic approach, and we accordingly designed the crowdsourcing task. Concerning the task definition, to achieve our goal, two alternative tasks can be defined: *i)* The contributor will be asked to detect the spoken dialect using a proposed multi-choice selection, *ii)* or he will be asked to validate the available semi-automatic dialect annotation. For the sake of ease and to avoid contributor workload, we have chosen the second alternative.

4.3.2 Data Preparation

In this step, we build the project and manage the interfaces. One can choose among the existing free platforms: PyBossa[3], Zooniverse[4], and Crowdcrafting [5]. We have opted for a simple and open source platform, which is *CrowdCrafting* based on the well known PyBossa platform.

A CrowdCrafting project, as in PyBossa, has two main components that allow requester to customize his project: *Task Presenter* and *Task Creator*. The contributors are called *crafters* while the project creator is named *Requester*. We have designed a form containing a question about a given speech segment with buttons for three possible responses (Yes, No, Unknown). In addition, this form contains a map of Algerian dialects to help the crafters. It provides some links to the project social networks accounts (Facebook, Twitter) which can give more details on the dialects and allow engaging and recruiting more crafters.

Our task is restricted mainly to Algerian users for that the form is written in Arabic. In order to attract contributors of all levels and classes, the user interface design was established and validated to work cross-platforms and devices employing BootStrap[6] design framework. Figure 4.2 illustrates our desktop user interface.

Additionally, our crowdcrafting project employs stateless control where the tasks get delivered to the contributors regardless of the conditions of their status profile of connexion (authenticated or anonymous). Only the task link is used to contribute to the project.

[3]PyBossa: is a free, open-source framework for crowdsourcing, developed by Sci-Fabric company, http://pybossa.com/

[4]Zooniverse: is largest and most popular crowdsourcing platform for people powered research, https://www.zooniverse.org/

[5]The platform is 100% open source making scientific research accessible to everyone, https://crowdcrafting.org/

[6]Bootstrap, http://getbootstrap.com/

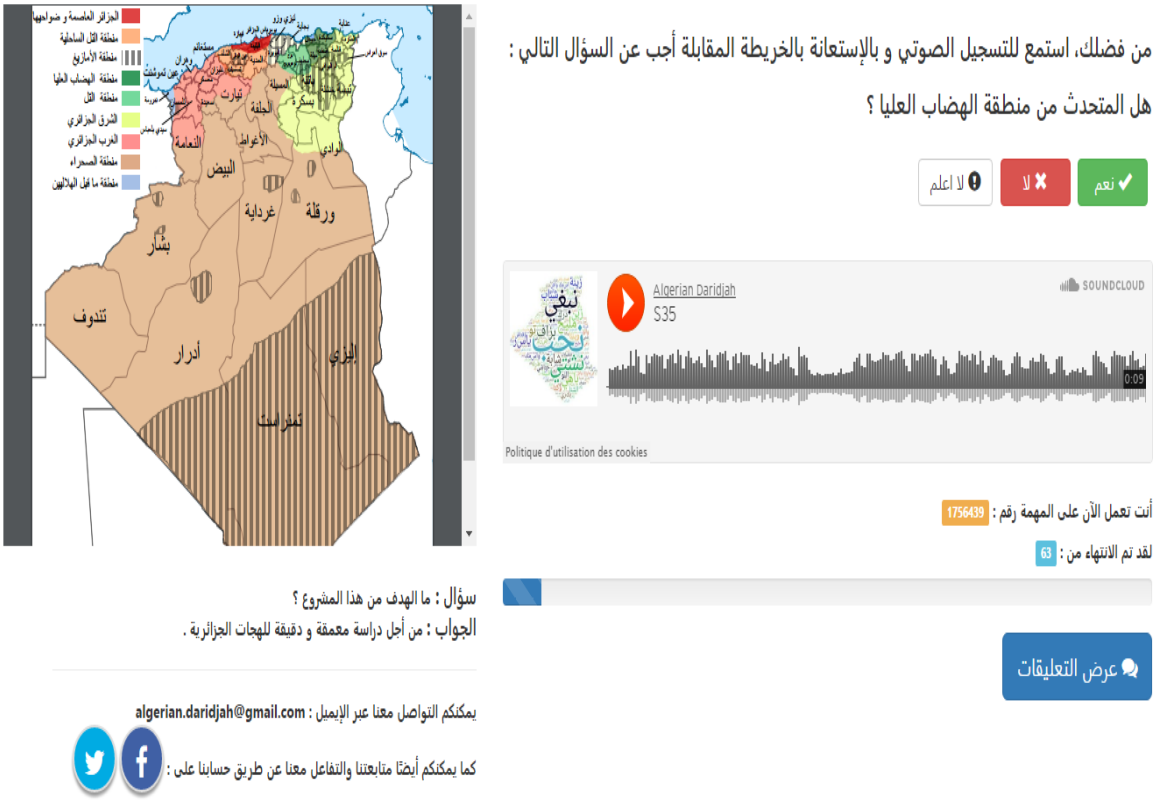


FIGURE 4.2: Desktop User Interface.

Regarding the tasks, for this pilot experiment, we have selected a sample of 10% of KALAM'DZ corpus, totaling more than 10 hours which takes into consideration the inter-dialect variability (Li et al. 2013). The utterances and segments selected are used to create our 1012 tasks which is also the number of speakers ensuring the maximum coverage of speakers as well as the dialects.

Table 4.2 shows general information about the tasks. We have tried to balance the contribution of each dialect to be close to the original distribution, knowing that in KALAM'DZ corpus, some dialects (Sahel-Tell, Pre-Hilālī, and Hilali-Tellian) are less represented. Each annotation task is presented to five independent annotators. The multi tasks reading is performed to gather more observations for each segments and it will be used to draw inter-annotator measurements between the users. The contributors were asked to listen only as long as necessary to determine the dialect being spoken before making a judgment and decision.

Dialect	# of Utterances	Average time per utterance (s)	Total Time
Sulaymite	222	36	2h 12 mn
Ma'qilian	126	38	1h 20 mn
Hilālī-Saharan	126	38	1h 18 mn
Hilali-Tellian	112	28	53 mn
High-plains	126	38	1h 20 mn
Algiers-Blanks	126	50	1h 41mn
Sahel-Tell	98	36	59 mn
Pre-Hilālī	76	32	41 mn
Total	1012	37	10h 24 mn

TABLE 4.2: Distribution of Tasks per Dialect.

4.3.3 Project Execution

In this phase the contributors must be recruited to participate with some motivation. In fact, contrary to paid crowdsourcing, the recruitment consists in a set of primarily advertising activities to attract contributors to the crowdsourcing project. In our case, we have relied on two mechanisms. First, we have used Social Media power by creating dedicated Facebook page and Twitter account. In addition, we have done an open call by posting an ad in all Facebook groups of Algerian universities through the Facebook page. In fact, our dedicated Facebook page contains and shares a lot of historical and linguistic information on dialects which is an attraction for potential contributors. As a second mechanism, we have invited directly people to participate by using mailing lists that contain students, some Algerian NLP experts, and academic researchers. As filtering step, we have restricted the advertisement of the project to only Algerian communities.

A core challenge of all altruistic crowdsourcing is how to motivate contributors to participate. For that purpose, we have relied on a campaign motivation that uses a motto *A Call for Community Help to Survey Algerian dialects.*

4.3.4 Data Aggregation & Evaluation

In unpaid crowdsourcing, imposing constraints to workers is not recommended. Thus, all quality control mechanisms are applied after project execution step. In our case to eliminate malicious works, we have excluded all contributions from outside of Algeria. In addition, all answers from crafter who has listened to the speech less than one second.

Concerning the data aggregation, the resulting annotations for each task are gathered using majority voting which represents an Inter-Annotator Agreement (IAA). The evaluation of the whole corpus resulting annotations is performed by contrasting the results with those of Gold Standard method. For which we selected 10% of the tasks, which is about 106 tasks, that are annotated manually by experts.

4.4 Results and Discussion

The Dz كـلام[7] project, our altruistic crowdsourcing framework, was designed and developed between March and April 2017. It was launched on 17th April. All Annotations were completed within a period of 17 days with 218 crafters participated in the exercise. The quality control formerly described is than applied. As such, we have ensured that all contributors are from Algeria using IP address information. According to our advertising strategy, all invited workers have participated from Algeria. On the other hand, a surprising result was that only 9 answers (less than 2%) were rejected due to the respect of one second listening threshold. This fact prove that the designed task iq attractive.

In order to give some interpretation to Dz كـلام crowdsourcing results, we have targeted to answer the following questions: *i)* What can be concluded on the results quality of our system? *ii)* Is altruistic crowdsourcing efficient for dialect annotation/validation corpus? How about needed time? *iii)* What we have learned from that experience in order to determine good practices for altruistic dialect annotation crowdsourcing?

Annotation Quality

For evaluating the altruistic crowdsourcing annotation, we have compared the Inter-Annotator Agreement (IAA) results with both the semi-automatic annotation and a baseline resource provided by an expert. Table 4.3 illustrates these results, ADT column reports the average time that the worker takes to get his decision.

We observe that the crowd agree with the semi-automatic annotation with more than 82% and disagree with 12.27% while 5.53% of them give neutral judgment. This result confirms that altruistic crowdsourcing is efficient for such NLP task.

These results provide also some interesting results on linguistic characteristics of Algerian dialectal varieties. In fact, the Sulaymite, Ma'qilian, and Hilālī-Saharan dialects are the most

[7]Dz كـلام: https://crowdcrafting.org/project/DZDaridjah/

Dialect	ADT (s)	IAA (%)			GSA (%)
		Yes	No	NotKnown	
Ma'qilian	72.2	**94.44**	**5.56**	0.0	92.8
Hilālī-Saharan	74.7	**93.60**	6.40	0.0	85.7
Sulaymite	88.5	**90.99**	8.56	0.45	81.8
Algiers-Blanks	97.6	80.95	17.46	1.59	92.3
High plains	95.8	73.81	10.31	15.87	76.9
Hilālī-Tellian	108.9	65.18	18.75	16.07	72.7
Pre-Hilālī	90.0	62.34	**27.27**	10.38	**55.5**
Sahel-Tell	120.9	79.38	13.40	7.22	70.0
Average	-	82.19	12.27	5.53	81.1

TABLE 4.3: Statistics on Average Decision Time (ADT), and Comparative Results with IAA and Gold Standard per Dialect. (GSA= Gold Standard Agreement)

easiest to detect as the contributors validate them with more than 90%. The High-plains of Constantine, Sahel-Tell, and Algiers-Blanks dialects are bit less identifiable. While, the Pre-Hilālī and Hilālī-Tellian dialects are hard to be recognized. The measured Average Decision Time per Dialect confirm these linguistic features of dialects.

Concerning the comparison with the expert annotation, the results show that the accuracy of crowd annotation is about 81.1 %, where expert agree crowds in 86 gold tasks from a total of 106 tasks.

Altruistic vs. Paid Crowdsourcing

In order to get a rough idea about altruistic crowdsourcing efficiency compared with paid crowdsourcing genre for the targeted NLP task, we have compared our results with those of Wray & Ali (2015). They performed a large scale speech corpus dialect annotation using the well known paid crowdsourcing platform CrowdFlower.

Table 4.4 reports some details on this comparison.

Despite the fact that our unpaid crowdsourcing is relatively slower (see the average answers per day), these results can be explained by the fact that our advertising scheme is limited and modest when compared to large CrowdFlower membership. In addition, we observe that for both frameworks the average answers per contributor are similar.

Figure 4.3 shows the distribution of answers according to connexion user profile (authenticated or anonymous). The distribution of answers according to the connexion worker's profile

	Wray and Ali	Our annotation
Corpus size	57 hours	10h 24 mn
Project duration (days)	21	17
Average Answers per contributor	23	24
Average Answers per day	2271	288
Number of segments	47696	5060

TABLE 4.4: Comparison between Wray & Ali (2015) and Dz كلام Annotation.

is about 91% and 9% for authenticated and anonymous connexion respectively. Even Crowdcrafting platform allows authenticated connexion, this result shows that crowd prefer working in anonymous way. So, the requester must rethink the quality control strategies that involve worker' profile knowledge. In our case, we have rely on location detection technologies to guess worker' region.

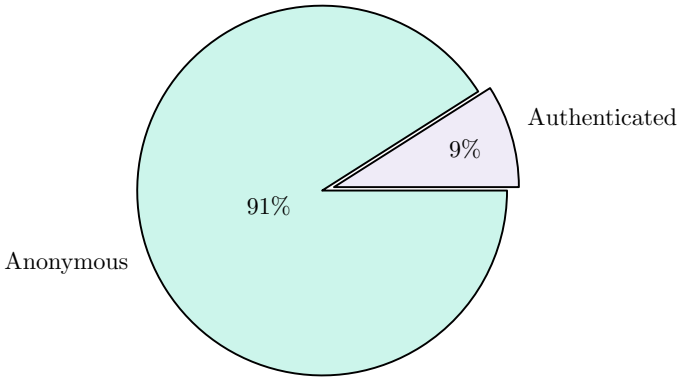


FIGURE 4.3: The Distribution of Answers.

Figure 4.4 reports the distribution of crafters according to the region from where they have participated. It confirms that all crafters are from Algeria. We observe that most crafters are from the south of Algeria, it is predicted as the open calls are launched from this region. Concerning some crafters (19%), their IP address information allows only the country location.

Figure 4.5 reports the distribution of answers per day according to their connexion profile. These results show that in the five first days a highest participation rate is reported. It is due to its closeness from the massive first open call. Then, we have observed that the number of participation has relapsed from the sixth day. This situation has pushed us to relaunch more calls both by emails and via Social Media posts whenever it is necessary. Thus, the requester has to examine periodically the rate of participation.

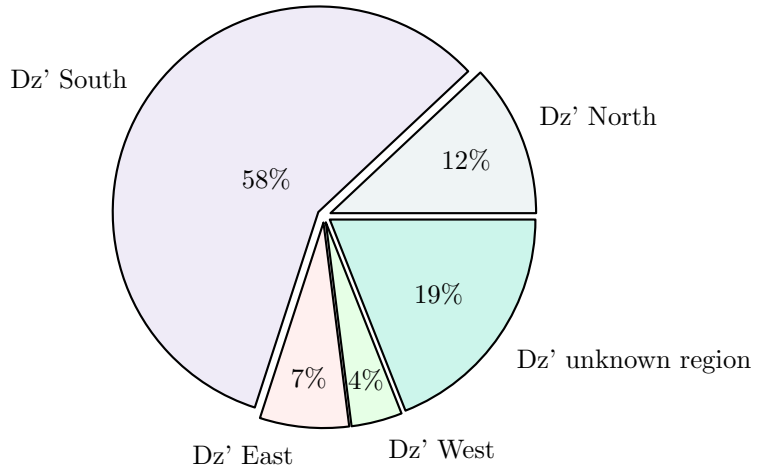


FIGURE 4.4: Location of Crafters.

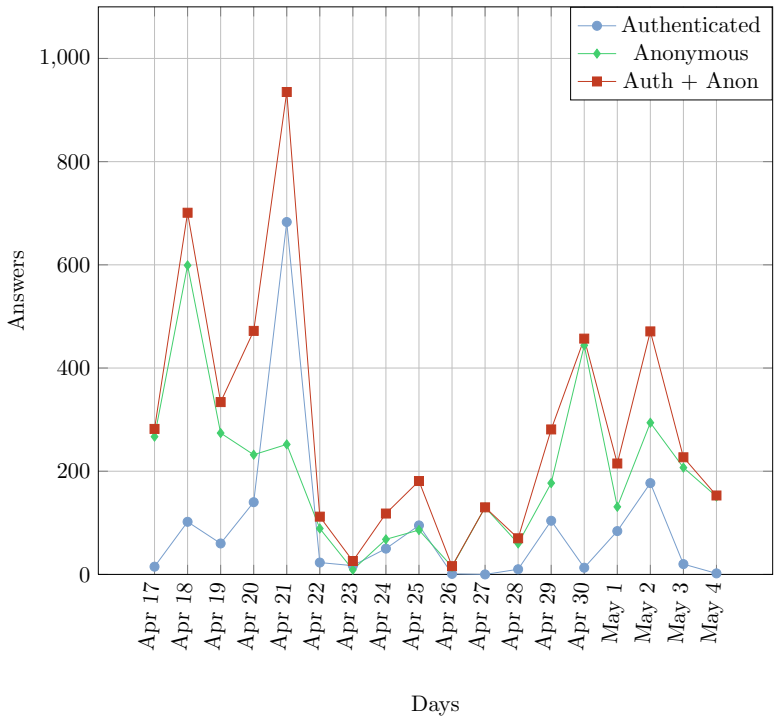


FIGURE 4.5: Answers per Day.

Figure 4.6 illustrates the distribution of answers per time of day. We note that, at least for Algerian community, between 4 PM and 8 PM, the participations are higher. It leads us to advice that the periodic open calls will be more efficient when launched in this time period.

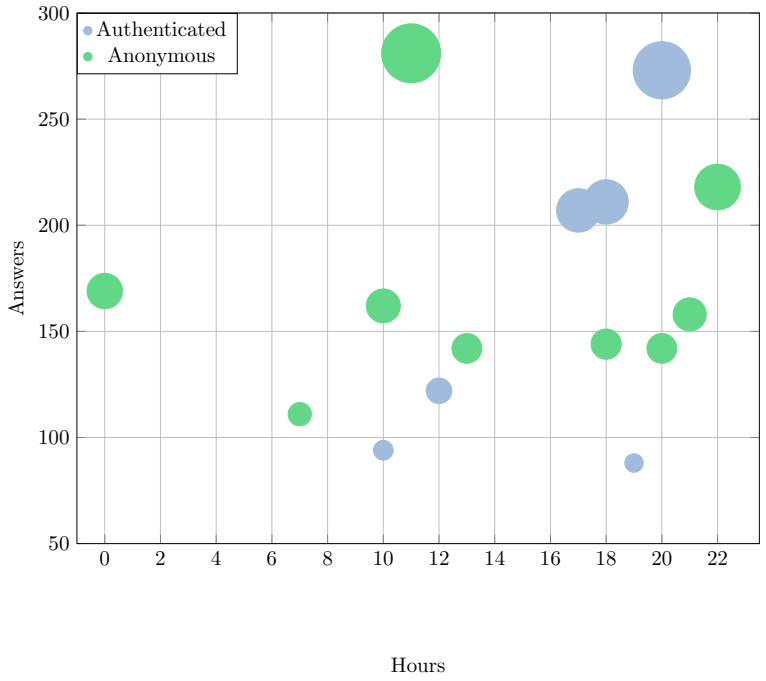


FIGURE 4.6: Answers per Time of Day.

4.5 Best Practices

Relying on this experience with altruistic crowdsourcing, we propose a set of guidelines to follow when designing crowdsourcing project using free platforms:

- *Make it easier, Clear & Reduce burdens*: In unpaid crowdsourcing, imposing constraints to workers is not recommended as it is a voluntary work. The contributor must find all instructions that should be followed to do correctly the task in both open call and Interface. In addition, the crowd prefer to work as anonymous contributor. For that, do not oblige them to authenticate.
- *Motivation Tricks*: The requester must well motivate the potential workers. For example, the requester has to address the open call to the most motivated community to a domain. Choosing "Community-help" motivation is often a worthwhile cause.
- *Select Dialect Area and Users Skills*: The requester must open call to users located in targeted dialect Area and with the language skills needed for dialectal annotation.
- *Stronger Advertising Campaign*: Contrary to paid crowdsourcing platforms where workers labor is rewarded which attracts a large number of contributors. The latter, large

number of contributors, is the key of Atruistic crowdsourcing success, it can be achieved using the power of Social Media.

- *Permanent Monitoring of the Project Execution*: As unpaid platform, daily observation must be done to check the progress of completed tasks to recall new volunteers when it is needed. In addition, the time of launching such open calls must be considered to get large participation.

- *Fill in the lack of worker' profile information* Information on worker' profile are not offered by unpaid platforms, for that, the requester has to consider that when designing quality control metrics in order to avoid malicious inputs and get best annotation results.

4.6 Conclusion

In this chapter, we have shown that using altruistic crowdsourcing with a simple quality control mechanism is an effective way for speech dialect annotation. In fact, 81% of annotations were accurate and similar to the Gold Standard annotation done by experts.

However, dealing with altruistic crowdsourcing needs to consider its specificities. First, unpaid crowdsourcing completes slower compared to paid crowdsourcing. Second, the requester has to take care of many concerns such project advertising, permanent monitoring of the project execution and implementing efficient quality control technique. For that, we have determined a list of best practices for crowdsourcing project when using free platforms.

Part II

Spoken Dialect Identification

Chapter 5

Preliminaries and Related Work

Contents

This chapter recalls some preliminaries on speech processing and language identification fields. In addition, we focus on reviewing some related work that have investigated Arabic dialect identification.

5.1 Introduction

Language IDentification (LID) problem can be viewed as a language recognition or a language detection (or verification) task. LID is the process where the language is identified automatically, whereas the task of language detection is to decide if the claimed language is the true language (Hanani 2012). LID can be text-based or speech signal-based. The text-based LID is considered a solved problem compared to Signal-based one (SLID), which is an active area of research (Rao et al. 2015). In the context of this thesis, LID is done from speech segments. The earliest SLID system was done by Leonard and Doddington (Leonard & Doddington 1974).

Automatic Spoken Dialect IDentification (SDID) can be considered as a special case of spoken LID, but it is featured by its complexity because it deals with the subdivisions of the same language, as opposed to SLID (Torres-Carrasquillo et al. 2004). In fact, the dialects of the same language basically use the same vocabulary, syntax, and semantics, while the variations is observed only in pronunciation patterns and phonology, and a slight variation is observed in grammar (Chittaragi et al. 2018). Most studies define dialect as accent. However, the accent refers to the variations in pronunciation only, and is thus distinct from dialect, which is characterized by variations in grammar, syntax, pronunciation, and lexical (Crystal 2008).

The potential applications for SDID are similar to those in the area of SLID. In fact, SDID is useful for both academia and industry for its promising impact on society. It also provides important benefits for speech technology. Robust SDID is used to improve speech recognition systems, which exist in most today's electronic devices. SDID is also enhancing the human-computer interaction applications. In addition, it should also prove useful in new services for e-health and telemedicine, in text-to-speech synthesis to produce regional speech, and in spoken dialogue systems' development (Biadsy et al. 2009, Etman & Beex 2015) .

5.2 Challenging Issues in Automatic Arabic Dialect Identification

In contrast to the language identification scenario, dialects typically share a common phonetic inventory[1] and written language (Shon et al. 2017). Thus for generating dialect models for SDID, it should be important to insure that no test speaker's information is presented in the train set to avoid their influence on the decision. Indeed, between two utterances there may exist differences in text, speaker, environment, and dialect. Hence, the main problem is to find a reliable SDID system deriving the dialect differences apart from the text, speaker, and environment differences. Rao et al. (2015) give some important issues in SLID, which are also issues for SDID. More details are given in what follows:

- *Variation in speaker characteristics*: Within the same language, there is speaker variability as results of the presence of different speakers and speaking styles. For that when creating language modeling, it is essential to ignore the speaker variation.
- *Variation in environment and channel characteristics*: Characteristics of speech signal are strongly affected by where the data are collected (the environmental conditions of recording) and on where they are transmitted (channel). Thus, to achieve the robustness for SDID system, it is necessary to select the features, which are less influenced by the channel and environment.
- *Similarities between dialects*: There is a lot of similarity in Arabic dialects as most of the dialects are derived from the Arabic language. Therefore, developing an SDID system for Arabic dialects is a challenging task.
- *Extraction and representation of dialect-specific prosody*: dialects have discriminate prosodic features, which are intonation, duration, intensity, stress, and rhythm patterns. For speech processing, there are no standard proper techniques available to represent the prosody, which is high level source knowledge. Therefore, it is difficult to extract and represent the prosodic information of specific language/dialect.

[1]Inventory: in linguistics, this term is used to refer to an unordered listing of items that describe a language; e.g. the listing of the phonemes of English would constitute that language's "phonemic inventory" (Crystal 2008).

5.3 Overview of an Automatic Spoken Dialect Identification System

In this section, we given an overview of SLID system as described by Ambikairajah et al. (2011), which is the same architecture for SDID system. In fact, as indicated previously, SDID can be considered as a special case of SLID with some additional challenges.

As described by Figure 5.1, an SLID system based on machine learning is divided into two steps: the *front-end*, which is the extraction of speech information that characterized the language, thus parameterizing the speech waveform, and a *back-end* step that generates the language models, $\{\lambda_l \mid l = 1, 2, ..., L\}$, where L is the total number of languages, as illustrated in Figure 5.1. The language models are created using a modeling technique to capture the characteristics of each language.

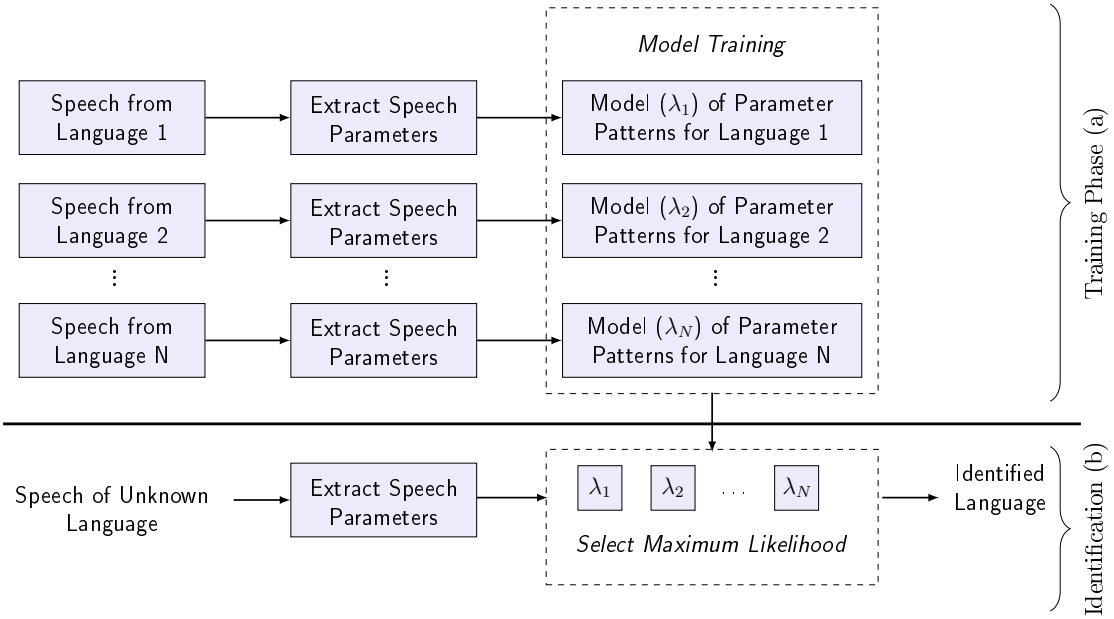


FIGURE 5.1: Basic Block Diagram of an LID System.

The main purpose of the front-end step is to extract the most discriminate information between languages from the speech waveform and discard as much of the redundant information such as as possible speaker and noise.

Like other pattern recognition problems, SLID consists of two phases (a) Training: Building language/class model, and (b) Testing: Scoring the models and the selection of the highest scored language. During the training phase, Figure 5.1 (a), each training speech utterance is converted into a stream of feature vectors X, which represent the result of front-end step,

which are distinctive features extracted from a given training dataset. These feature vectors are then used to train a separate model, λ_l for each language l, by using a modeling technique.

In the identification phase, Figure 5.1 (b), the same kind of speech distinctive features are extracted from the unknown speech utterance. This feature set is then evaluated by the trained models, $\{\lambda_l \mid l = 1, 2, ..., L\}$, to extract scores. With the evaluation scores, the system makes the language identity decisions

Although the research in recent years has focused on SLID, it is still a challenge to build a reliable SLID system. These latters are categorized by the features they use: acoustic–phonetic, phonotactic, prosodic, and lexical approach (Li et al. 2013). In the following section, we discuss the possible features that can be deployed in language identification.

5.4 Discriminative Speech Information for Language Identification

A fundamental issue of automatic language identification is to determine the efficient discriminative features to characterize languages. Humans are the best accurate LID system in the world today (Rao et al. 2015). Simply, within a one or two seconds of listening to a speech of their familiar languages, we are able to identify the language by using the higher level knowledge on languages, such as vocabulary, syntax, grammar and sentence structure. In the absence of the higher level knowledge, humans listeners relies on lower level features such as the prosody, phonetic and phonotactics (Muthusamy et al. 1994, Van Leeuwen et al. 2008).

A dialect or a language can be distinguished from a another by means of many characteristics extracted from the knowledge of a language: acoustic/phonetic, phonotactic, prosodic, lexical and syntactic (Li et al. 2013). These levels are from the lowest to the highest language knowledge. Figure 5.2 illustrates the levels of language knowledge used for an SLID. Generally, the language knowledge can be divided into the spoken level and word level. At the spoken level, the acoustic, phonetic, phonotactic and prosodic information are integral components of spoken aspect of a language. At the word level, differences between languages can be exploited based on morphology and sentence syntax. The levels are elaborated below, most information are extracted from the tutorial of SLID (Ambikairajah et al. 2011).

- *Acoustic Phonetics*: Acoustic information is one of the most primitive information which can be extracted directly from the raw speech when analyzed in physical terms,

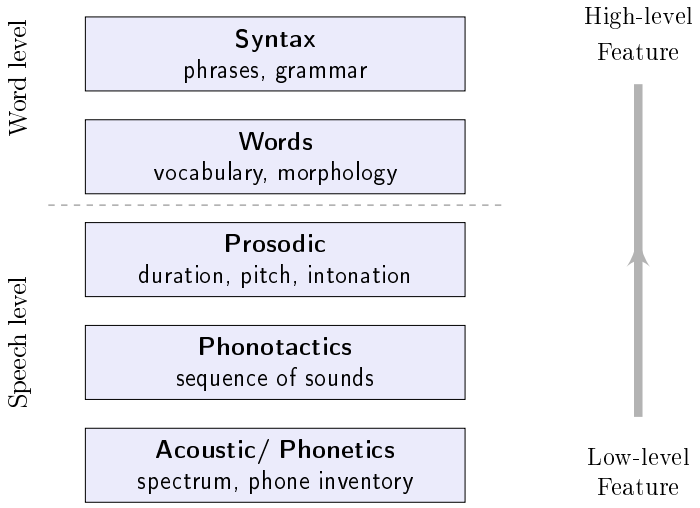


FIGURE 5.2: Levels of LID Features.

e.g. fundamental frequency, amplitude, harmonic structure (Crystal 2008, Wong 2004). The acoustic information is typically done by extracting the spectral features from segments of speech (Navrátil 2006). The most widely used acoustic features are Mel Frequency Cepstral Coefficient (MFCC) (Davis & Mermelstein 1980), Perceptual Linear Prediction (PLP) and Linear Prediction Cepstral Coefficient (LPCC). To add the temporal aspects to the acoustic feature vector, Delta and Shifted Delta Cepstrum (SDC) features are used (Rao et al. 2015). In linguistic system, speech sounds, as entities, are the phonemes. Each language has its unique set of phonemes. For instance, Dutch and German have the phoneme /x/, whereas French and Italian do not (Kirchhoff 2006).

- *Phonotactics*: refer to the combinations of a set of phonemes in a language. Some phoneme sequences common in one language could be rare in another language. For example, the phoneme sequence /st/ is very common in English, while impermissible in Japanese.

- *Prosody*: As mentioned before, the prosody is the musical expression of speech that a speaker produces when he speaks. The main aspects of prosody is the tone, stress, and rhythm. In fact, the fundamental frequency, intensity, and duration sequence are used for indicating tone, stress, and rhythm respectively. The prosodic information can be used to discriminate Arabic dialects. More details are mentioned in Section 1.4.2.

- *Words*: At the word level, the information are extracted from the structure of words.In fact, each different language has its own sets of vocabularies and its own morphology rules.

- *Syntax*: Syntax is the study of the rules that govern the way words are combined to form sentences.The sentence patterns vary across different languages.

Morphological and syntactic information are currently not common used information in LID systems.

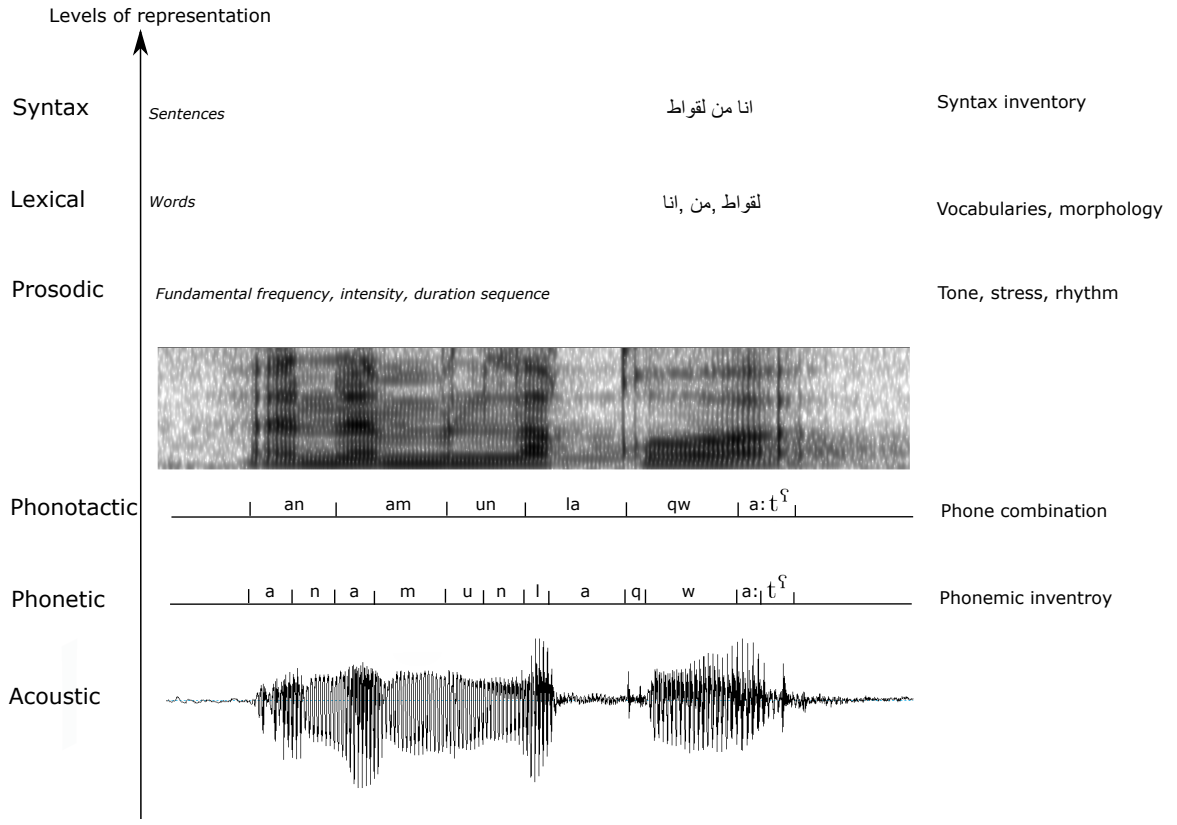


FIGURE 5.3: Illustration of Language Information at Different Levels of Representation.

Figure 5.3 gives an example of language information at different levels of the spoken Algerian dialect of the sentence أنا من لقواط /'ana mun laqwaṭ/ (I am from Laghouat).

SLID systems can use some or all of the above language knowledge to estimate the similarity between a speech and different targeted language and then combined these measures to make the final decision. Figure 5.4 reports a block diagram of a generic LID system that use all levels of information. However, it is not necessary that an SLID system use all language knowledge. The most popular approaches use the combination between the acoustic and phonotactic information (Ambikairajah et al. 2011).

5.5 Spoken Arabic Dialect Identification

In the middle of nineties, Zissman et al. (1996) presented the pioneered research on SDID. They employed a phone recognition system combined with language modeling to classify

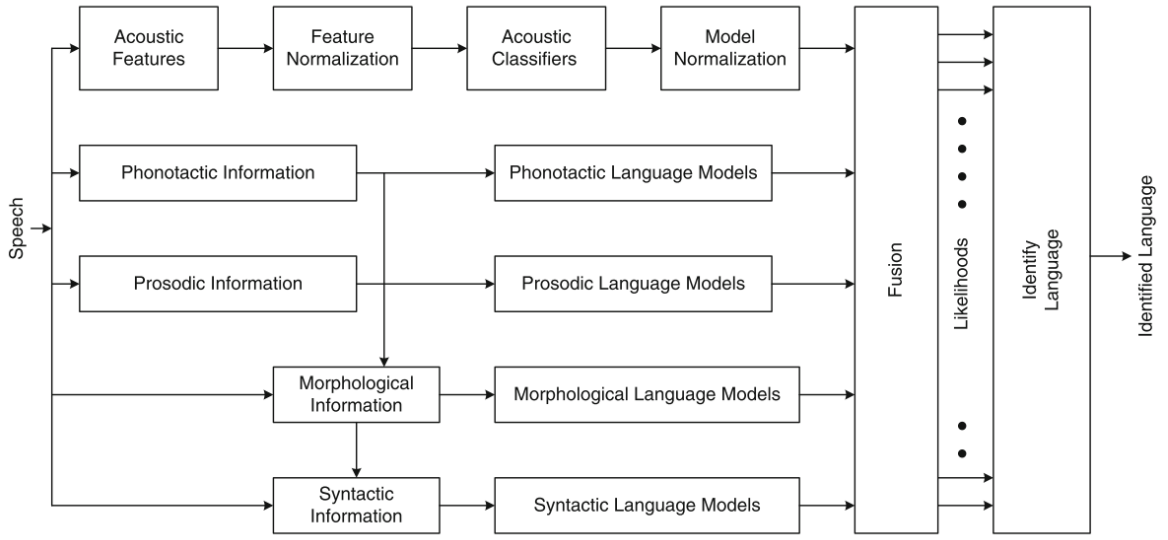


FIGURE 5.4: Block Diagram of Generic LID System using the Features from Various Levels (Rao et al. 2015).

Cuban and Peruvian dialects of Spanish. The first Arabic SDID (ASDID) system was authored by Rouas et al. (2006), were carried using Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM), which has a history extending back to the 2006s. Lack of speech datasets/corpora for Arabic dialects dedicated to scientific research purposes was the main cause of the scarce research on ASDID (Bougrine et al. 2016, Zaghouani 2014). Furthermore, there is even less standard datasets. For this reason in what follows, we describe the studied approaches without considering their achieved performances intentionally as they deal with different datasets.

Narrowing our focus into Arabic SDID, we could categorize prior world based on two aspects: *speech feature Level* and *Intra/Inter country dialect*. The first aspect indicates from which level the speech features are extracted. In general, the pre-lexical levels are the most frequently used to identify dialects from speech. Hence, we consider acoustic/phonetic (Acous/-Phone), phonotactic (Phono), and prosodic (Proso) levels that are exploited alone or combined. The second criterion, distinguishes the origin of targeted dialects from either *Intra-country* or *Inter-country* context, which means that the studied dialects are from the same country/region or dialects from many countries. This criterion is chosen due to the level of difficulty and complexity. Indeed, SDID dealing with dialects from the same country or geographical area is much more difficult than cross-countries. In Figure 5.5, we categorized the main existing Arabic SDID systems according to these criteria.

Firstly, we summarize the ASDID systems where acoustic/phonetic models are used alone (Al-Ayyoub et al. 2014, Alorfi 2008, Biadsy et al. 2011, Lachachi & Adla 2015, Moftah et al. 2018).

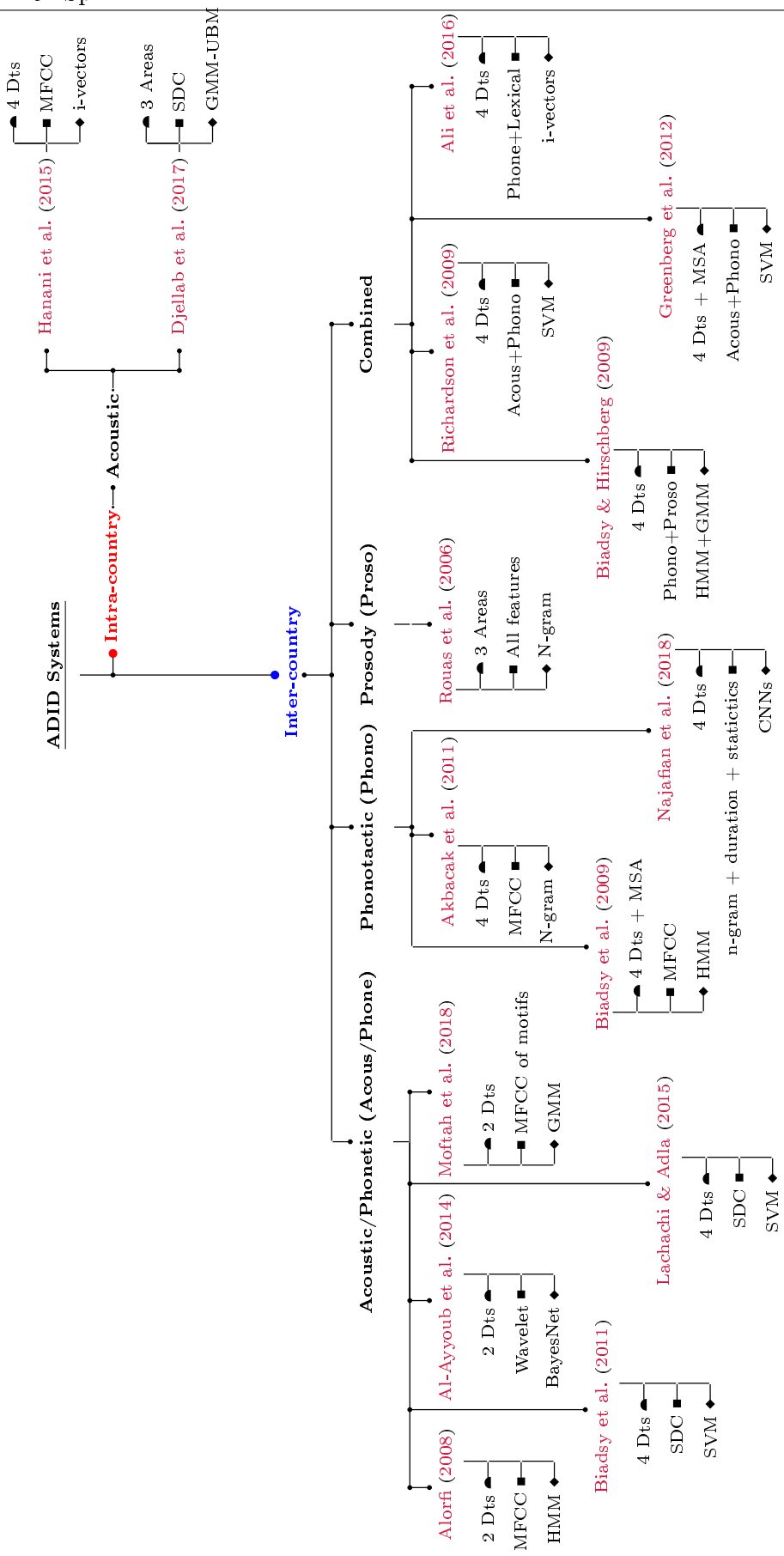


FIGURE 5.5: Classification of ADID Systems According to two Criteria: Speech Feature Level and Intra/Inter Country.

Most of the reviewed work are based on spectral features, which are Mel Frequency Cepstral Coefficient (MFCC) or Shifted Delta Cepstrum (SDC), with Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) used for dialect modeling.

Alorfi (2008) proposed a model that exploits different acoustic/phonetic features using HMM. The model associates two states for each dialect representing common and unique sounds respectively. The model was used to identify two Inter-country dialects: Egyptian and Gulf.

Biadsy et al. (2011) used the phone labels segmentation to constrain the acoustic models. The dialect models are generated using the SVM classifier with special Kernel function, and validated the approach on four Arabic Inter-country dialects: Iraqi, Gulf, Levantine, and Egyptian.

In Al-Ayyoub et al. (2014), an acoustic model using fixed size segmentation was used to extract the selected wavelet features. They dealt with two dialects Jordanian and Egyptian. The results achieved were not conclusive due to the limited size and the quality of the dataset used in their experiments.

Recently, Moftah et al. (2018) present a technique based on pattern. In fact, they characterize each dialect by the most repeated acoustic sequences. Then, they extract their corresponding MFCC and use GMM as classifier. They applied their approach on two Arabic dialects: Levantine and Egyptian.

For the context of Magrebian dialects, Lachachi & Adla (2015) deployed the reducing Universal Background Model (UBM) to support special SVM classification. They reduced the size of dataset using the Minimal Enclosing Ball method by means of a fuzzy C-mean clustering algorithm. They used a dataset containing five sub dialects spoken in: the three Algerian departments (Oran, Algiers, Constantine), Morocco and Tunisia.

In contrary to other acoustic/phonetic approaches, Djellab et al. (2017) and Hanani et al. (2015) were the only relevant work that proposed ASDID systems for Intra-country context. In Hanani et al. (2015), the authors employed i-vectors method on acoustic information for regional accents recognition. They experimented the approach on Arabic Palestinian accents from four regions, namely: Jerusalem, Hebron, Nablus and Ramallah. On the other hand, Djellab et al. (2017) designed a GMM-UBM and an i-vectors-based framework for accent recognition. They used the implemented approach on a set of data spoken in three Algerian areas: the East, Center and West of Algeria.

Otherwise, using phonotactic cues, few attempts were made for building ASDID by deploying Parallel Phone Recognition and Language Modeling (PPRLM) (Akbacak et al. 2011, Biadsy

et al. 2009), and recently Najafian et al. (2018) use the deep learning technique. In Biadsy et al. (2009), the authors applied PPRLM using nine (Arabic and non-Arabic) phone recognizers. They performed the experiments on a large dataset of four Arabic dialects (Egyptian, Gulf, Iraqi, Levantine) together with MSA. Akbacak et al. (2011) designed an approach that combines three models to identify four Arabic dialects (Iraqi, Gulf, Levantine and Egyptian). The models are cepstral GMM, PPRLM, and Phone Recognition modeled via SVM. The combination is carried-out at the score-level.

Najafian et al. (2018) proposed a novel feature representation for phonotactic approach, which add to the same phone n-gram sequence the phone duration and it probability statistics that characterize the phonemes. They used SVM and Convolutional Neural Networks (CNNs) as classifiers. They applied this approach on four Arabic Inter-country dialects: Maghrebi, Gulf, Levantine, and Egyptian.

Some other frame work exploited combinations of acoustic/phonetic and phonotactic features to perform ASDID (Greenberg et al. 2012, Richardson et al. 2009). In Greenberg et al. (2012), the authors based on three spectral similarities and n-grams they designed a combined approach using four core classifiers. The combination is done at the back-end level of the system using Bayes classifier. They targeted a set of 24 languages containing four variations of Arabic language, which are MSA and three dialects: Iraqi, Levantine and Maghrebi. Richardson et al. (2009) presented a method that use the discriminative n-gram selection algorithm. They evaluated the classifiers on different group dialects, which are English (Indian/American), Mandarin Chinese (Taiwan/mainland) and Arabic (Gulf/Iraqi/Levantine). They concluded that SVM classifier has achieved best results for Arabic dialects.

Ali et al. (2016) designed an approach based on *i*-vectors method that combined phonetic and lexical features. The Arabic Broadcast speech dataset was used in their experiments, which includes four Arabic dialects Egyptian, Gulf, Levantine, and Maghrebi .

On the other side, we have observed that there are few attempts for prosody-based ASDID. Based on a previous work of Ghazali et al. (2002), it was shown that some Arabic dialects (Syria, Jordan, Morocco, Algeria, Tunisia and Egypt) can be grouped using rhythmic information in three dialectal areas: Maghreb (Morocco, Algeria), Middle-East (Syria, Jordan), and an intermediate one (Tunisia, Egypt). Rouas et al. (2006), in a folow-up work, have designed an ASDID system for three previous dialectal areas. The approach collected all prosodic information: intonation, rhythm and stress. Thus, they used a segmentation which is based on consonant/vowel location to get the approximative structure of the syllable. They have utilized a special codification to represent the duration of phonemes and the energy instead of the real values. Then, they classified dialect areas using multi-gram models where

grams are their pseudo-syllables and each area's dialect is represented by the most frequent sequences of n-gram. Unfortunately, they tested their system on a small dataset.

In Biadsy & Hirschberg (2009), the authors combined the prosodic and phonotactic approaches for competing each-other to perform more accurate identification, where they augmented their phonotactic system, described above, by adding some prosodic features like durations and fundamental frequency measured at n-gram level where grams are syllables. They tested their system on four Arabic dialects: Gulf, Iraqi, Levantine, and Egyptian.

Recently, Arabic NLP shared tasks are included SDID to discriminate between five Arabic varieties, namely MSA and four Arabic dialects: Egyptian, Gulf, Levantine, and Maghrebi, by combined text and acoustic speech information. Those competitions include Multi-Genre Broadcast (MGB-3) challenge (Ali et al. 2017) and VarDial evaluation campaign (Zampieri et al. 2017).

Table 5.1 presents the literature overview of ASDID according the targeted dialect and level of speech information. In the light of this exhaustive review of the most relevant ASDID systems, we highlight the fact that a little attention has been paid to Algerian ASDID, only two researches one in intra-country dialects (Lachachi & Adla 2015) and the other in inter-country dialects (Djellab et al. 2017). In addition, we confirm that there is a lack of ASDID system that exploit the prosodic information. In fact, only Rouas et al. (2006) and Biadsy & Hirschberg (2009) have considered using such information. In Rouas et al. (2006), the authors investigated dialects within the same area. While in Biadsy & Hirschberg (2009), the authors focused on inter-country dialects.

All of the proposed ASDID systems use conventional classification method in occurrence SVM, HMM-GMM, N-gram. However, Deep Learning is still a novelty in this research area. Especially as result of their is not investigated despite their provided efficiency for spoken language identification (Ferrer et al. 2016, Lopez-Moreno et al. 2016).

5.6 Conclusion

In this chapter, we have introduced the Spoken Dialect IDentification (SDID) problem and its challenging issues. Second, we have given an overview of an automatic SDID system and we have explored the discriminative features to build an efficient SLID system. Then, we have attempted to provide a summary of the literature relating to Arabic SLID. It has been shown from this review that only one research has considered intra-country Algerian dialects. In addition, the existing literature on Arabic SDID is extensive and has particularly investigated

Dialect	Acous/Phone	Phono	Proso	Acoust+Phono	Phone+Lexical	Phono+Proso
Egyptian	Alorfi (2008), Biadsy et al. (2011), Al-Ayyoub et al. (2014), Moftah et al. (2018)	Biadsy et al. (2009), Akbacak et al. (2011), Najafian et al. (2018)			Ali et al. (2016), Ali et al. (2017), Zampieri et al. (2017)	Biadsy & Hirschberg (2009)
Gulf	Alorfi (2008), Biadsy et al. (2011)	Biadsy et al. (2009), Akbacak et al. (2011), Najafian et al. (2018)		Richardson et al. (2009)	Ali et al. (2016), Ali et al. (2017), Zampieri et al. (2017)	Biadsy & Hirschberg (2009)
Iraqi	Biadsy et al. (2011)	Biadsy et al. (2009), Akbacak et al. (2011)		Greenberg et al. (2012), Richardson et al. (2009)		Biadsy & Hirschberg (2009)
Levantine	Biadsy et al. (2011), Moftah et al. (2018)	Biadsy et al. (2009), Akbacak et al. (2011), Najafian et al. (2018)		Greenberg et al. (2012), Richardson et al. (2009)	Ali et al. (2016), Ali et al. (2017), Zampieri et al. (2017)	Biadsy & Hirschberg (2009)
Jordanian	Al-Ayyoub et al. (2014)					
Palestinian	Hanani et al. (2015)					
Morrocan	Lachachi & Adla (2015)					
Tunisian	Lachachi & Adla (2015)					
Algerian	Lachachi & Adla (2015), Djellab et al. (2017)					
Maghrebi [a]		Najafian et al. (2018)	Rouas et al. (2006)	Greenberg et al. (2012)	Ali et al. (2016), Ali et al. (2017), Zampieri et al. (2017)	
Middle-East Area[b]			Rouas et al. (2006)			
intermediate Area[c]			Rouas et al. (2006)			

[a] The Maghrebi gathered group of dialects, we kept it because the authors of related work were not specific.
[b] Syria + Jordan.
[c] Tunisia + Egypt.

TABLE 5.1: Literature Overview of Spoken Arabic Dialect Identification.

on acoustic/phonetic and phonotactic based approach, contrary to the prosodic one which has less attention. In the chapter that follows, we will describe some statistic analysis of prosodic information for Algerian dialects.

Chapter 6

Statistic Analysis of Prosodic Information For Algerian Dialects

Contents

In this chapter, we describe some prosodic features related to intonation and rhythm phenomena, which are used in the chapter that follows to build our SDID systems. Then, we explain how we have extracted them and the used Open Source softwares. Finally, we perform some statistical analysis of these prosodic phenomena for Algerian dialects to support their use motivation.

6.1 Prosodic Information for Dialects

In order to identify a language/dialect, many features and measurements are developed in literature to capture prosody inherent to a speech. In general, the pitch and duration sequence are used for indicating intonation and rhythm respectively.

6.1.1 Rhythm Features

Rhythm derives from the repetition of elements, which are in speech: syllables, or stressed syllables. Rhythm also refers to aspects of temporal organization of speech. Historically, James (1940) is the first research claimed that the languages have different rhythms. Pike (1945) and Abercombie (1967) classified languages into two rhythm groups: stress-timed and syllable-timed languages. In stress-timed languages, as in English, Arabic, and German, the stressed syllables are repeated at regular intervals of time (stress-timing). In syllable-timed languages, as in French and Spanish, the syllables are occurred at regular intervals of time. A third rhythm group was later added by Peter (1975), which is mora-timed languages, such as Japanese and Tamil, it is due to the recurrence of mora, which is a rhythmic unit that can be a sub-syllabic[1] that includes onset and nucleus, or a coda (Crystal 2008, Nespor et al. 2011).

In Dauer (1983), the authors observed that various phonological properties, such as vowel reduction and syllable complexity, can distinguish between the two group languages: syllable-timed languages have a smaller variety of syllable types than stress-timed languages, and they do not display vowel reduction. These phonological properties allow to use vocalic and consonantal durations to classify languages by rhythm phenomenon. Thus, different Rhythm Measures (RMs) have been developed to successful classifying the languages into three groups: stress-timed, syllable-timed, and mora-timed (White & Mattys 2007).

[1] Onset, nucleus and coda are component of a syllable.

The first and the most popular RM used to classify a language is the Interval Measures (IM) (Ramus et al. 1999). IM metrics include three separate measures: ΔV, the standard deviation of vocalic intervals; ΔC, the standard deviation of consonantal intervals; %V, the duration proportion of vocalic interval. The results of this study showed that the combination of %V and ΔC provides the best correlate feature to separate traditional rhythm categorizations of languages (stress-timed, syllabic-timed, mora-timed). Figure 6.1 reports the results of this combination for eight languages. Stress-timed languages, such as English and Dutch, has full and reduced vowels. However, syllable-timed languages, such as French, and Spanish does not have vowel reduction. In other words, the percentage of vocalic sequences (%V) of stress-timed language is smaller than in syllable-timed language. Moreover, ΔC is larger in stress-timed language and reflect more complex syllable structure.

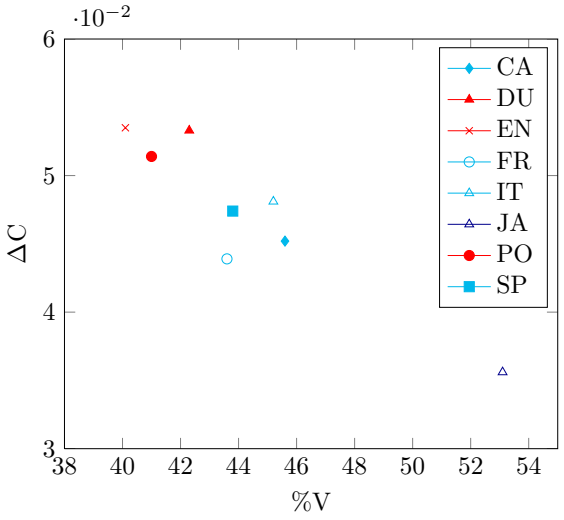


FIGURE 6.1: %V versus ΔC for 8 Languages, reproduced from Ramus et al. (1999). (CA= Catalan, DU= Dutch, EN= English, FR= French, IT= Italian, JA= Japanese, PO= Polish, SP= Spanish)

Dellwo (2006) noted that ΔC and ΔV metrics were strongly correlated with speaking rate, for that he proposed their normalized version the VarcoC and VarcoV metrics for ΔC and ΔV respectively. VarcoV and VarcoC are the standard deviation of vocalic (ΔV) and consonantal intervals (ΔC) divided by the mean vocalic and consonantal interval duration respectively.

Grabe & Low (2002) proposed an additional metric, the Pairwise Variability Index (PVI), which focuses on the temporal succession between the consonantal and vocalic intervals of the global utterance, which is used to capture the variability in syllable duration. They shown also that, the best combination to classify languages according to the rhythm is the raw PVI ($rPVI$) for the consonantal intervals and the normalized PVI ($nPVI$) for vocalic intervals, $rPVI$ and $nPVI$ are defined by equations (6.1) and (6.2) receptively, where m is the number

of vocalic/consonantal intervals in the global utterance, and d_k is the duration of the k^{th} interval.

$$rPVI = \sum_{k=1}^{m-1} |d_k - d_{k+1}| / (m-1) \tag{6.1}$$

$$nPVI = \sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| / (m-1) \tag{6.2}$$

Figure 6.2 illustrates the distribution of intervocalic $rPVI$ and vocalic $nPVI$ of six languages reproduced from Grabe & Low (2002). They found that vocalic $nPVI$ scores are clearly discriminative between stress-timed and syllabic-timed languages compared with intervocalic $rPVI$ scores. For example, English, German, and Dutch (stress-timed) have higher vocalic $nPVI$ scores than for Spanish and French (syllabic-timed). Mora-timed language (Japanese) is not clearly distinguished from syllabic-timed and stress-timed languages by using PVI scores, in contrast with Ramus et al. (1999).

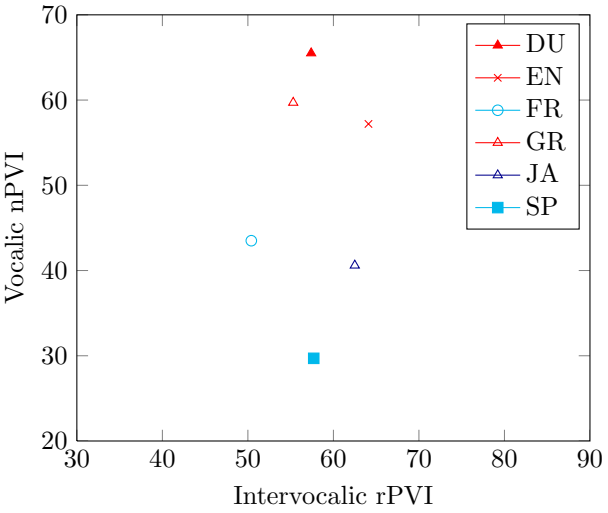


FIGURE 6.2: Intervocalic rPVI versus Vocalic nPVI for 6 Languages, reproduced from Grabe & Low (2002). (DU= Dutch, EN= English, FR= French, GR= German, JA= Japanese, SP= Spanish).

These RMs differ in three respects. First, the nature of targeted intervals, which can be the durations of vocalic or consonantal intervals. Second, they assume *global* or *local* forms. Global RMs capture the time period variation of determined intervals over an entire utterance. Local measures focus on differences between two consecutive intervals and then average those differences over the utterance. Third, some RMs include a normalization step, which is used

to avoid the changes in speech tempo, while others keep raw durations (Loukina et al. 2011). In Table 6.1, we report the used RMs, their description, and their characteristics according to the above aspects.

RM	Description	Type of interval	Scope	Normalization	Main Reference
%V	Percent of total utterance duration composed of vocalic intervals.	Ratio	Global	Yes	Ramus et al. (1999)
ΔV	The standard deviation of vocalic interval duration.	V	Global	No	Ramus et al. (1999)
VarcoV	Coefficient of variation of vocalic interval duration (i.e., standard deviation of vocalic interval duration divided by the mean), multiplied by 100.	V	Global	Yes	Dellwo (2006)
nPVI-V	Normalized pairwise variability index of vocalic intervals. Mean of the differences between successive vocalic intervals divided by their sum, multiplied by 100.	V	Local	Yes	Grabe & Low (2002)
ΔC	The standard deviation of consonantal interval duration.	C	Global	No	Ramus et al. (1999)
VarcoC	Coefficient of variation of consonantal interval duration (i.e., standard deviation of consonantal interval duration divided by the mean), multiplied by 100.	C	Global	Yes	Dellwo (2006)
rPVI-C	Raw Pairwise variability index for consonantal intervals. Mean of the differences between successive consonantal intervals.	C	Local	No	Grabe & Low (2002)

TABLE 6.1: The Classification of Seven Used Rhythm Metrics.

In addition, we consider an other metric: *Speech Rate*, which is the number of syllable per second.

6.1.2 Intonation Features

The pitch, or fundamental frequency (F_0), is used for indicating the intonation. The statistical modeling of intonation are used for each utterance. We take into consideration two groups of intonation metrics: Global information and total size of pitch trajectory. Global information metrics of the pitch are measured using quantiles of nucleus.

This group has four pitch values which are: *Bottom* is 2^{nd} quantiles, *Median* is 50^{th} quantiles, *Top* is 98^{th} quantiles of pitch nucleus in Hz, and *Range* is the difference between *Top* and *Bottom* in SemiTones (ST). Concerning the total size of pitch trajectory, this group has two metrics (TrajIntra, TrajInter) calculated using pitch trajectory, which is the sum of absolute intervals within/between (TrajIntra/TrajInter) syllabic nuclei divided by duration (in ST/s). Table 6.2 resumes these features.

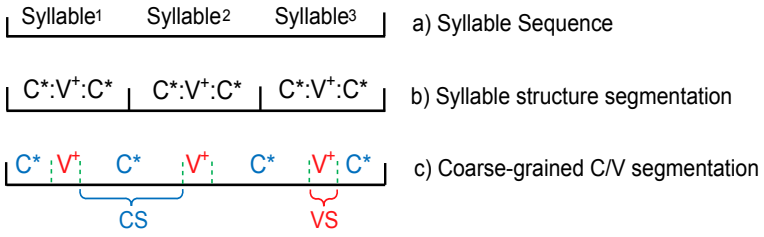
6.2 Segmentation and Extraction of Prosodic Features

The extraction of the prosodic features needs a consonant/vowel segmentation. However, for Arabic dialect the problem of phoneme segmentation is not completely solved and unfortunately there is no related decoders such as the case for other dialects. To cope with this problem, we rely on the specific Arabic syllable structure to perform a consonant/vowel segment as we explain in what follows.

Metric	Include	Definition
Global Information	*Bottom*	2^{nd} quantiles pitch nucleus
	Top	98^{th} quantiles pitch nucleus
	Median	50^{th} quantiles pitch nucleus
	Range	Difference between *Top* and *Bottom*
Total Size of Pitch Trajectory	*TrajIntra*	Pitch trajectory of nuclei / duration
	TrajInter	Pitch trajectory between nuclei / duration

TABLE 6.2: The Used Intonation Metrics.

A syllable is the unit of pronunciation between a phoneme and a word. It is divided into three components: the opening *Onset*, the central *Nucleus* and the closing segment *Coda*. The nucleus and coda are called rhyme (or rime) (Crystal 2008). In Arabic dialect, each syllable contains at least a nucleus while coda, and onset are optional. The nucleus is imperatively one or many vowels, and the others are consonants. Thus, an Arabic syllable structure can be modeled by the regular expression $C^*V^+C^*$ where C is a consonant and V is a vowel. We exploit this fact to perform a consonant/vowel segmentation. Thus, we sketched up these measures by considering nuclei as vowel segments and what remains as consonant segments. We name this segmentation by coarse-grained segmentation. For example, we consider the following sequence of the three syllables (a). The syllable structure segmentation is presented in (b). Then, we get our coarse-grained segmentation (c), where *VS* (resp. *CS*) represents vowel (resp. consonant) segment.



Once the consonant/vowel segmentation is performed, our prosodic features are extracted at utterance level. Whereas, previous prosodic-based ADID researches leveraging prosody extracted prosodic features at pseudo-syllable level (Biadsy & Hirschberg 2009, Rouas et al. 2006).

6.3 Used Tools

In order to extract prosodic information, we have involved a set of Open Source softwares. We have used Praat tool[2] to remove noise. Our coarse-grained segmentation is done automatically using Praat tool enhanced by Prosogram[3] script. The latter performs syllable segmentation using the intensity of the band-pass filtered signal. The syllable' nuclei are delimited using spectral and amplitude changes. The script can achieve over 80% according to segmentation accuracy (Salselas & Herrera 2011).

The intonation metrics and speech rate are extracted using Prosogram script. While, rhythm features are calculated using Correlatore[4], which is mainly designed for rhythm analysis.

[2] Praat v 5.3.47, Online: http://www.fon.hum.uva.nl/praat
[3] Prosogram v 2.9, Online: http://bach.arts.kuleuven.be/pmertens/prosogram/
[4] Correlatore v 2.1, Online: http://www.lfsag.unito.it/correlatore/

In addition, we have scripted the combination of all these tools by using Python[5] language and the statistical analysis tests is done by using R[6] language.

6.4 Prosodic Statistic Analysis of Algerian Dialects

In this section, we first described our selected Algerian Arabic dialects data. Then, we describe and analyze the statistic analysis of prosody in Arabic Algerian dialects utterances on both prosodic phenomena: rhythm and intonation. In order to study the discriminative power of prosody, we present in this section the description of our statistic analysis of prosody in Algerian dialects. we have used two statistical analysis tests: one way ANalyses Of VAriance (ANOVA) and Tukey HSD (Tukey Honest Significant Differences).

6.4.1 Speech Material

The Algerian dialects are under investigation in this study. As mentioned in previous Section 2.5, there is no standard corpus available for Algerian Arabic dialects such as for some other Arab countries (Alsulaiman et al. 2013). For this reason, we have collected own corpora Alg-Daridjah and Kalam'DZ.

For this statistical analysis, we have chosen Alg-Daridjah corpus, described before in Section 3.1, as test bed as it provides a representation of phonetic, prosodic, and orthographic varieties of Algerian Arabic dialects. It contains six Arabic Algerian dialects: Algiers-Blanks, Hilālī, Ma'qilian, Pre-Hilālī, Sahel-Tell, and Sulaymite. Table 6.3 gives more details on the sample chosen for the current analysis. The total number of utterances is 1892, each one is about 6s duration in average.

The speakers are chosen from adult population from 18 to 50 years old. The 41 talkers are native from their dialect region, and both parents of each speaker were also native from the same dialect region. The speeches gather both spontaneous and sub-spontaneous utterances.

6.4.2 Prosodic Statistic Analysis

In this statistical analysis, we have considered the prosodic features presented in the previous.After their extraction, we identify the rhythmic class of each Algerian dialects. Finally,

[5] The used packages are *subprocess*, *os*, *sys*, and *shutil*.
[6] The used functions are *aov* and *TukeyHSD*.

Dialect	#Utterances	#Speakers	#Female	Department
ALgiers-Blanks (ALB)	51	01	-	Algiers
HILālī (HIL)	657	14	09	Adrar, Djelfa, Ghardaïa, Laghouat
MA'Qilian (MAQ)	351	08	06	Sidi Bel Abbès, Mostaganem, Oran
PRe-Hilālī (PRH)	143	03	02	Tlemcen
SAhel-Tell (SAT)	187	04	04	Médéa
SULaymite (SUL)	503	11	05	El-Oued, Annaba
Total	1892	41	26	

TABLE 6.3: Speech Material Details.

we have performed two statistical analysis tests: one-way ANOVA and Tukeys HSD. ANOVA gives a single answer to the question *Do the data sets differ overall?* While, Tukeys HSD test can be run afterwards to answer the individual questions *Does data set A differ from data set B? Does data set A differ from data set C?* (Omer et al. 2018).

Table 6.4 presents the results of one-way ANOVA testing with $\alpha = 0.05$[7] (p-value[8]).

The results also reports the mean values and the standard error (in parentheses) calculated for each rhythm features: average proportion of the duration of vocalic intervals (%V), the average standard deviation of the duration of consonantal intervals (ΔC) and of vocalic intervals (ΔV) and their normalized version VarcoΔC and VarcoΔV respectively, raw pairwise variability index in consonantal interval duration (rPVI-C), and normalized pairwise variability index in vocalic interval duration (nPVI-V) and speech rate calculated across all utterances of each dialect.

Cross-dialectal comparison shows that the proportion of vocalic intervals (%V) represents less than 50% of the total duration of an utterance in all Algerian dialects. This result supports the claim of Hamdi et al. (2004) that all Arabic dialects have less than 50% of vocalic intervals (%V).

Particularly, Pre-Hilālī dialect exhibit the smallest proportion of vocalic intervals (%V) and the greatest variability in consonant interval duration (rPVI-C). This fact shows that it has the highest degree of vowel reduction and more complex syllable structure. The findings

[7] α is the significance level.

[8] Probability value (p-value), when the p-value $\leq \alpha$ indicates the reject the null hypothesis that all the data are sampled from populations with the same mean.

Feature	ALB	HIL	MAQ	PRH	SAT	SUL	F	p-value
% V	**47.5** (1.1)	40.4 (0.5)	43.3 (0.3)	**39.1** (0.7)	44.2 (0.6)	43.6 (0.3)	$F(5, 1886) = 17.1$	$p < .000$
ΔC	55.4 (3.3)	**97.1** (8.3)	58.3 (1.3)	86.5 (7.2)	59.8 (2.4)	**46.7** (1.2)	$F(5, 1886) = 10.3$	$p < .000$
ΔV	44.0 (2.6)	36.7 (0.7)	40.9 (1.2)	38.7 (1.4)	**45.7** (1.6)	**33.9** (0.7)	$F(5, 1886) = 14.8$	$p < .000$
VarcoC	62.9 (2.6)	63.2 (0.7)	63.1 (0.8)	**71.0** (1.7)	63.8 (1.3)	**58.1** (0.7)	$F(5, 1886) = 13.5$	$p < .000$
VarcoV	50.7 (2.3)	**50.2** (0.6)	52.7 (1.0)	52.4 (1.4)	**57.0** (1.3)	**49.5** (0.7)	$F(5, 1886) = 6.8$	$p < .000$
rPVI-C	61.4 (3.8)	**99.1** (8.2)	62.1 (1.3)	88.7 (6.2)	63.4 (2.3)	**51.5** (1.1)	$F(5, 1886) = 9.8$	$p < .000$
nPVI-V	54.0 (2.4)	51.86 (0.6)	51.9 (0.8)	53.9 (1.3)	**56.2** (1.2)	**49.5** (0.7)	$F(5, 1886) = 5.9$	$p < .000$
Speech Rate	**6.1** (0.1)	6.8 (0.1)	6.3 (0.1)	6.3 (0.1)	6.5 (0.1)	**7.2** (0.1)	$F(5, 1886) = 14.0$	$p < .000$

TABLE 6.4: Mean (Standard Errors) Values of Rhythm Feature for Arabic Algerian Dialects and the Results of One-way ANOVA Testing Effect of Dialects.

of the current study are consistent with those of Marçais (1986) who found that Pre-Hilālī dialect characterized by the presence of vowel reduction.

However, Sulaymite dialect shows the smallest variability in consonant intervals duration (ΔC, VarcosC, rPVI-C) and in vocalic intervals duration (ΔV, VarcosV, and nPVI-V). Thus, Sulaymite dialect is characterized by the absence of vowel reduction and simple syllable structure. These findings further support the note of Hamdi et al. (2005) that Tunisian dialect has a simple syllable structure. Sulaymite dialect shared many characteristic with Tunisian dialect because they are geographically close.

The vocalic interval measure (ΔV, VarcosV, and nPVI-V) resulting of the Sahel-Tell dialect are higher than the rhythm metric scores of others dialects.

According to the speech rate metric, we noticed a faster overall articulation rate for Sulaymite dialect, whereas a slower articulation rate for Algiers-Blanks dialect.

In addition to this attempt study to identify the rhythmic class of each Algerian dialects, we have compared these results to other languages (Algerian-MSA, English, and French) existed rhythmic features.

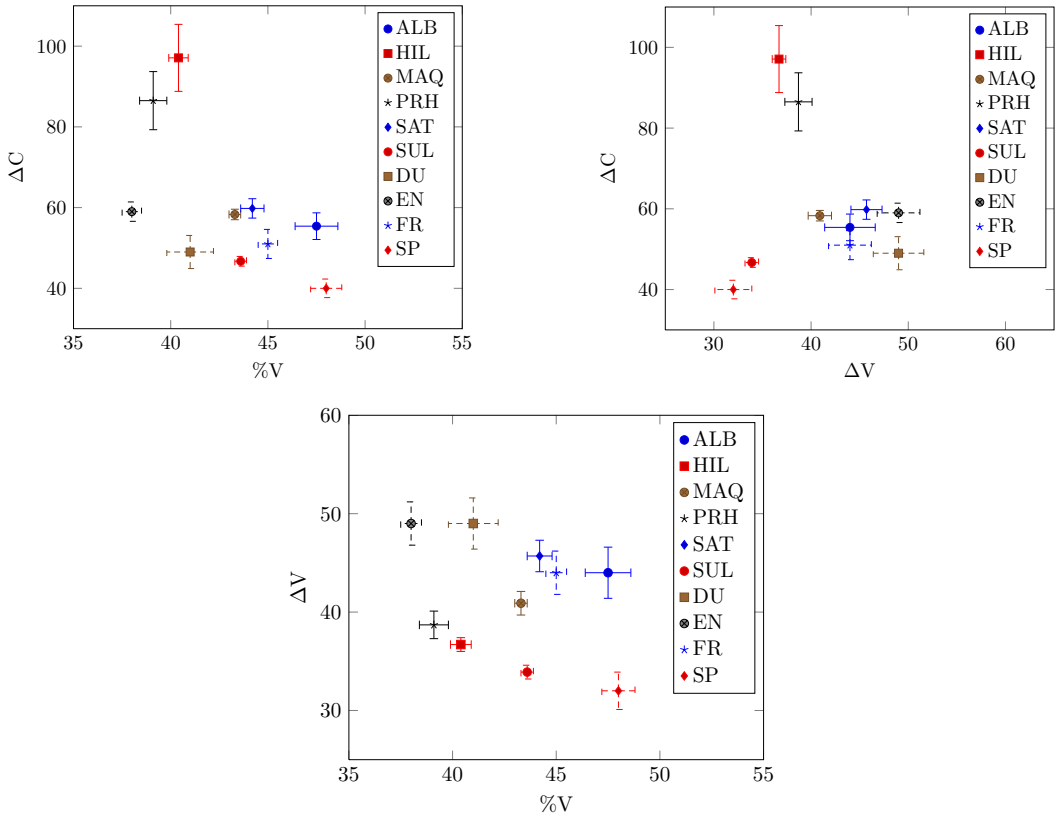


FIGURE 6.3: Distribution of languages & Algerian dialects over the %V and ΔC plane (top left), ΔV and ΔC (top right), and %V and ΔV plane (bottom). Error bar represents ±1 Standard Deviation. (DU= Dutch, EN= English, FR= French, SP= Spanish)

Figure 6.3 demonstrates the classification of the Arabic Algerian dialects with Dutch, English, French, and Spanish languages by plotting the graph with the values of %V, ΔV, and ΔC. English, French, Spanish, and Dutch languages features are taken from White & Mattys (2007).

As mentioned in Section 6.1.1, the (%V, ΔC) plane is very popular in the rhythmical categorization of languages/dialects. In this study, we observe from the (%V, ΔC) plane, that Algerian Arabic dialects can be classified into two groups. The first dialect group has the highest ΔC and the lowest %V such as English and Dutch languages where those are classified as stress-timed. It includes Pre-Hilālī and Hilālī dialect. The second group represents the syllable-timed rhythm class, such as Spanish and French language. It has lower values of ΔC but higher values of %V, which includes Ma'qilian, Algiers-Blanks, Sulaymite, and Sahel-Tell dialect. The other two graphs illustrate more scattered patterns and, thus, do not show language groupings as clearly as the %V-ΔC graph does. For example, Sulaymite dialect is not grouped with other dialects.

Figure 6.4 shows the Algerian dialects and four languages plotted on (rPVI-C, nPVI-V) plane. English, French, Spanish, and Dutch languages values of features (rPVI-C and nPVI-V) are taken from White & Mattys (2007). All dialects are clustered only on one group close to syllable-timed rhythm class, such as French and Spanish languages, which is characterized by having lower nPVI-V scores than for syllabic-timed (Duch, English) rhythm class.

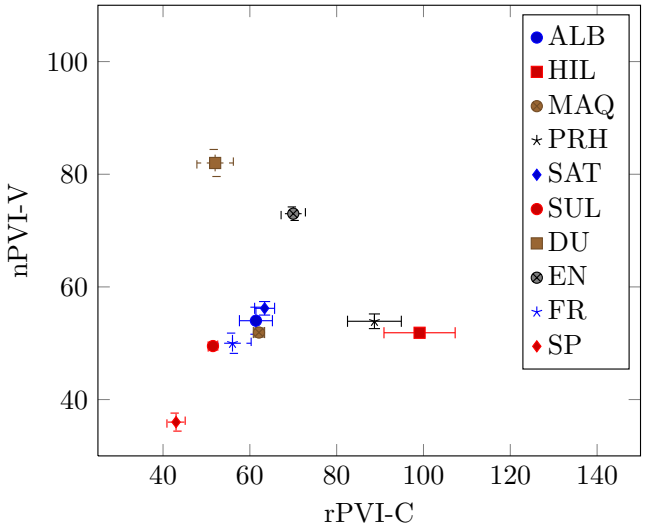


FIGURE 6.4: Distribution of Languages & Dialects along the rPVI-C (x axis) and nPVI-V Dimensions (y axis). Error bar represents ±1 Standard Deviation. (DU= Dutch, EN= English, FR= French, SP= Spanish)

In order to find whether the rhythm features and speech rate were significantly different between Algerian dialects, ANOVA tests were performed followed by the post-hoc (Tukey HSD) test.

In fact, the results of one-way ANOVA ($\alpha = 0.05$) that test for differences in rhythm metrics for six Algerian dialects, reported in Table 6.4, mention that the seven rhythm metrics: the three interval measure metrics (%V, ΔV, and ΔC), both time-normalized metric (VarcosC and VarcosV), and both pairwise variability indices (rPVI-C and nPVI-V), are sensitive to speaker' dialect.

Statistical analysis, one-way ANOVA, shows significant effects of rhythm features and speech rate between these Algerian dialects (all p-value $< \alpha$). In one-way ANOVA test, a significant p-value indicates that some of the dialect means are different. To determine which dialects are different, we rely on Tukey test for performing multiple pairwise-comparison.

The summary from the plots for Tukey test on all vocalic interval features of Arabic Algerian dialects is reported in Table 6.5. We observe that the pair dialect (SAT-ALB) has not

Feature	% V	ΔV	VarcoV	nPVI-V	ΔC	VarcoC	rPVI-C	Speech Rate
HIL-ALB	✓							
MAQ-ALB	✓							
PRH-ALB	✓							
SAT-ALB								
SUL-ALB	✓	✓						✓
MAQ-HIL	✓	✓			✓		✓	✓
PRH-HIL						✓		✓
SAT-HIL	✓	✓	✓	✓	✓		✓	
SUL-HIL	✓				✓	✓	✓	
PRH-MAQ	✓					✓		
SAT-MAQ				✓				
SUL-MAQ		✓				✓		✓
SAT-PRH	✓	✓				✓		
SUL-PRH	✓			✓	✓	✓	✓	✓
SUL-SAT		✓	✓	✓		✓		✓

TABLE 6.5: Summary of Significant Differences in the Rhythm Features between Algerian Dialect Pairs.

significant difference in all rhythm metrics and speech rate. While, the both pairs (SAT-HIL) and (SUL-PRH) have significant difference in six metric. In addition, the feature that is discriminate between the most pairs (ten from fifteen pairs) is the proportion of vocalic interval (%V). For that, it can be used for binary classification. While, the less discriminative feature (only two pairs) is VarcoV feature.

To analyze cross-language differences according to pitch variation, the following distributional measures were calculated per utterance: four global pitch information (Bottom, Top, Median, Range) and both total size of pitch trajectory (TrajIntra, TrajInter). Table 6.6 reports the mean and standard error (in parentheses) of these intonation features and the results of one-way ANOVA testing effect. The main observation is that all p-value are smaller than .005 for each ANOVA test, which means that all the intonation features have signification effect of speakers' dialect.

In addition, Algiers-blanks dialect exhibits the smallest variability in three pitch features, which are *Bottom*, *Mean*, and *TrajIntra* of pitch information. In contrary, the highest variability in both pitch global information features (*Bottom* and *Mean*) for Sahel-Tell dialect and in pitch *TrajIntra* for Hilālī dialect. Pre-Hilālī dialect has the highest variability in pitch *Top* and in *Mean* feature. In addition, Ma'qilian is characterized by the highest variability

Feature	ALB	HIL	MAQ	PRH	SAT	SUL	F	p-value
Pitch Range	12.7 (4.9)	9.0 (4.8)	**14.7** (5.8)	12.8 (5.0)	9.6 (5.52)	**8.0** (3.9)	F(5, 1886) = 99.3	p< .000
Pitch Top	213.0 (55.3)	267.7 (79.4)	315.8 (76.5)	**330.3** (109.7)	299.5 (50.0)	**259.0** (67.2)	F(5, 1886) = 50.1	p< .000
Pitch Bottom	**100.4** (14.8)	159.8 (46.7)	135.8 (37.1)	156.3 (54.1)	**175.1** (40.2)	163.7 (46.0)	F(5, 1886) = 41.6	p< .000
Pitch Mean	**148.2** (19.3)	206.7 (52.2)	208.9 (40.8)	**231.9** (74.2)	**231.4** (21.0)	211.4 (49.7)	F(5, 1886) = 29.8	p< .000
Pitch TrajIntra	**6.9** (6.7)	**9.9** (6.2)	7.4 (2.7)	8.2 (5.1)	8.8 (4.7)	**9.9** (4.8)	F(5, 1886) = 18.5	p< .000
Pitch TrajInter	14.1 (38.5)	9.9 (5.8)	**9.8** (3.5)	9.9 (4.3)	10.2 (4.4)	**18.2** (4.4)	F(5, 1886) = 8.1	p< .000

TABLE 6.6: Mean (Standard Error) Values of Intonation Features for Arabic Algerian Dialects and the Results of One-way ANOVA Testing Effect of Speakers' Dialect.

in pitch *Range* and by the smallest variability in pitch $TrajInter$, contrary to Sulaymite dialect. Furthermore, Sulaymite dialect exhibits the highest variability in pitch $TrajIntra$ and by the smallest variability in pitch *Top*.

In summary, it is clear that for all prosodic features has a significant different of dialect by using ANOVA test. To determine which dialects are significant, we have computed the Tukey test.

According to Tukey test of the significant differences in the intonation features between Algerian dialects, reported in Table 6.7. We observe that the pair dialect (PRH-MAQ) has the less significant difference. While, the both pairs (HIL-ALB) and (SUL-MAQ) have significant difference in six intonation features. In addition, the feature that is discriminate between the most pairs (twelve from fifteen pairs) is all global information (Bottom, Top, Median, and Range). For that, it can be used for binary classification. While, the total size features of pitch trajectory are discriminative features for more than half dialect pairs.

As a global summary, it is clear that for almost all prosodic features are a significant differences between dialect by using ANOVA test. As results, the global pitch information (Range, Top, Bottom, and Mean) are the highest features (twelve from twelve) that have discriminative power effect of dialect pairs compared to other prosodic information.

Feature	Range	Top	Bottom	Mean	TrajIntra	TrajInter
HIL-ALB	✓	✓	✓	✓	✓	✓
MAQ-ALB		✓	✓	✓		✓
PRH-ALB		✓	✓	✓		✓
SAT-ALB	✓	✓	✓	✓		✓
SUL-ALB	✓	✓	✓	✓	✓	
MAQ-HIL	✓	✓	✓	✓	✓	
PRH-HIL	✓	✓			✓	
SAT-HIL		✓	✓	✓		
SUL-HIL	✓			✓		✓
PRH-MAQ	✓		✓			
SAT-MAQ	✓		✓	✓	✓	
SUL-MAQ	✓	✓	✓	✓	✓	✓
SAT-PRH	✓	✓	✓			
SUL-PRH	✓	✓		✓	✓	✓
SUL-SAT	✓	✓	✓	✓		✓

TABLE 6.7: Summary of the Significant Differences in the Intonation Features between Algerian Dialect Pairs.

6.5 Conclusion

In this chapter, we presented the targeted prosodic information (rhythm and intonation) for dialects. First, we presented the different rhythm measures, which are developed in literature to successful language classification into tree groups: stress-timed, syllable-timed, and mora-timed. Second, we described both groups of intonation metrics. Then, we explained how we have segment and extract these prosodic features and then the used Open Source softwares. Finally, we analyzed statistically Algerian dialects in order to capture their specificities related to prosody information using two statistical analysis tests: one way ANOVA and Tukey tests.

Chapter 7

Arabic Algerian Dialect Identification

Contents

In this chapter, we present our third contribution, which is the development of Arabic Spoken Dialect IDentification (ASDID) systems on intra-country context, specially for Algerian dialects. We start by describing the first system: a flat classification-based approach. Then, we present our second system that relies on hierarchical classification. For both systems, we leverage prosodic information.

The majority of Arabic SDID work, presented in previous Section 5.5, are mostly based on acoustic/phonetic, phonotactic, or lexical information. Moreover, they need a massive amount of data for experimentation. Algerian dialects are under-resourced language, it is necessary to think outside the box. For that, we will exploit the prosodic information to identify Algerian dialects. As results from the statistic analysis, as previously described in the previous chapter, the prosodic information can be used for *binary classification* between Algerian dialects.

In addition, most researchers investigating ASDID have deployed the *shallow* or traditional machine learning methods, such as: Support Vector Machine (SVM), Hidden Markov Model (HMM), or Gaussian Mixture Model (GMM). Deep learning methods have already made impressive advances in fields, such as computer vision and pattern recognition. Following this trend, recent NLP research is increasingly focusing on the use of deep learning methods. For example, the percentage of deep learning papers in the Association for Computational Linguistics[1] (ACL) conference is more than 70% of accepting papers (Young et al. 2017). This is yet to happen for Arabic NLP (Al-Ayyoub et al. 2018).

Before presenting our proposed systems, let us introduce and recall the deep learning classification method.

7.1 Deep Learning

Deep Learning (DL) is also sometimes called feature or representation learning. Du & Shanker (2009) mentioned that DL has emerged as a new area of machine learning research since 2006s. DL tries to mimic the human brain, Figure 7.1 reports biological neural network, by constructing an architecture that consists of input layer and an output layer with many hidden layers between them (Al-Ayyoub et al. 2018).

When compared DL with traditional approaches, or *shallow* machine learning such as SVM, two point are required to be clarified. First, traditional approaches require samples with

[1]Is the international scientific and professional society for NLP community.

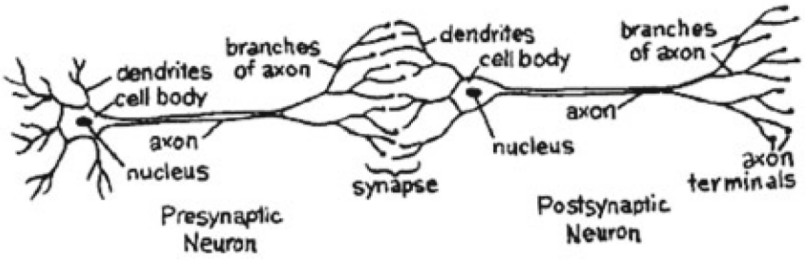


FIGURE 7.1: Biological Neural Network (Aggarwal 2018).

well-studied discriminative features extracted by experts (Pasupa & Sunhem 2016). Also, a single layer feed-forward network is a shallow learning. In contrast, deep learning is known as representation learning, which has been shown to perform better at extracting global relationships in the data (Di et al. 2018). Figure 7.2 illustrates the comparison of shallow and deep machine learning architectures. Second, the learning capacity of deep learning algorithms is proportional to the size of data, the performance of deep learning models just keep on improving with additional data, whereas, for traditional learning algorithms, the performance reaches a plateau after a certain amount of data is provided (Di et al. 2018, Goyal et al. 2018). Figure 7.3 gives a better representation of this comparison.

There are different types of Deep Neural Networks (DNNs). It can be classified according on

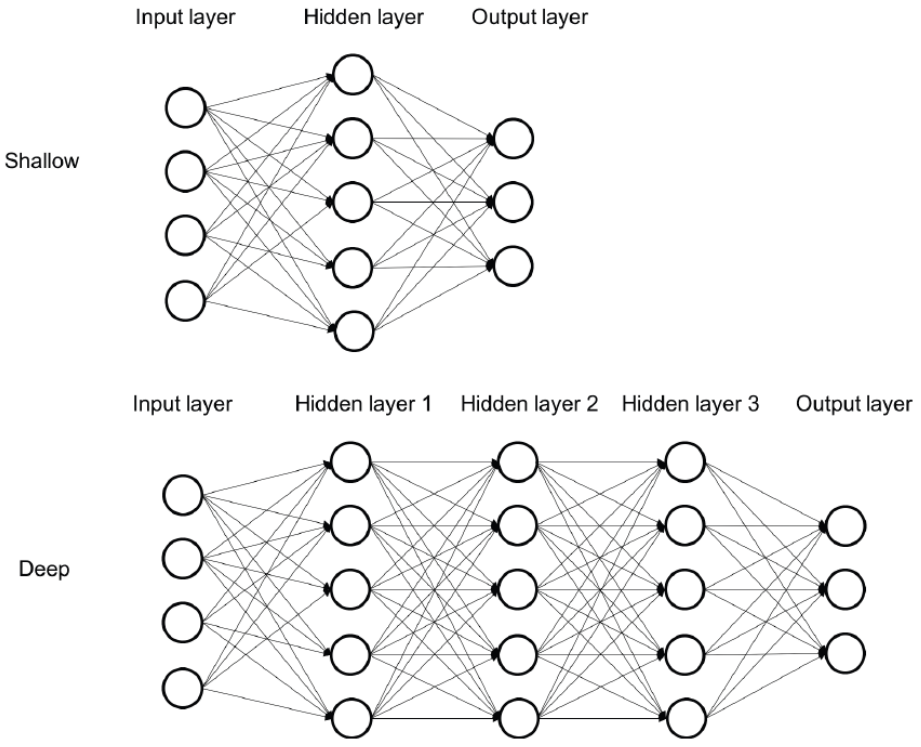


FIGURE 7.2: Comparing Deep and Shallow Architecture (Di et al. 2018).

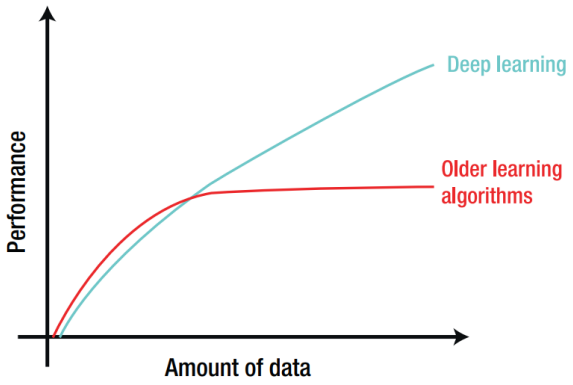


FIGURE 7.3: Scaling Data Science Techniques to Amount of Data (Goyal et al. 2018).

architecture and usage. If the neurons are taking the form of cycles, they form networks such as recursive neural networks. Else, they form networks such as feed-forward and convolutional neural networks (Goyal et al. 2018).

A. Feed-Forward Neural Networks

Feed-Forward Neural Networks (FFNN) —also called MultiLayer Perceptron (MLP)— is an artificial neural network that has more than one hidden layer[2] between its inputs and outputs. The name feedforward comes from the fact that layers *feed* into one another in the *forward* direction from input to output (Aggarwal 2018). Figure 7.4 reports an example of Feed-Forward neural network with three hidden layers.

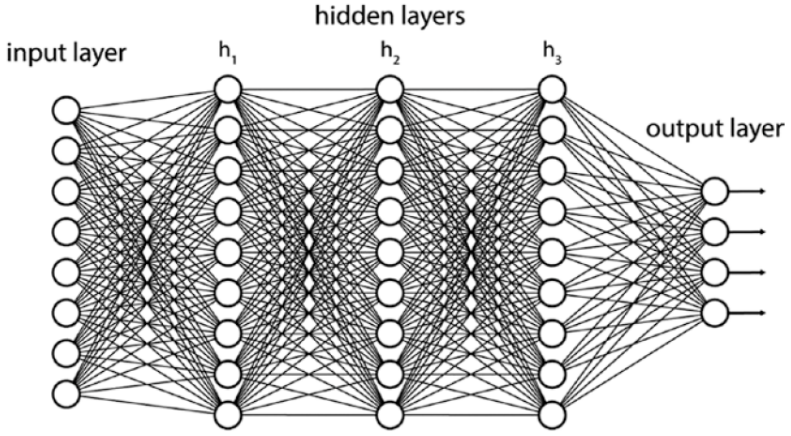


FIGURE 7.4: Feed-Forward Neural Network (Beysolow 2018).

[2] *Hidden Layer*: Because the computations performed are not visible to the user.

B. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are well adapted for image recognition and handwriting recognition. Their architecture is based on sampling a window or portion of the input data (image, speech), detecting its features, and then using the features to build a representation (Goyal et al. 2018).

Figure 7.5 illustrates a sample of CNNs architecture. CNNs is widely used in image problems, such as: object, image, handwriting recognition, and other computer vision problems. It also achieves state-of-the-art results in NLP problems, such as speech recognition (Manaswi 2018).

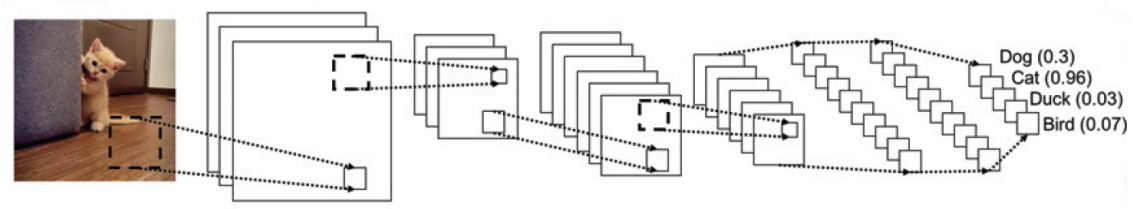


FIGURE 7.5: Sample of CNNs Architecture (Di et al. 2018).

C. Recursive Neural Networks

Recursive Neural Network (RNN) architecture a deep feed-forward neural network with addition of loops to the architecture. It is best suited to recognizing patterns in sequences of data, particularly time series and text. This architecture can be used for the detection of the sentiment of a sentence, in which the results is not dependent on the individual works only, but also on the order in which the words are syntactically grouped in the sentence (Goyal et al. 2018).

Figure 7.6 reports the structure of an RNN and its time-layered representation. This kind of architecture ensures that the output at time "t" (o_t) depends on the inputs at time "t" (x_t), "t-1" (x_{t-1}), ..., and "1" (x_1). In other words, the output depends on the sequence of data rather than a single piece of data (Manaswi 2018).

In most recent studies in Language IDentification (LID), Deep Neural Networks (DNNs) has been deployed in different architectures. We chosen for our system the Feed-Forward

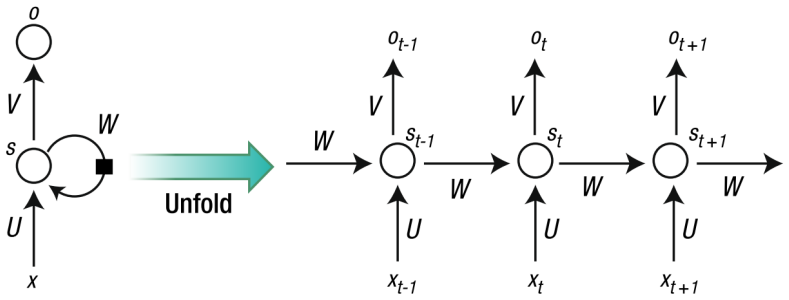


FIGURE 7.6: The Structure of an Recurrent Neural Network (Beysolow 2018).

Neural Network (FFNN) architecture, because the prosodic features are not images and are not correlated in time.

FFNN architecture is considered as the state-of-the-art in LID (Koolagudi et al. 2017, Lopez-Moreno et al. 2016). Lopez-Moreno et al. (2016) extract the acoustic information to identify 24 languages. They showed that DNNs system improves their superiority compared to i-vector system. For Dravidian language, Koolagudi et al. (2017) exploit the DL method by modeling the cepstral coefficients and prosodic features. As results, they mentioned that DL are considered to classify the four Dravidian languages.

Let us recall that Feed-Forward Neural Network (FFNN) belongs to the category of feed-forward neural networks and are made up of three types of layers: an input[3] layer, one or more hidden layers, and a final output [4] layer (Goyal et al. 2018).

Figure 7.7 shows a full-connected FFNN with only one hidden layer, which means every neuron of one layer is connected to all neurons of the next layer. Every connection between neuron j in layer k and neuron m in layer $k+1$ has a weight denoted by w_{jm}. Each neuron really computes two functions within the node: pre-activation (z) and post-activation (σ) function.

The pre-activation function of $\mathtt{z}_{ij}$ at neuron `i` from layer `j` is represented by Equation (7.1). For example, the zoomed neuron (neuron 3 from layer 2) reports the pre-activation function of $\mathtt{z}_{23}$.

$$\mathtt{z}_{ij} = b_j + \sum_i w_{ij} \ y_i \tag{7.1}$$

where i is an index over the units of the layer below and b_j is the bias of the unit j.

[3]The number of neurons present across the input layer is equal to the number of features.

[4]The number of neurons present across the output layer is usually equal to the number of classes we want to classify the target value.

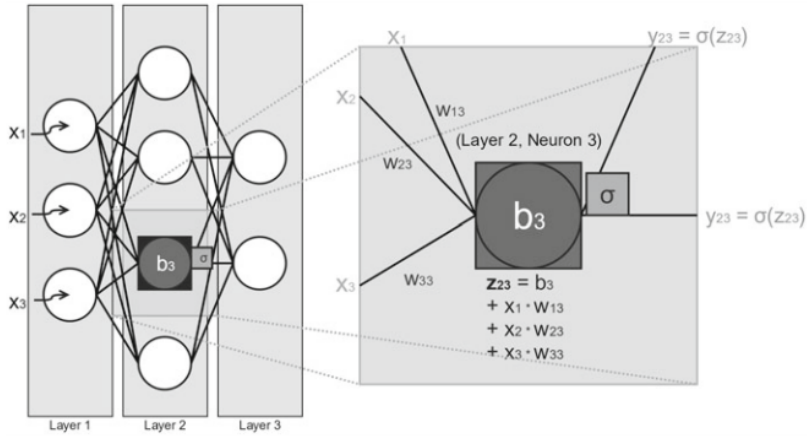


FIGURE 7.7: A Simple Feed-Forward Neural Network (Skansi 2018).

The result of this pre-activation function is mapped to its corresponding post-activation function $\mathtt{y}_{ij}$ represented by Equation (7.2). The activation function in each neuron decides whether the incoming signals have reached the threshold and should output signals for the next level (Di et al. 2018).

$$\mathtt{y}_{ij} = \sigma(z_{ij}) \tag{7.2}$$

Equation 7.3 reports a few common activation functions, such as: Sigmoid, Tanh, and Rectified Linear Units (ReLU).

$$Sigmoid(z_{ij}) = \frac{1}{1+e^{-z_{ij}}},\ Tanh(z_{ij}) = \frac{e^{2z_{ij}}-1}{e^{2z_{ij}}+1},\ ReLU(z_{ij}) = \max(0, z_{ij}) \tag{7.3}$$

For classification problems, the post-activation function of the output layer is a softmax output function, which converts the outputs of the final hidden layer $\mathtt{y}_{nj}$ into a class probability $\mathtt{p}_j$ of the following form:

$$\mathtt{p_j} = \frac{\exp(\mathtt{y_{nj}})}{\sum_{\mathtt{d=1}}^{\mathtt{k}} \exp(\mathtt{y_{nd}})}, \quad \forall j \in \{1, ..., k\} \tag{7.4}$$

where $\mathtt{k}$ is an index over all of the target classes and $\mathtt{n}$ is number of the output layer.

In order to measure how far the output of softmax function to the true target class, the loss function of the network, also called the objective function, is used to compute a distance score, capturing how well the network has done on this specific parameters. For classification,

as cost function for back-propagating gradients in the training stage, cross entropy can be the useful loss function, defined as:

$$C = -\sum_{j} t_j \log(p_j) \tag{7.5}$$

where t_j represents the target probability of the class j for the current evaluated example, it takes a value of either 1 (true class) or 0 (false class).

The fundamental trick in deep learning is to use the loss score as a feedback signal to adjust the value of the weights a little, which the optimization step, which is the key to how a network learns (Chollet 2017). The most common optimization approach to minimize the *Loss score* is still *Stochastic Gradient Descent (SGD)* and its variations, for example, *momentum-based methods*, *AdaGrad*, *Adam*, and *RMSProp* (Di et al. 2018).

Figure 7.8 reports the graph that gives the relationship between the network, layers, loss function, and optimizer.

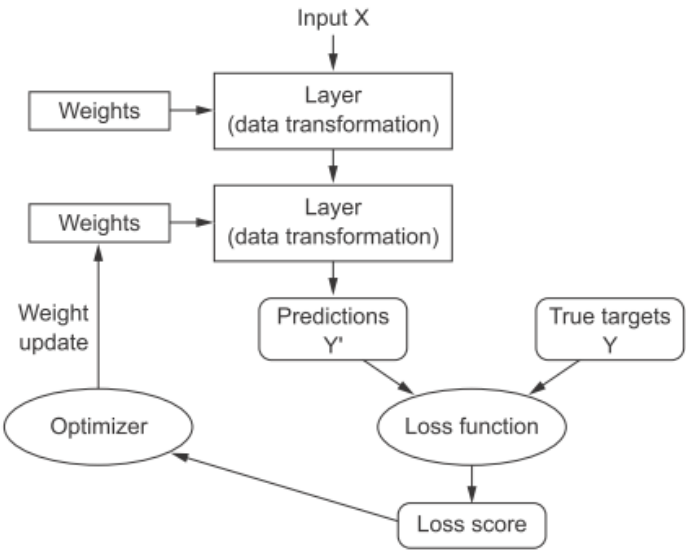


FIGURE 7.8: Relationship between the network, layers, loss function, and optimizer (Chollet 2017).

7.2 Flat Dialect Identification Approach

Our first proposed Algerian Arabic Spoken Dialect Identification (AADID) is baplised the *flat* approach, which is a machine learning based method that leverage prosodic information and deep learning technique.

7.2.1 Approach Overview

Figure 7.9 gives an overview of this approach. It processes as follows. Given a dataset of speech from different dialects, first a preprocessing step is performed where we remove almost noise and silence from all speeches of the dataset. After that a *coarse-grained* segmentation is performed, as shown in Section 6.2.

Then, *prosodic information* are extracted. Let us recall, we have used two prosodic information: intonation and rhythm, which are measured by using the pitch and duration sequence respectively. We have selected seven rhythm metrics: $\%V$, $VarcoV$, $nPVI\text{-}V$, ΔC, $VarcoC$, $rPVI\text{-}C$, and *Speech Rate*. In addition to that, six intonation features: four global information and two pitch trajectory features. We have already detailed these features in Table 6.1 and 6.2.

Our Prosodic Features (Pf) are extracted at utterance level. Whereas, previous prosodic-based ADID researches leveraging prosody extracted prosodic features at pseudo-syllable level (Biadsy & Hirschberg 2009, Rouas et al. 2006).

After that, the *dialect models* are generated using Deep Neural Networks (DNNs), specially Feed-Forward Neural Network (FFNN) architecture, which is considered as the state-of-the-art in language identification (Koolagudi et al. 2017, Lopez-Moreno et al. 2016). The resulting Algerian Arabic ADID system, we call AADID-DNN.

The input to the DNNs classifier consists of the prosodic features described above (14 Pf) and the output layer size is 4 (number of dialects). Figure 7.10 reported details on DNNs topology and description. For complete this architecture, we need to fix the: i) Number of layers in a network, ii) Number of neurons in the hidden layer, iii) Neurons Activation function. iv) Optimizer to updating the model parameters. These parameters are tuned empirically. More details on that are given in the experiment section.

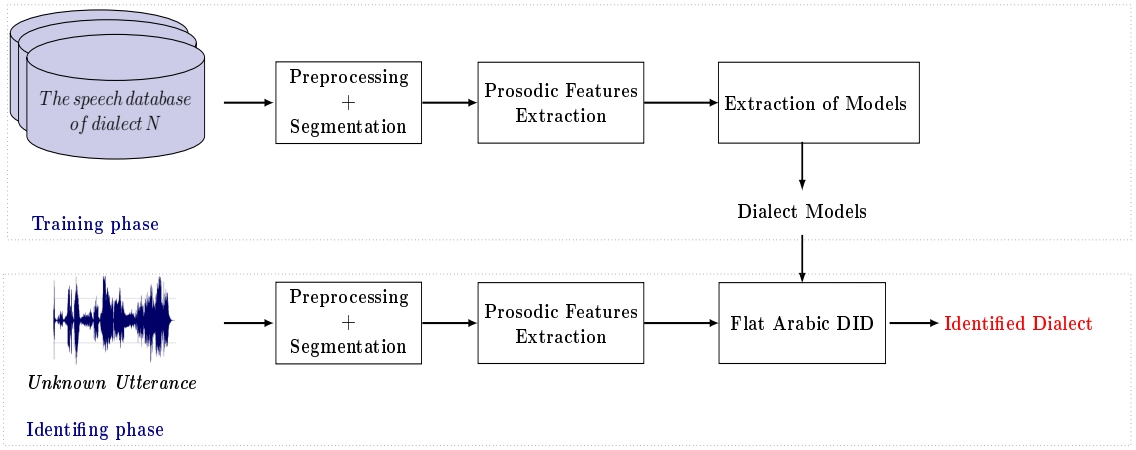
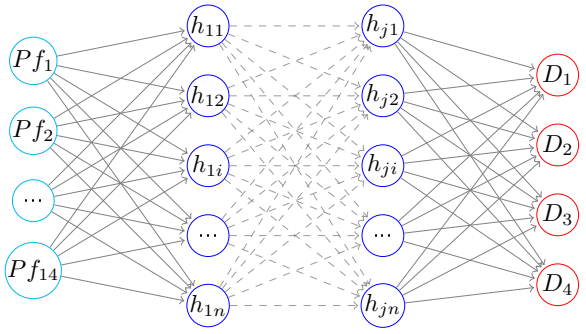


FIGURE 7.9: Overview of the Flat System.



- **Topology :**
  - 14 Pf : # of input Prosodic features.
  - j : # *of hidden layers.*
  - n : # *of neurons in each hidden layer.*
  - *Neurons Activation Function*: ?
  - Softmax output layer.
- **Training :**
  - *Optimizer*: ?
  - A dropout factor of 0.5 in each hidden layer.

FIGURE 7.10: DNNs Topology and Description.

7.2.2 Experiments

In order to evaluate the performances of our AADID-DNN approach, we have performed a batch a number of experiments:

1. The evaluation of our proposed AADID-DNN performances using a number of features and settings.
2. The comparison between DNN and SVM classifiers. For that, we designed a prosody-based using SVM modeling (AADID-SVM).
3. The comparison between the prosodic and acoustic speech information, which is the most used information to identify ADID. Our baseline acoustic-based approach is based on Djellab et al. (2017) that identify Algerian areas.
4. The sensitivity measurement of the utterance size variation for prosody-based and acoustic-based systems systems.

Before presenting our experiments and results, let us first present the dataset, the DNNs tuning, the used tools, and the baseline system.

7.2.2.1 Data

As mentioned in the literature review of Arabic speech corpus, as presented in Section 2.5, these studies indicate that more attention has been focused on some Arabic dialects, such as Egyptian and Levantine, compared to Algerian dialects. For that, to evaluate our AADID-DNN approach, we have selected the data from our KALAM'DZ corpus.

As described on the previous pages, KALAM'DZ is a large speech corpus dedicated to Algerian Arabic dialectal varieties. It covers eight Arabic dialects spoken in Algeria. This corpus is collected from web sources namely YouTube, Online Radio stations, and TV channels. The size of the corpus is about 104 hours with 4881 speakers. This corpus provides a representation of phonetic, prosodic and orthographic varieties of Algerian Arabic dialects. It represents a good sample for within-language variability, among these aspects: scope and the size of the corpus, the richness of speech topics and content, number of speakers, gender, regional dialects, recording environment, and content (Li et al. 2013) .

A sample of 42 hours was chosen from the aforementioned collection. The selection was based on the most commonly spoken dialect, coverage for all dialect varieties as well and the maximum number of utterances. The dialects include: Hilālī, Sulaymite, Ma'qilian, and Algiers-Blanks.

Table 7.1 provides further details on the selected sample for the current study. The total number of utterances is 17036, with an average duration of 8.1 s.

Dialect	# of Utterances	Average time (s)	Total Time
Algiers-Blanks (ALG)	4845	6.7	09h 06min
Ma'qilian (MAQ)	3470	8.9	12h 03min
Hilālī-Saharan (SAH)	4560	8.7	10h 57min
Sulaymite (SUL)	4152	8.4	10h 37min
Total	17036	8.1	42h 43min

TABLE 7.1: Speech Data Composition.

7.2.2.2 DNNs Implementation

The problem of DNNs architectures optimization is to find the optimal number of hidden layers in the network, the number of neurons within each layer, the good activation function and optimizer, in order to maximize the performance of DNNs. For that, we have used is Genetic Algorithms (GA) which is commonly used to generate high-quality solutions to solve optimization problem (Beasley et al. 1993).

GA is an optimization meta method that applies three operators similar to natural genetic operators: reproduction, crossover, and mutation, in order to archive the best solutions (Sheppard 2018).

To find the optimal topology, the GA has been successfully applied in several domains for the purpose of solving problems with various degrees of complexity. Chiroma et al. (2017) give a summary of DNNs topology optimization through a GA search.

The explored parameters of the problem must be defined for the GA, Table 7.2 reports the range of DNNs hyper-parameters, such as: activation functions, number of of hidden layers and neurons per layer, optimizer, and the GA parameters, such as: the number of population in each generation (PopSize), the maximum number of generations (MaxGeneration), and fitness function (Cost-Fcn).

GA Hyper-parameters	Explored Values
Activation Function	ReLU, Sigmoid, Tanh
Hidden Layers	2, 3, 4, 5, 6, 7, 8, 9, 10
# of Neurons	64, 128, 256, 512, 768, 1024, 2000, 2560, 3000, 3500, 4000, 4500, 5000
Optimizer	rmsprop, adam, SGD, nadam, adagrad, adadelta, adamax
PopSize	20
MaxGeneration	10
Cost-Fcn	Accuracy

TABLE 7.2: Explored Hyper-parameters for Genetic Algorithm Modeling.

We have modified the Python codes[5] implement by Matt Harvey that evolve a neural network with a genetic algorithm, more details are found in this blog post[6]. The steps of genetic algorithm to evolve the network are as follows:

[5] https://github.com/harvitronix/neural-network-genetic-algorithm.
[6] https://blog.coast.ai/lets-evolve-a-neural-network-with-a-genetic-algorithm-code-included-8809

1. Initialize 10 random networks to create our population.
2. Score each network by using a fitness function, which is in our case the accuracy for classification task.
3. Sort all the networks in our population by score of fitness function (accuracy).
4. Selects and breeds the best members of the population to produce more like them: Keep 40% percentage of the top networks to become part of the next generation and to breed children. Also, randomly keep a 10% of percentage of remaining (non-top networks) in the population, which helps to find potentially lucky combinations between worse-performers and top performers, and also helps keep us from getting stuck in a local maximum.
5. Mutates randomly some of the parameters on some of the networks (20 %) to attempt to find even better candidates.
6. Kills off the rest, to keep our population at 10 networks.
7. Repeats from step 2 until 20 generation. Each iteration through these steps is called a generation.

As results of the genetic algorithm execution which predicted the best hyper-parameters values, Table 7.3 illustrates the optimal parameters that we have archived to complete the DNNs architecture.

Hidden Layers	# of Neurons	Activation Function	Optimizer
4	64	Sigmoid	SGD

TABLE 7.3: Hyper-parameters Values Predicted for Deep Neural Networks.

7.2.2.3 Acoustic-based Baseline System

In order to contrast our AADID-DNN approach, we have chosen the most used speech information, to identify ADID in literature, which is the acoustic one. There is a panoply of acoustic-based approaches, our baseline system is based on Djellab et al. (2017) because they identify Algerian area dialects. They extracted the basic acoustic feature, Mel-Frequency Cepstral Coefficients (MFCCs), with the commonly utilized additional features of the temporal aspects, which is Shifted Delta Cepstral (SDC) coefficients. This approach has carried out using Gaussian Mixture Models-Uniform Background Model (GMM-UBM) modelings.

To extract MFCCs acoustic features, a given input utterance gets segmented into windows of 25ms with 10ms overlap. Then, Mel-scale filter bank analysis is applied to the signal, respectively before converting the obtained log-energy filter bank trajectories to Cepstral coefficients. The first 7 Mel-frequency cepstral coefficients, including $C0$, are computed on each frame (for more details see Appendix C). These MFCCs are subject to some normalization (vocal tract length and cepstral mean) in addition to other filter (RASTA).

After the normalization of MFCCs features, additional features, which are SDC coefficients, are added, the temporal aspects of the speech signal, to each feature vector. These coefficients are extracted, using the conventional 7-1-3-7 scheme, and appended to the MFCC features to produce a 56-dimensional feature vector.

Then, a dialect-independent Universal Background Model (UBM) is constructed from the training data from each of the dialects. MAP adaptation of means and weights is then performed on the UBM to generate dialect dependent Gaussian Mixture Models (GMMs), one for each dialect. These GMMs are used to evaluate the likelihood of test utterances as belonging to a particular dialect. In our case, we have deployed UBM-GMM with 1024 mixtures. For more details on these features extraction, normalization and reduction refer to Djellab et al. (2017).

7.2.2.4 Used Tools

In addition to tools that extract prosodic information, we have involved a set of Open Source softwares.

- To tune the DNNs architectures and other hyper-parameters, we have adopted the Python code of Matt Harvey[7] which is developed to evolve a neural network with a genetic algorithm.
- Our DNNs learning models were implemented using Keras[8], an open source deep learning library written in Python.
- For implementing the acoustic system, we use SIDEKIT[9], a publicly available Speaker and Language Recognition toolkit for acoustic feature extraction, and the UBM-GMM modeling is done using Spear[10] (Khoury et al. 2014), an open-source speaker recognition toolbox based on Bob tools (Anjos et al. 2012).

[7] https://github.com/harvitronix/neural-network-genetic-algorithm
[8] https://keras.io/
[9] http://www-lium.univ-lemans.fr/sidekit/
[10] https://pypi.python.org/pypi/bob.bio.spear/3.1.1

7.2.2.5 Results and Discussion

For all experiments, we use speaker-independent DID. This means that the speaker set used for training is disjoint from speaker set used for testing which leads to real dialect characterization capture to avoid potential speaker characterization capture. The dataset is split into training (3/4) and testing (1/4) sets.

Our discussion of the results begins with the study of the performance of our AADID-DNN system. To evaluate this system, we use the k-fold cross validation technique. To be more confident in our system performance, we create five different models and test it on five different test sets, results are reported in Table 7.4.

From Table 7.4, we observe that AADID-DNN reaches an average precision of 41.6%. We highlight that prosody gives a precision acceptable compared to state-of-art of ADID system. In fact, Ali et al. (2016) worked on *inter-country* dialects identification, they reached, with lexical-based system, 45.8% and when they combined with the acoustic information they have reached 60%. While, Djellab et al. (2017), who works on intra-country context (only 3 regions dialect) with acoustic they reached 57%. For our case prosodic-based gives similar result if we take only 3 dialects.

System	ALG	SAH	MAQ	SUL	Average
Fold 1	26	37	18	100	47
Fold 2	34	33	24	97	41
Fold 3	36	32	00	99	34
Fold 4	37	32	26	98	42
Fold 5	36	31	32	100	44
	33.8	33	20	98.8	

TABLE 7.4: AADID-DNN System Performance on Different Cross-validation Folds.

The confusion matrix, got from the best AADID-DNN fold system, is shown in Figures 7.11. In general, the AADID-DNN system performs worst in case of MAQ and SUL dialects which are most often confused with ALG and SAH, let mention that SAH is part of Hilālī group, that is related to the greater prosodic similarity between these pair dialects, which is confirmed by our prosodic statistic analysis. In fact, we have shown that the dialects pairs: MAQ-ALG, MAQ-SAH, SUL-ALG, and SUL-SAH, have less discriminative features (less than eight from fourteen).

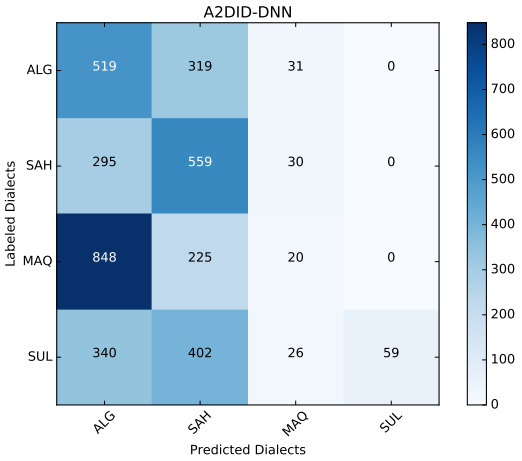


FIGURE 7.11: Confusion Matrix for the AADID-DNN System.

AADID-DNN vs. AADID-SVM Performances

In order to measure the power of DNN classifier, we have compared our approach with the robust SVM classifier (AADID-SVM). This later system uses similar pre-processing and prosodic features extraction to AADID-DNN, while it uses SVM classifier to build dialect models. To reach the optimal hyper-parameters for the SVM (hyperplane and penalty), the grid-search method is used, which is a meta-classifier in Weka tool (Braga et al. 2013).

We have selected the data of train and test from the Fold-1, which is the best performance of AADID-DNN system, to created and evaluated the AADID-SVM system. Table 7.5 gives the performances in term of precision and recall for both systems AADID-DNN and AADID-SVM per dialect.

System		ALG	SAH	MAQ	SUL	Average
AADID-DNN	Precision	26	37	18	100	47
	Recall	60	63	2	7	32
AADID-SVM	Precision	25.9	34.5	37.0	100	47.8
	Recall	40	21.9	66.1	7.0	33.5

TABLE 7.5: AADID-DNN vs. AADID-SVM.

We observe that both approaches have reached comparable performances with 47% and 47.8% of precision for AADID-DNN and AADID-SVM respectively. However, the performance of AADID-SVM system is more suitable as there is no precision less than 26% contrary to AADID-DNN system (MAQ is about 18 %).

In addition, SVM-based system recognizes well MAQ dialect. This is due certainly to the size of dataset as we shown that DNNs reaches an improvement with additional data compared

to traditional classifiers such as SVM classifier. As result, we mention that we need more data to implement the AADID-DNN system.

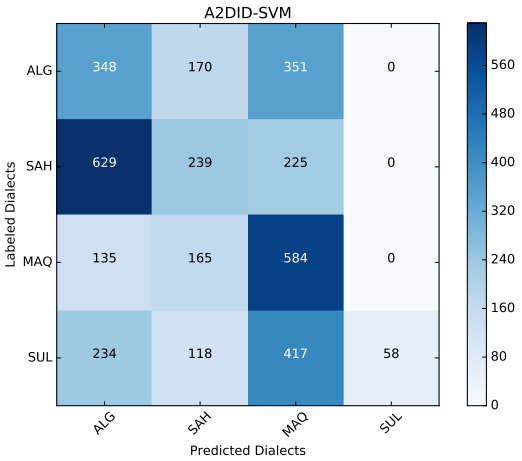


FIGURE 7.12: Confusion Matrix for the AADID-SVM System.

Looking at results, shown in the confusion matrix (Figure 7.12), it can be inferred that only SUL is the most challenging ones, which is confused with MAQ, contrary to AADID-DNN which is confused with SAH. This highlights difficulty to distinguish the SUL dialect using prosody. In addition, their recall is very low regardless of their precision for both systems.

Prosody vs. Acoustic Characterization

Turning now to the comparison between the prosodic and acoustic information, note that we use the AADID-SVM for prosodic-based system which is more suitable compared to AADID-DNN one.

Table 7.6 reports results of acoustic-based and AADID-SVM system in term of precision and recall. We observe that acoustic system reaches an average 57 % for both precision and recall. It is clear that acoustic-based system outperforms prosody-based one. The SUL is also the less identified dialect, which out up that this dialect share may acoustic and prosodic information to other dialects.

Figure 7.13 reports the confusion matrix of the acoustic-based system. It shows that the system recognizes relatively well the instances of the SUL and MAQ dialects. Contrary, the ALG dialect is less identified compared to the AADID-DNN system. This leads to conclude that for the case of dialect identification, the acoustic-based system has better discriminative power. However, a combination can be considered to exploit their complementary discriminative power.

System		ALG	SAH	MAQ	SUL	Average
Acoustic-Based	Precision	51	79	51	42	57
	Recall	43	76	57	45	57
Prosodic-BasedAADID-SVM	Precision	25.9	34.5	37	100	47.8
	Recall	40	21.9	66.1	7	33.5
Combination	Precision	52	51	41	85	59
	Recall	43	57	53	71	57

TABLE 7.6: Acoustic-based, Prosodic-based, and Combination Systems in Term of Precision and Recall.

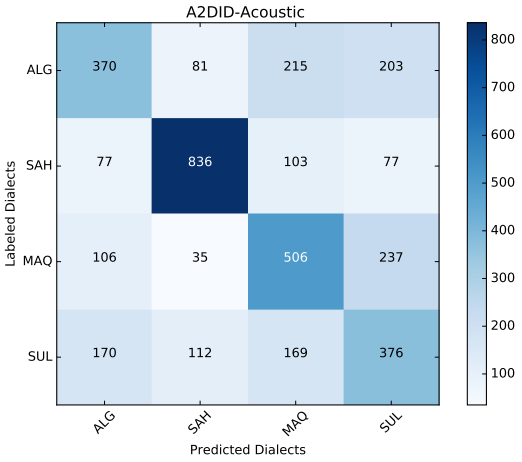


FIGURE 7.13: Confusion Matrix for the Acoustic-Based system.

As a supplementary investigation, we fused both AADID-SVM and acoustic-based systems. In the fusion step, we have applying the random forest classifier as top layer models by using the predictions of the both bottom layer models (AADID-SVM and acoustic-based). This approach reaches an average precision of 59% and recall of 57%, which is the best performance we achieved.

Comparing the performances of the three systems (see Table 7.6), it can be seen that the SUL dialect archives an average of 58% and 71 % for precision and recall respectively for combined system, which is better compared to the other systems. Surprisingly, the SAH and MAQ dialects have less performance than acoustic system, which means that the prosodic information has a negative effect.

Prosody vs. Acoustic: Test Utterance Size Sensitivity

For any DID system, the optimal performance would be its accuracy on shorter utterances. For that, we have measured the sensitivity of to the size of the utterance our prosodic-based

and acoustic systems.

Table 7.7 reports performance in term of precision of both AADID-DNN and Acoustic-based systems and their deviation from the average precision (measured above) according to test utterance sizes. Where *Short* column represents utterances with a size less than 3 seconds, *Medium* size represents utterances with a size between 5s and 10s and *Long* for utterances with a size that exceeds 10 seconds.

	Short (Len $< 3s$)		Medium ($3s \leq$ Len $< 10s$)		Long ($10s \leq$ Len)	
	Acoustic	Prosodic	Acoustic	Prosodic	Acoustic	Prosodic
Precision	42.5	57	51.3	44	60.9	06
Deviation	-14.5	$+6$	-5.7	-3	$+3.9$	-41

TABLE 7.7: AADID-DNN vs. Acoustic-based System According to Utterance Size Variation.

For Acoustic-based approach, it is obvious that the system needs longer utterances to get an accurate identification. For shorter utterances, the precision drops by -14.2. Contrary to the Prosody-based system, it behaves well for short utterances which is a requirement for real-live applications. The performance for *Short* utterances is $+6$ above the average precision. The drawback of such system it that is degrades rapidly with the length of the utterances (-41) which is quite large. Such features of both approaches lead to believe that a combination can be envisaged in which Prosody characterization improves the performances when acoustic one fail to identify short utterance.

7.3 Hierarchical Arabic Algerian Dialect Identification

In the flat approach, we haven't harness the hierarchy structure for Algerian dialects, mentioned in previous Section 1.5. For this fact, in the next section, we will present our Hierarchical Arabic Algerian Dialect Identification system.

7.3.1 Approach Overview

Many classification problems are cast as Hierarchical Classification (HC) problems, where the classes to be predicted are organized into a class hierarchy, where classes which are more

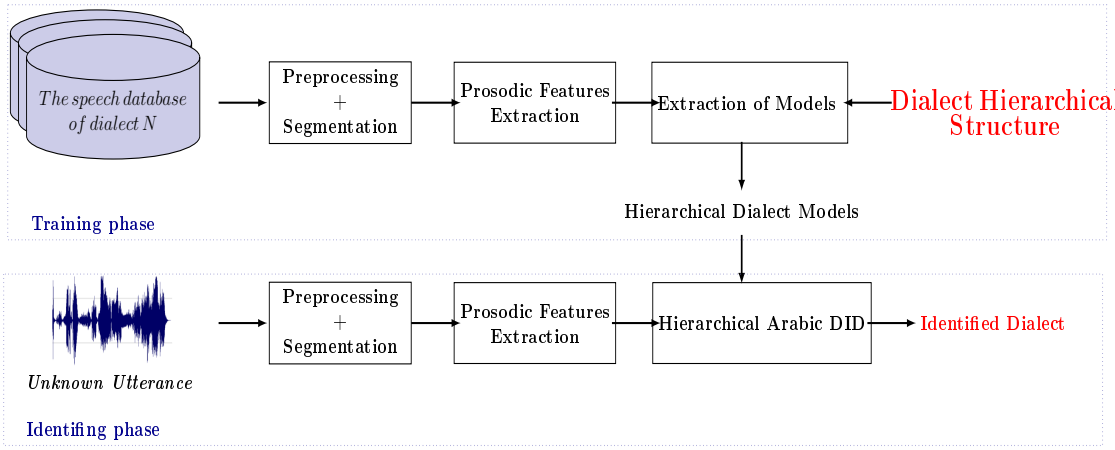


FIGURE 7.14: Overview of the HADID system.

similar to one another are grouped together. HC is a machine learning method that consists of placing new items into classes on the light of a predefined hierarchical structure of such classes Silla et al. (2011). It follows a divide to conquer technique, where the classification is divided into many sub-problems in a hierarchical way where classes that are more similar are grouped together in groups.

To our knowledge there is no deployment of HC in Arabic DID. However, Hierarchical LID (HLID) have been already proposed for others languages. For Indian languages, Jothilakshmi et al. (2012) designed a two level classification system using acoustic features. On the other hand, Yin et al. (2007) proposed a HLID framework where speech signal is characterized at acoustic level and some prosodic features. The fusion of these classifiers is performed using modern GMM fusion system. Likewise, Wang et al. (2009) suggested a hierarchical system using bayesian logistic regression models as score generators. The final identification is performed by a score based-likelihood merger. For Philippine languages, Laguna & Guevara (2015) developed a HLID system via GMM, in which speeches are characterized by means of acoustic and prosodic features.

Our proposed approach relies on prosodic features to identify Arabic dialect in intra-country context. In order to deal with the large dialectal varieties, our approach employs the dialect hierarchy structure to efficiently derive well training models. Figure 7.14 sketches the global scenario of our HADID approach. As inputs, we provide a database of speech from different dialects. Within the preprocessing phase, we remove noise and silence from all the database speeches. Afterward, the coarse-grained consonant/vowel segmentation is applied and then prosodic information are extracted. The latters processes are explained in Section 6.1. For the extraction of HADID models, we have used as an input our derived hierarchical structure for Algerian dialects which is presented in Figure 1.4.

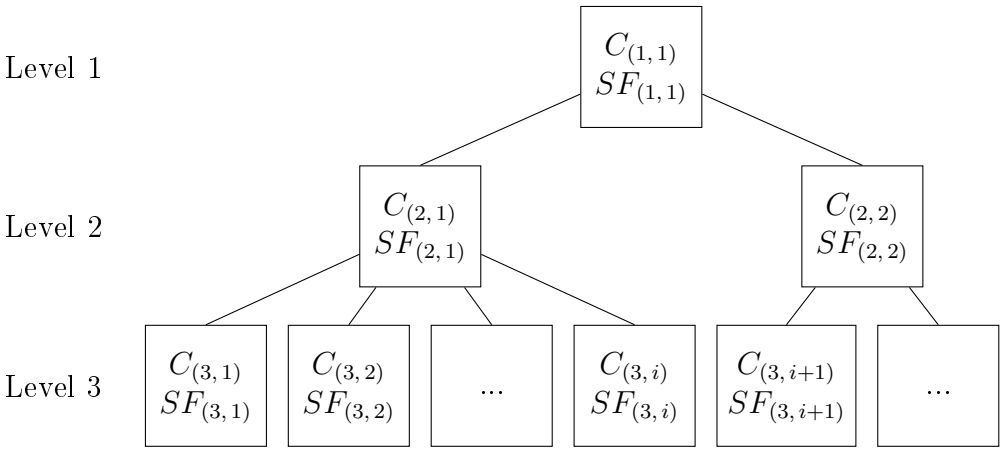


FIGURE 7.15: An Example of a Classifier per Parent Node Hierarchical Classification.

In order to build HADID models, many hierarchical classification methods can be taken into consideration. The most used are global and top-down approaches Silla et al. (2011). We opt to use the last one. It is also called local classifiers approach, where the hierarchy is taken into account using local information. More precisely, we built a set of dialect models, a Local Classifier per Parent Node (LCPN) from the top to down according to the dialect hierarchical structure. We consider for each model only discriminative features. Hence, a different Set of Features($SF_{(level,\#node)}$) and different Classifier ($C_{(level,\#node)}$) can be tuned separately for each node in order to reach the best performances. Figure 7.15 illustrates such hierarchical classification. This tree shape follows the hierarchical structure.

In other words, we choose among prosodic features those that have the best discriminative power for the embedded dialects by the node. Thus, each LCPN model dedicated to a node X is trained using the dataset that represents dialects for which X is an ancestor.

During the identification phase, an utterance is determined by passing through the different DID systems within the Hierarchical DID models. The target utterance is firstly classified to the most likely dialect group, proceeding level by level from the top until the final dialect becomes identified.

LCPN dialect models are generated using the state-of-the-art in language identification, Deep Learning technique (Ferrer et al. 2016, Lopez-Moreno et al. 2016), in this case Deep Neural Networks (DNNs).

7.3.2 Experiments

In order to evaluate the performances of our HADID approach, we have performed various experiments. First, we evaluate and analyze the performance of our HADID system. Secondly, we measure, also here, the effect of using DNNs for dialect modeling. In fact, we compare HADID performance to HADID system that uses SVM instead of DNNs modeling. Finally, we perform comparison of HADID with a baseline Flat classification approach, such done in the flat approach, which is built without considering hierarchical dialect structure. Before presenting our experiments and results, we present the dataset, used tools, and evaluating metrics.

7.3.2.1 Dataset and Tools

For the current analysis purpose, Table 7.8 provides details on the selected samples extracted from KALAM'DZ corpus. One can observe that we have dealt with a supplementary dialect (Pre-Hilālī). This firth dialect allow us to deal with the hierarchical structure of our studied dialects. We recall that we have performed speaker independent DID system. This means that the speaker set used for training is disjoint from speaker set used for testing which leads to real dialect characterization capture.

Dialect	# of Utterances	Average time (s)	Total Time
Pre-Hilālī (PRH)	449	8.7	01h 05min
Algiers-Blanks (ALG)	4845	6.7	09h 06min
Ma'qilian (MAQ)	3470	8.9	12h 03min
Hilālī-Saharan (SAH)	4560	8.7	10h 57min
Sulaymite (SUL)	4152	8.4	10h 37min
Total	17458	8.3	43h 48min

TABLE 7.8: Speech Data Details.

As indicated previously, we have deployed a set of Open Source softwares to implement and evaluate our HADID system. To extract prosodic features, we have used the same tools (Praat[11], Prosogram[12], and Correlatore[13]) as mention in the previous chapter (Section 6.3).

Concerning the SVM dialect models generation, we have used Weka tool[14], which is one of the most commonly tools in Machine Learning. The SVM kernel deployed is a Radial Basis

[11]Praat v 5.3.47, Online: http://www.fon.hum.uva.nl/praat
[12]Prosogram v 2.9, Online: http://bach.arts.kuleuven.be/pmertens/prosogram/
[13]Correlatore v 2.1, Online: http://www.lfsag.unito.it/correlatore/
[14]Weka v 3.7.13, Online: http://www.cs.waikato.ac.nz/ml/weka/

Function. We have got the best C and γ parameters of SVM, thanks to GridSearch script, which is a meta-classifier (Braga et al. 2013). In other side, for DNNs dialect modeling and test purposes, we have used *H2O*[15] deep learning package scripted with *R* language[16].

7.3.2.2 Hadid Implementation

According to the available hierarchical dialect structure, presented previously in Figure 1.4, and the selected dialects, we have built two LCPNs based on DNNs classifier ($DNNs_1$, $DNNs_2$). Figure 7.16 illustrates the deployed hierarchical dialect models: $DNNs_1$ and $DNNs_2$. $DNNs_1$ classifier is used to classify the first level dialect groups (Pre-Hilālī, Bedouin). It can be seen as a regional dialect identification. Whereas, $DNNs_2$ classifier is used to classify the four Bedouin sub-dialects: Hilālī-Saharan (SAH), Sulaymite (SUL), Ma'qilian (MAQ), Algiers-Blanks (ALB) dialect.

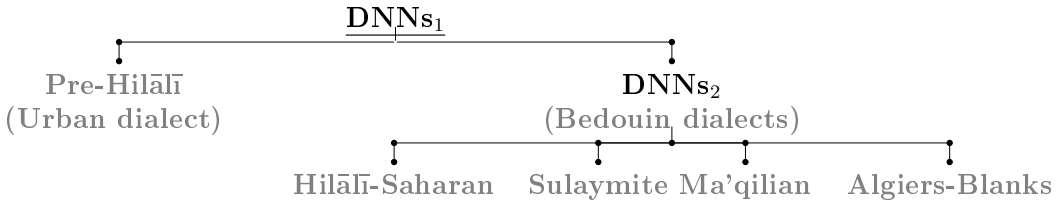


Figure 7.16: Hierarchical Dialect Models.

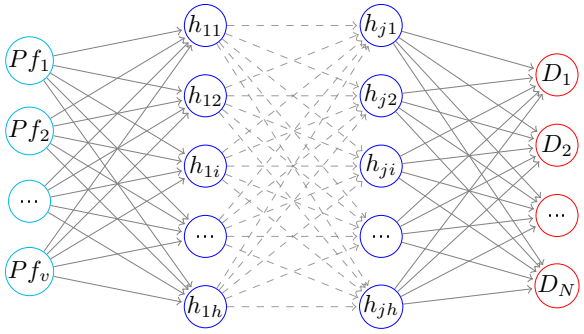
In order to choose the appropriate prosodic feature set for each classifier, we have used the feature selection method: ANOVA ranking (Wu 2009). The features with smallest p-values are selected to be part of the feature set:

1. For $DNNs_1$ classifier, six features are selected. These features are: the duration proportion ($\%V$), the standard deviation of vocalic interval (ΔV), and four pitch values (*Range*, *Mean*, *TrajIntra*, *TrajInter*). Thus, the input layer has 6 linear units. Whereas, the output layer has two units (Pre-Hilālī, Bedouin).

2. For $DNNs_2$ classifier, seven features are selected. These features are: the duration proportion ($\%V$), the standard deviation of consonantal interval (ΔC), and five pitch values (*Range*, *Top*, *Bottom*, *TrajIntra*, *TrajInter*). Thus, the input layer has 7 linear units. Whereas, the output layer has four units.

[15]H2O, Online: http://www.h2o.ai/
[16]R is a language and environment for statistical computing.

Now let us describe both topologies within $DNNs_1$ and $DNNs_2$. Figure 7.17 illustrates the generic DNNs topology and its related description. Each DNNs has four hidden layers with 560 units. A dropout factor of 0.5 is used for each hidden layer. These hyper-parameters are chosen using genetic algorithm.



- **Topology**
  - v : # of input Prosodic features (Pf).
  - j : 4 hidden layers.
  - h : 560 hidden Rectified Linear Units in each hidden layer.
  - N : number of output dialects.
  - Softmax output layer.
- **Training**
  - Stochastic Gradient Descend.
  - A dropout factor of 0.5 in each hidden layer.

FIGURE 7.17: DNNs Topology and Description.

7.3.2.3 Evaluation Metrics

Concerning the performance measure of HADID system, we use the extended version of the well known metric *Precision*, but tailored to the hierarchical classification scenario, namely *hierarchical Precision* (hP) proposed by Kiritchenko et al. (2006). It is defined as follows:

$$hP = \frac{\sum_i |\hat{C}_i \cap \hat{T}_i|}{\sum_i |\hat{C}_i|}$$

where $\hat{C}_i$ is the set consisting of the most specific class(es) **predicted** for test example i and all its ancestor classes, $\hat{T}_i$ is the set consisting of the **true** most specific class(es) of test example i and all its ancestor classes.

The performance of the whole HADID system is measured using Hierarchical micro-precision, which is the average of hP on the entire test utterances. The performance of the baseline Flat classification approach is measured using the standard micro-Precision.

7.3.2.4 Results and Discussion

For all experiments, we use speaker-independent DID and k-folds cross-validation technique ($k = 5$), in order to ensure that our results are reliable.

Our results discussion begins with the performance study of our HADID system. Table 7.9 gives the Hierarchical micro-precision of five folds for HADID-DNN.

	Level-1 (%)	Whole System (%)
Fold 1	98.8	67.1
Fold 2	98.9	57.5
Fold 3	98.7	62.2
Fold 4	98.8	62.3
Fold 5	98.3	56.8
Average	98.7	61.2

TABLE 7.9: HADID-DNN System Precision on Different Cross-validation Folds.

These results confirmed that prosody is suitable to separate region dialects (according to results in Level-1). In fact, $DNNs_1$ classifier separates between Pre-Hilālī dialects and Bedouin ones with $hP = 98.7$ %. The average precision remains acceptable (61.2 %) for all the five dialects in spite of their closeness. We can also observe that the precision of the system is quite stable. In fact, the deviation between the precision for each fold and the average precision doesn't exceed 4.4 %.

HADID-DNN vs. HADID-SVM System

Turning now to the comparison of HADID-DNN with HADID-SVM system. This latter designed as HADID system where dialect modeling uses conventional SVM classifier. Table 7.10 gives the comparative results in term of their hP for 5-fold cross-validation.

Concerning the first level, regional dialect identification, HADID-SVM precision is of the same magnitude as HADID-DNN system. However, it is apparent from that HADID-SVM system is better than HADID-DNN system with an improvement of 5.9% in term of precision. However, detailed results by dialects (see Figure 7.18) proves that HADID-SVM is more suitable than HADID-DNN system. In addition, the precision of all dialects is more than 56%. Indeed, the deviation between best and worst precisions by dialect is about 32% for both HADID systems.

HADID-DNN vs. Flat System

The designed baseline Flat classification system is built using the same preprocessing and segmentation phases as for HADID approach. The speech is also characterized using the best

	Level-1 (%)	Whole System (%)
HADID-SVM	99.4	67.1
HADID-DNN	98.7	61.2

TABLE 7.10: HADID vs. HADID-SVM Performance in Term of Precision.

prosodic features according to ANOVA ranking method. However, one model is generated for all targeted dialects. We have built two flat systems by using DNN and SVM modeling respectively.

Table 7.11 reports comparative results of both Flat classifications and HADID-SVM system in term of precision.

The average precision is about 40% and 27.1% for Flat-SVM and Flat-DNN system, respectively. It is clear that HADID-SVM outperforms this baseline Flat classification system by an improvement more than 27.1%. This main finding proves that for ADID, Hierarchical classification is more suitable than Flat classification. Furthermore, from Figure 7.18, we can observe that the precision of Maaqilian dialect is the less for Flat-SVM system. In addition, we can note that the Flat-SVM classifier confused between Ma'qilian, Hilālī-Saharan and Algiers-Blanks utterances. While, Pre-Hilālī and Sulaymite dialects can be discriminated to others ones because they have their own characteristics.

System	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	**Average**
Flat-SVM	46.3	36.5	45.6	36.4	35.6	**40.0**
Flat-DNN	30.9	30.7	34.9	**05.3**	33.8	**27.1**
HADID-SVM	67.2	67.3	67.2	66.8	67.2	**67.1**

TABLE 7.11: Flat Classification vs. HADID System Precision on Different Cross-validation Folds.

Results by Dialect: HADID-SVM vs HADID-DNN vs. Flat System

Figure 7.18 reports more details on both HADID systems and Flat-SVM results. In fact, it reports average hP by dialect.

The best result for HADID-SVM system is observed for Hilālī-Saharan and Pre-Hilālī dialects with 89% and 82% precision respectively. While, for HADID-DNN, Algiers-Blanks and Sulaymite are the best classified, compared to HADID-SVM, with 70% and 69% precision

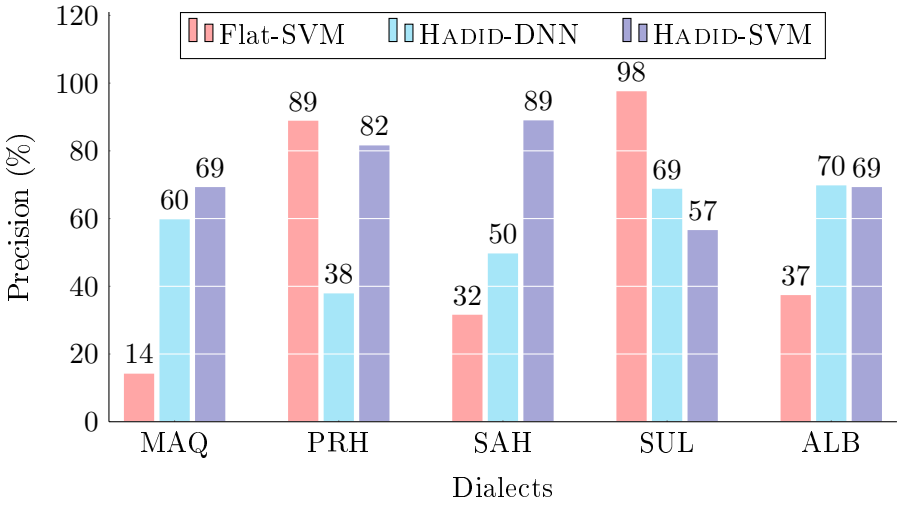


FIGURE 7.18: Comparative Results of Flat Classification, HADID-DNN, and HADID-SVM Systems by Dialect. (SAH= Hilālī-Saharan, SUL= Sulaymite, MAQ= Ma'qilian, ALB= Algiers-Blanks)

respectively. The worst classification result is observed for Pre-Hilālī dialect with 38% precision. This result can be explained by the fact that the DNN classifier need more data. while, for HADID-SVM, it gives better results about 82% precision, which means an improvement of more than 44% compared to HADID-DNN.

A deep examination of results of Flat-SVM classification shows that Ma'qilian utterances are mis classified in others Bedouin dialects (Hilālī-Saharan and Algiers-Blanks). This fact suggests a refinement of the hierarchical dialect structure which needs more dialectologists efforts.

7.4 Discussion and Conclusion

In this chapter, first, we have described our proposed Algerian Arabic Dialect IDentification system (AADID-DNN) that identify Algerian Arabic dialects from a speech characterized at utterance level and leveraging prosody. AADID-DNN was evaluated on Algerian Arabic dialects corpus totaling 42 h. We demonstrated that prosody can be used for Algerian dialect identification. Our system showed its superiority to acoustic-based system in performance when the test utterance sizes are short. When compared with SVM modeling, the AADID-DNN system achieved roughly the same precision, which proves that both SVM and DNN modeling are robust for ADID. In addition, a contrastive acoustic-based system using occurrence UBM-GMM, which is considered as state-of-the-art, has shown that acoustic characterization

outperforms prosodic for almost dialects. But both systems can be complementary to each other.

In another investigation, we have designed a prosody-based Hierarchical Arabic Dialect IDentification (HADID) that identifies Algerian Arabic dialects from a speech. Within HADID system, a top-down method is involved where a Local Classifier per Parent Node (LCPN) is built according to the given predefined hierarchical dialect structure. The LCPN dialect models are constructed using Deep Neural Networks method. HADID-DNN performances are evaluated on Algerian Arabic dialects corpus with 17458 utterances, each one is about 8.3 s duration in average. For region dialects identification, HADID reaches a precision of 98.7% while for dialect identification it gives 61.2%. These results prove its suitability for Arabic Dialect IDentification (ADID). In addition, we have also implemented HADID by using SVM classifier, which gives about 67.1 % of precision for dialect identification. Then, we have compared with Flat-SVM classification system, HADID-SVM gives an improvement of 27.1% in term of precision.

Certainly our dataset covers only a part of Arabic Algerian dialects and the hierarchical dialects structure needs more refinement, this framework can be considered as a kernel for more complete systems. It can deal with all Arabic dialects identification, as long as, we have an efficient hierarchical dialect structure.

Conclusion and Perspectives

Arabic Natural Language Processing (NLP) has received little attention compared to other languages. We have addressed especially to Arabic Dialect IDentification (ADID) problem, when dialects are under resourced and belong to the same country (intra-country), such Algerian one. First, we have investigated the resources language building. In fact, language resources (corpora) are necessary for most of the front-end NLP researches. Such corpora have to be large to achieve NLP communities expectations. In fact, the notion of "More data is better data" was born with the success of modeling based on machine learning and statistical methods.

In our first contribution, we have built two spoken Arabic Algerian corpora: Alg-Daridjah and Kalam'DZ. First, Alg-Daridjah is a parallel rich corpus that provides a representation of phonetic, prosodic and orthographic varieties of Algerian dialects. The rich text content of the speech is designed to highlight the features of the dialect. Indeed, a well chosen words to express all discriminative phones. Text narrations in dialect variations to reflect features of dialects. Second, Kalam'DZ is Web-based corpus. In fact, the used methodology makes building a Web-based corpus that shows the within-language variability. In addition, we have devised a recipe in order to facilitate building large-scale speech corpus which harnesses Web resources.

In the second contribution, we have used altruistic crowdsourcing technique to validate the dialect annotation of Kalam'DZ corpus, which is semi-automatic one created by using web-source meta-data. The results shows that the annotation of crowds agree with 81% to this done by expert. From this experience with altruistic crowdsourcing, we have determined a list of guidelines to follow when using free platforms and crowdsourcing.

In the third contribution, we have addressed the problem of Arabic Dialect IDentification (ADID) in an intra-country context where dialects are very similar and close. First, we have explained which speech information are selected and how we have extracted to characterize Algerian dialects. Then, we have performed two statistical analysis tests: one way ANOVA

and Tukey tests. As results, we observed that the intonation information are more discriminative than rhythm one. Then, on the light of this analysis, two Algerian ADID systems are proposed by using prosodic information:

1. We designed a flat Algerian Arabic Dialect IDentification system (AADID-DNN) that identify dialects from a speech characterized at utterance level leveraging prosody. AADID-DNN was evaluated by using a part from KALAM'DZ corpus (42 h). Then, we have compared the obtained results of DNN classifier with SVM one AADID-SVM. We observe that both approaches have reached comparable performances. To compare the discriminate power between Prosodic information and Acoustic one, we have built the acoustic approach, as Djellab et al. (2017), which is based on spectral features and Gaussian Mixture Models-Uniform Background Model (GMM-UBM) for dialect modelings. Finally, we have combined the both approaches.
2. We have also designed a prosody-based Hierarchical Arabic Dialect IDentification (HADID) that identifies Algerian Arabic dialects from a speech. Within HADID system, a top-down method is involved where a Local Classifier per Parent Node (LCPN) is built according to the given predefined hierarchical dialect structure. The LCPN dialects models are constructed using SVM classifier.

The main findings of our whole study are:

1. The prosodic features are discriminative information for ADID.
2. Compared to acoustic features, the prosodic ones are less discriminative than acoustic one but its showed their superiority in performance when the test utterance sizes are short.
3. The combination of acoustic and prosodic features improves the ADID systems.
4. SVM modeling is still robust compared to DNNs modeling at least for our data size.
5. The HADID system beats flat ones and can be gives better results.

This research has thrown up several improvements such as:

- Our both corpora include only Algerian Arabic dialects, which is used by 70 to 80% of the population. In addition to Arabic dialects, the Algerian's population speaks also the Tamazight (Berber language) with its different varieties, which is mother tongue of 25% to 30% of population (Saadane & Habash 2015). We will extend our corpora by collecting Algerian Tamazight uttered by berber native speakers.

- Algerian dialects is included to Maghrebi ones which has been affected by the same both Arabization and colonization phases and share many linguistic features. We will built Web-based corpora for each Maghrebi dialect. Then, we will apply our intra-country ADID approaches to Maghrebi dialects to exploit inter-dialect context.

- Ali et al. (2017) perform the combination of lexical and acoustic approach to evaluate ADID for inter-country context, which reaches an average about 75.1% of precision. We will investigate to combine the lexical, prosody, and acoustic approach where the combination is performed at the scoring level.

- We have performed our HADID approach by using top-down method, which is involved Local Classifier at each Parent Node (LCPN). There are more than one possible ways to perform HADID implementation. We plan to investigate other global hierarchical methods (Silla et al. 2011) .

Bibliography

Abercombie, D. (1967), *Elements of General Phonetics*, Aldine Pub. Company.

Abuata, B. & Al-Omari, A. (2015), 'A Rule-based Stemmer for Arabic Gulf Dialect', *Journal of King Saud University - Computer and Information Sciences* **27**(2), 104 – 112.

Abushariah, M. A. M., Ainon, R. N., Zainuddin, R., Elshafei, M. & Khalifa, O. O. (2012), 'Phonetically Rich and Balanced Text and Speech Corpora for Arabic Language', *Language Resources and Evaluation* **46**(4), 601–634.

Aggarwal, C. C. (2018), *Neural Networks and Deep Learning - A Textbook*, Springer.

Akbacak, M., Vergyri, D., Stolcke, A., Scheffer, N. & Mandal, A. (2011), Effective Arabic Dialect Classification Using Diverse Phonotactic Models, *in* 'Interspeech, Italy', pp. 737–740.

Al-Ayyoub, M., Nuseir, A., Alsmearat, K., Jararweh, Y. & Gupta, B. (2018), 'Deep Learning for Arabic NLP: A Survey', *Journal of Computational Science* **26**, 522 – 531.

Al-Ayyoub, M., Rihani, M. K., Dalgamoni, N. I. & Abdulla, N. A. (2014), Spoken Arabic Dialects Identification: The Case of Egyptian and Jordanian Dialects, *in* 'International Conference on Information and Communication Systems (ICICS), Jordan', pp. 1–6.

Al-Sharkawi, M. (2016), *History and Development of the Arabic Language*, Taylor & Francis.

Al-Walaie, M. A. & Khan, M. B. (2017), Arabic Dialects Classification Using Text Mining Techniques, *in* 'International Conference on Computer and Applications (ICCA)', pp. 325–329.

Alghamdi, M., Alhargan, F., Alkanhal, M., Alkhairy, A., Eldesouki, M. & Alenazi, A. (2008), 'Saudi Accented Arabic Voice Bank', *Journal of King Saud University-Computer and Information Sciences* **20**, 45–64.

Ali, A., Najim, D., Patrick, C., Sameer, K., Sree Harsha, Y., Glass, J., Bell, P. & Renals, S. (2016), 'Automatic Dialect Detection in Arabic Broadcast Speech', *arXiv preprint arXiv:1509.06928v2* .

Ali, A., Vogel, S. & Renals, S. (2017), 'Speech Recognition Challenge in the Wild: Arabic MGB-3', *arXiv preprint arXiv:1709.07276* .

Almeman, K., Lee, M. & Almiman, A. A. (2013), Multi Dialect Arabic Speech Parallel Corpora, *in* 'Communications, Signal Processing, and their Applications (ICCSPA)', pp. 1–6.

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

Alsulaiman, M., Muhammad, G., Bencherif, M. A., Mahmood, A. & Ali, Z. (2013), 'KSU Rich Arabic Speech Database', *Journal of Information* **16**(6), 4231–4253.

Ambikairajah, E., Li, H., Wang, L., Yin, B. & Sethu, V. (2011), 'Language Identification: A Tutorial', *IEEE Circuits and Systems Magazine* **11**(2), 82–108.

Ammar, Z., Fougeron, C. & Ridouane, R. (2014), A la recherche des traces dialectales dans l'arabe standard: production des voyelles et des fricatives inter-dentales par des locuteurs tunisiens et marocains, *in* '30^{mes} Journées d'Etudes sur la Parole (JEP), Le Mans, France', pp. 684–693.

Anjos, A., El-Shafey, L., Wallace, R., Günther, M., McCool, C. & Marcel, S. (2012), Bob: a Free Signal Processing and Machine Learning Toolbox for Researchers, *in* 'Proceedings of the 20th ACM International Conference on Multimedia', pp. 1449–1452.

Barkat, M., Ohala, J. & Pellegrino, F. (1999), Prosody as a Distinctive Feature for the Discrimination of Arabic Dialects, *in* 'Eurospeech, Budapest, Hungary', pp. 395–398.

Beasley, D., Bull, D. R. & Martin, R. R. (1993), 'An overview of genetic algorithms: Part 1, fundamentals', *University computing* **15**(2), 56–69.

Behnstedt, P. & Woidich, M. (2013), Dialectology, *in* J. Owens, ed., 'The Oxford Handbook of Arabic Linguistics', pp. 300–325.

Benali, I. (2004), Le rôle de la prosodie dans l'identification de deux parlers algériens: l'algérois et l'oranais, *in* 'MIDL', Paris, France, pp. 128–132.

Beysolow, T. (2018), *Applied Natural Language Processing with Python: Implementing Machine Learning and Deep Learning Algorithms for Natural Language Processing*, Apress.

Biadsy, F. & Hirschberg, J. (2009), Using Prosody and Phonotactics in Arabic Dialect Identification, *in* 'Interspeech', Brighton, UK, pp. 208–211.

Biadsy, F., Hirschberg, J. & Ellis, D. P. W. (2011), Dialect and Accent Recognition Using Phonetic-Segmentation Supervectors, *in* 'Interspeech', Florence, Italy, pp. 745–748.

Biadsy, F., Hirschberg, J. & Habash, N. (2009), Spoken Arabic Dialect Identification Using Phonotactic Modeling, *in* 'Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages', Athens, Greece, pp. 53–61.

Borromeo, R. M. & Toyama, M. (2016), 'An Investigation of Unpaid Crowdsourcing', *Human-centric Computing and Information Sciences* **6**(1).

Boudraa, M., Boudraa, B. & Guerin, B. (2000), 'Twenty Lists of Ten Arabic Sentences for Assessment', *Acta Acustica united with Acustica* **86**(5), 870–882.

Bougrine, S., Cherroun, H. & Abdelali, A. (2017), 'Altruistic Crowdsourcing for Arabic Speech Corpus Annotation', *Procedia Computer Science* **117**, 137 – 144.

Bougrine, S., Cherroun, H. & Abdelali., A. (2018), Spoken Arabic Algerian Dialect Identification, *in* 'IEEE the International Conference on Natural Language and Speech Processing (ICNLSP)', Algiers, Algeria.

Bougrine, S., Cherroun, H. & Ziadi, D. (2017), 'Hierarchical Classification for Spoken Arabic Dialect Identification using Prosody: Case of Algerian Dialects', *arXiv preprint arXiv:1703.10065* .

Bougrine, S., Cherroun, H. & Ziadi, D. (2018), 'Prosody-based Spoken Algerian Arabic Dialect Identification', *Procedia Computer Science* **128**, 9 – 17.

Bougrine, S., Cherroun, H., Ziadi, D., Lakhdari, A. & Chorana, A. (2016), Toward a Rich Arabic Speech Parallel Corpus for Algerian sub-Dialects, *in* 'LREC'16 Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT)', pp. 2–10.

Bougrine, S., Chorana, A., Lakhdari, A. & Cherroun, H. (2017), Toward a Web-based Speech Corpus for Algerian Dialectal Arabic Varieties, *in* 'Workshop on Arabic Natural Language Processing (WANLP), Association for Computational Linguistics', pp. 138–146.

Bouziri, R., Nejmi, H. & Taki, M. (1991), L'accent de l'arabe parlé à Casablanca et à Tunis: étude phonétique et phonologique, *in* 'ICPhS, Aix-en-Provence', pp. 134–137.

Braga, I., do Carmo, L. P., Benatti, C. C. & Monard, M. C. (2013), A note on parameter selection for support vector machines, *in* F. Castro, A. Gelbukh & M. González, eds, 'Advances in Soft Computing and Its Applications', Berlin, Heidelberg, pp. 233–244.

Canavan, A. & Zipperlen, G. (1996), 'CALLFRIEND Egyptian Arabic LDC96S49.', Philadelphia: Linguistic Data Consortium.

Canavan, A., Zipperlen, G. & Graff, D. (1997), 'CALLHOME Egyptian Arabic Speech LDC97S45', Philadelphia: Linguistic Data Consortium.

Caubet, D. (2000), 'Questionnaire de dialectologie du Maghreb (d'après les travaux de W. Marçais, M. Cohen, GS Colin, J. Cantineau, D. Cohen, Ph. Marçais, S. Lévy, etc.)', *Estudios de Dialectología Norteafricana y Andalusí (EDNA)* **5**, 73–90.

Chen, N. F., Shen, W. & Campbell, J. P. (2010), A Linguistically-informative Approach to Dialect Recognition using Dialect-discriminating Context-dependent Phonetic Models, *in* 'International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 5014–5017.

Chiroma, H., Mohd Noor, A. S., Abdulkareem, S., I. Abubakar, A., Hermawan, A., Qin, H., Hamza, M. & Herawan, T. (2017), 'Neural Networks Optimization through Genetic Algorithm Searches: A Review', *Applied Mathematics & Information Sciences* **11**, 1543–1564.

Chittaragi, N. B., Prakash, A. & Koolagudi, S. G. (2018), 'Dialect Identification Using Spectral and Prosodic Features on Single and Ensemble Classifiers', *Arabian Journal for Science and Engineering* **43**(8), 4289–4302.

Chollet, F. (2017), *Deep Learning with Python*, Manning Publications Co., Greenwich, CT, USA.

Choukri, K., Hamid, S. & Paulsson, N. (2005), 'Specifiaction of the Arabic Broadcast News Speech Corpus'. NEMLAR, Center for Sprogteknologi.

Choukri, K., Nikkhou, M. & Paulsson, N. (2004), Network of Data Centres (NetDC): BNSC - An Arabic Broadcast News Speech Corpus., *in* 'LREC', European Language Resources Association.

Crystal, D. (2008), *A Dictionary of Linguistics and Phonetics*, Vol. 6 of *Lecture Notes in Statistics*, Blackwell, Malden.

Dauer, R. M. (1983), 'Stress-timing and Syllable-timing Reanalyzed', *Journal of phonetics* **11**, 21–62.

Davis, S. & Mermelstein, P. (1980), 'Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences', *IEEE Transactions on Acoustics, Speech and Signal Processing* **28**(4), 357–366.

Dellwo, V. (2006), Rhythm and Speech Rate: a Variation Coefficient for deltaC, *in* 'Language and language-processing', pp. 231–241.

Di, W., Bhardwaj, A. & Wei, J. (2018), *Deep Learning Essentials: Your Hands-on Guide to the Fundamentals of Deep Learning and Neural Network Modeling*, Packt Publishing.

Diab, M. & Habash, N. (2008), Arabic Dialect Processing. [Online; accessed 01/05/2018].
URL: *http://www.lrec-conf.org/proceedings/lrec2008/workshops/T5.pdf*

Djellab, M., Amrouche, A., Bouridane, A. & Mehallegue, N. (2017), 'Algerian Modern Colloquial Arabic Speech Corpus (AMCASC): Regional Accents Recognition Within Complex Socio-linguistic Environments', *Language Resources and Evaluation* **51**(3), 613–641.

Droua-Hamdani, G., Selouani, S. & Boudraa, M. (2010), 'Algerian Arabic Speech Database (ALGASD): Corpus Design and Automatic Speech Recognition Application', *Arabian Journal for Science and Engineering* **35**, 158.

Du, T. & Shanker, V. (2009), 'Deep learning for natural language processing', *Eecis. Udel. Edu* pp. 1–7.

Ekkehard Wolff (2018), 'Afro-Asiatic Languages'. Encyclopædia Britannica, inc., [Online; accessed 30/05/2018].
URL: *https://www.britannica.com/topic/Afro-Asiatic-languages*

Embarki, M. (2008), 'Les dialectes arabes modernes: état et nouvelles perspectives pour la classification géo-sociologique', *Arabica* **55**(5), 583–604.

Etman, A. & Beex, A. L. (2015), Language and Dialect Identification: A Survey, *in* 'Intelligent Systems Conference (IntelliSys)', pp. 220–231.

Farghaly, A. (2010), 'The Arabic language, Arabic linguistics and Arabic computational linguistics', *Arabic Computational Linguistics* pp. 43–81.

Ferrer, L., Lei, Y., McLaren, M. & Scheffer, N. (2016), 'Study of Senone-Based Deep Neural Network Approaches for Spoken Language Recognition', *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **24**(1), 105–116.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. & Dahlgren, N. L. (1993), 'DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM'.

Ghazali, S., Hamdi, R. & Barkat, M. (2002), Speech Rhythm Variation in Arabic Dialects, *in* 'Proceedings of Prosody, Aix-en-Provence, France', pp. 331–334.

Ghazali, S., Hamdi, R. & Knis, K. (2005), Intonational and Rhythmic patterns Across the Arabic Dialect Continuum, *in* '19th Annual Symposium on Arabic Linguistics', Illinois, USA, pp. 97–121.

Gibbon, D., Moore, R. & Winski, R. (1997), *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter.

Goyal, P., Pandey, S. & Jain, K. (2018), *Deep Learning for Natural Language Processing*, Apress.

Grabe, E. & Low, E. L. (2002), 'Durational Variability in Speech and the Rhythm Class Hypothesis', *Papers in laboratory phonology* **7**, 515–546.

Graja, M., Jaoua, M. & Hadrich-Belguith, L. (2010), Lexical Study of A Spoken Dialogue Corpus in Tunisian Dialect, *in* 'The International Arab Conference on Information Technology (ACIT)', Benghazi, Libya.

Greenberg, C., Martin, A. & Przybocki, M. (2012), The 2011 NIST Language Recognition Evaluation, *in* 'Interspeech', Portland, Oregon, USA.

Habash, N. (2010), *Introduction to Arabic Natural Language Processing*, Synthesis Lectures on Human Language Technologies, Morgan and Claypool Publishers.

Haeri, N. (2003), *Sacred language, ordinary people: Dilemmas of culture and politics in Egypt*, Springer.

Hakouz, W., Ghanem, A., Wray, S. & Ali, A. (2015), LAHAJET: a Game for Classifiying Dialectal Arabic Speech, *in* 'ASRU'.

Halabi, N. (2016), Modern Standard Arabic Phonetics for Speech Synthesis, PhD thesis, University of Southampton.

Hamdi, R. (2007), La variation rythmique dans les dialectes arabes, PhD thesis, University Lumière Lyon 2 and University 7 Novembre, Tunisie.

Hamdi, R., Barkat-Defradas, M., Ferragne, E. & Pellegrino, F. (2004), Speech Timing and Rhythmic structure in Arabic dialects: a comparison of two approaches, *in* 'InterSpeech', pp. 1613–1616.

Hamdi, R., Ghazali, S. & Barkat, M. (2005), Syllable Structure in Spoken Arabic : a comparative investigation, *in* 'Interspeech', Lisbonne, Portugal, pp. 2245–2248.

Hanani, A. (2012), Human and computer recognition of regional accents and ethnic groups from British English speech, PhD thesis, The University of Birmingham.

Hanani, A., Basha, H., Sharaf, Y. & Taylor, S. (2015), Palestinian Arabic Regional Accent Recognition, *in* 'Speech Technology and Human-Computer Dialogue (SpeD)', pp. 1–6.

Harper, M. P. & Maxwell, M. (2008), Spoken language characterization, *in* 'Springer Handbook of Speech Processing', Springer, pp. 797–810.

Harrat, S., Meftouh, K., Abbas, M., Hidouci, K.-W. & Smaili, K. (2016), 'An Algerian dialect: Study and Resources', *International Journal of Advanced Computer Science and Applications (IJACSA)* **7**(3).

Harrat, S., Meftouh, K. & Smaïli, K. (2017), Maghrebi Arabic Dialect Processing: an Overview, *in* 'International Conference on Natural Language, Signal and Speech Processing (ICNLSSP)'.

Himmelmann, N. P. & Ladd, D. R. (2008), 'Prosodic Description: An Introduction for Fieldworkers'.

Holes, C. (2004), *Modern Arabic: Structures, Functions, and Varieties*, Georgetown University Press.

Jabbari, M. J. (2013), 'Levantine Arabic: A Surface Register Contrastive Study', *Studies in Literature and Language* **6**(1), 110–116.

James, A. L. (1940), *Speech signals in telephony*, Sir I. Pitman & sons, Limited.

Jothilakshmi, S., Ramalingam, V. & Palanivel, S. (2012), 'A hierarchical language identification system for Indian languages', *Digital Signal Processing* **22**(3), 544–553.

Kassem, L., Sabty, C., Sharaf, N., Bakry, M. & Abdennadher, S. (2016), tashkeelWAP: A Game With A Purpose For Digitizing Arabic Diacritics, *in* 'ACLing'.

Kaufmann, N. & Schulze, T. (2011), 'Worker Motivation in Crowdsourcing and Human Computation', *Education* **17**(2009), 6.

Khan, G., Streck, M. P. & Watson, J. C. (2012), *The Semitic languages: An international handbook*, Vol. 36, Walter de Gruyter.

Khoury, E., El Shafey, L. & Marcel, S. (2014), Spear: An open source toolbox for speaker recognition based on bob, *in* 'ICASSP', IEEE, pp. 1655–1659.

Kirchhoff, K. (2006), *in Multilingual Speech Processing*, Elsevier, chapter Language characteristics, pp. 5–31.

Kiritchenko, S., Matwin, S., Nock, R. & Famili, A. F. (2006), Learning and Evaluation in the Presence of Class Hierarchies: Application to Text Categorization, *in* 'Proceedings of the 19th Canadian Conference on Artificial Intelligence', Québec, Canada., pp. 395–406.

Koolagudi, S. G., Bharadwaj, A., Murthy, Y. S., Reddy, N. & Rao, P. (2017), 'Dravidian Language Classification from Speech Signal using Spectral and Prosodic Features', *International Journal of Speech Technology* **20**(4), 1005–1016.

Kumar, E. (2011), *Natural Language Processing*, IK International Pvt Ltd.

Kumar, G., Cao, Y., Cotterell, R., Callison-Burch, C., Povey, D. & Khudanpur, S. (2014), Translations of the CALLHOME Egyptian Arabic Corpus for Conversational Speech Translation, *in* 'IWLST'.

Kurdi, M. Z. (2016), *Natural Language Processing and Computational Linguistics 1: Speech, Morphology and Syntax*.

Lachachi, N. & Adla, A. (2015), 'GMM-Based Maghreb Dialect Identification System', *Journal of Information Processing Systems* **11**(1), 22–38.

Laguna, A. F. B. & Guevara, R. C. L. (2015), 'Development, Implementation and Testing of Language Identification System for Seven Philippine Languages', *Philippine Journal of Science* **144**(1), 81–89.

LaRocca, S. & Chouairi, R. (2002), 'West Point Arabic Speech LDC2002S02', Philadelphia: Linguistic Data Consortium.

Leclerc, J. (2012), 'Algérie dans l'aménagement linguistique dans le monde', Web. [Online; accessed 15/06/2018].
URL: *http://www.tlfq.ulaval.ca/axl/europe/danemark.htm*

Leonard, R. & Doddington, G. (1974), 'Automatic Language Identification, Technical report RADC-TR-74-200, Air Force Rome Air Development Center'.

Lewis, M. P. e. (2009), *Ethnologue: Languages of the World, Sixteenth edition*, Dallas, Tex.: SIL International.

Li, H., Ma, B. & Lee, K. (2013), 'Spoken Language Recognition: From Fundamentals to Practice', *Proceedings of the IEEE* **101**(5), 1136–1159.

Lindquist, H. (2009), *Corpus Linguistics and the Description of English*, Edinburgh University Press Series, Edinburgh University Press.

Lopez-Moreno, I., Gonzalez-Dominguez, J., Martinez, D., Plchot, O., Gonzalez-Rodriguez, J. & Moreno, P. J. (2016), 'On the Use of Deep Feedforward Neural Networks for Automatic Language Identification', *Computer Speech and Language* **40**, 46 – 59.

Loukina, A., Kochanski, G., Rosner, B., Keane, E. & Shih, C. (2011), 'Rhythm measures and dimensions of durational variation in speech', *The Journal of the Acoustical Society of America* **129**(5), 3258–3270.

Makhoul, J., Zawaydeh, B., Choi, F. & Stallard, D. (2005), 'BBN/AUB DARPA Babylon Levantine Arabic Speech and Transcripts', Linguistic Data Consortium (LDC). LDC Catalog Number LDC2005S08.

Manaswi, N. K. (2018), *Deep Learning with Applications Using Python: Chatbots and Face, Object, and Speech Recognition With TensorFlow and Keras*, 1st edn, Apress, Berkely, CA, USA.

Mansour, M. A. (2013), 'The Absence of Arabic Corpus Linguistics: A Call for Creating an Arabic National Corpus', *International Journal of Humanities and Social Science* **3**(12), 81–90.

Marçais, P. (1977), *Esquisse grammaticale de l'arabe maghrébin*, Lib. d'Amérique et d'Orient.

Marçais, P. (1986), *Encyclopaedia of Islam I*, chapter Algeria, pp. 374–379.

Mayer, M. (1969), *Frog, where are you?*, New York: Dial Press.

Meignier, S. & Merlin, T. (2010), LIUM SpkDiarization: an Open Source Toolkit for Diarization, *in* 'CMU SPUD Workshop'.

Mertens, P. (2004), The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model, *in* 'Speech Prosody', Nara, Japan, pp. 23–26.

Moftah, M., Fakhr, M. W. & El Ramly, S. (2018), Arabic Dialect Identification Based on Motif Discovery using GMM-UBM with Different Motif Lengths, *in* 'International Conference on Natural Language and Speech Processing (ICNSP)', Algiers, Algeria.

Muthusamy, Y. K., Jain, N. & Cole, R. A. (1994), Perceptual benchmarks for automatic language identification, *in* 'International Conference on Information and Communication Systems (ICICS)', Vol. 1, pp. 333–336.

Nair, V. G. (2014), *Getting Started with Beautiful Soup*, Packt Publishing.

Najafian, M., Khurana, S., Shon, S., Ali, A. & Glass, J. (2018), Exploiting Convolutional Neural Networks for Phonotactic Based Dialect Identification, *in* 'International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 5174–5178.

Navrátil, J. (2006), *Automatic Language Identification*, Elsevier, pp. 233–272.

Nespor, M., Shukla, M. & Mehler, J. (2011), 'Stress-timed vs. Syllable-timed Languages', pp. 1147–1159.

Omer, A. I., Zampieri, M. & Oakes, M. (2018), Phonetic Differences for Dialect Clustering, *in* 'International Conference on Information and Communication Systems (ICICS)', pp. 145–150.

Palva, H. (1991), 'Is There a North West Arabian Dialect Group', *Festgabe für Hans-Rudolf Singer* pp. 151–166.

Palva, H. (2006), *Encyclopedia of Arabic Language and Linguistics. Vol. I*, Leiden Boston: Brill, chapter Arabic dialects: Classification, pp. 604–613.

Pasupa, K. & Sunhem, W. (2016), A Comparison between Shallow and Deep Architecture Classifiers on Small Dataset, *in* 'International Conference on Information Technology and Electrical Engineering (ICITEE)', pp. 1–6.

Paul, L. M., Simons, G. F. & Fennig, C. D. (2009), *Ethnologue: Languages of the World, Twentieth edition*, Dallas, Texas: SIL International.

Pereira, C. (2011), Arabic in the North African Region, *in* S. W. et al., ed., 'Semitic Languages. An International Handbook', De Gruyter, Berlin, pp. 944–959.

Pereltsvaig, A. (2012), *Languages of the World: An Introduction*, Languages of the World: An Introduction, Cambridge University Press.

Peter, L. (1975), 'A Course in Phonetics', *University of California* .

Pike, K. L. (1945), *The Intonation of American English*, ERIC.

Postgate, J. N. (2007), *Languages of Iraq, Ancient and Modern*, British School of Archaeology in Iraq.

Quinn, A. J. & Bederson, B. B. (2011), Human Computation: A Survey and Taxonomy of a Growing Field, *in* 'Proceedings of the Conference on Human Factors in Computing Systems', pp. 1403–1412.

Ramus, F., Nespor, M. & Mehler, J. (1999), 'Correlates of Linguistic Rhythm in the Speech Signal', *Cognition* **75**(1), 265–292.

Rao, K. S., Reddy, V. R. & Maity, S. (2015), *Language Identification Using Spectral and Prosodic Features*, Springer.

Richardson, F., Campbell, W. & Torres-Carrasquillo, P. (2009), Discriminative N-Gram Selection for Dialect Recognition, *in* 'Interspeech', Brighton, UK, pp. 297–307.

Rodríguez-Fuentes, L. J., Penagarikano, M., Varona, A., Diez, M. & Bordel, G. (2016), 'KALAKA-3: a Database for the Assessment of Spoken Language Recognition Technology on YouTube Audios', *Language Resources and Evaluation* **50**(2), 221–243.

Rouas, J. L. (2005), Caractérisation et identification automatique des langues, PhD thesis, Université Paul Sabatier, Toulouse, France.

Rouas, J. L. (2007), 'Automatic Prosodic Variations Modeling for Language and Dialect Discrimination', *IEEE Transactions on Audio, Speech, and Language Processing* **15**(6), 1904–1911.

Rouas, J. L., Barkat-Defradas, M., Pellegrino, F. & Hamdi, R. (2006), Identification automatique des parlers arabes par la prosodie, *in* 'Journées d'Etude sur la Parole', Dinard, France, pp. 193–196.

Saadane, H. & Habash, N. (2015), A Conventional Orthography for Algerian Arabic, *in* 'Proceedings of the Second Workshop on Arabic Natural Language Processing (ANLP)', p. 69–79.

Sabou, M., Bontcheva, K., Derczynski, L. & Scharl, A. (2014), Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines, *in* 'Language Resources and Evaluation Conference (LREC)', pp. 859–866.

Salselas, I. & Herrera, P. (2011), 'Music and Speech in Early Development: Automatic Analysis and Classification of Prosodic Features from Two Portuguese Variants', *J. Portuguese Linguistics* **9**(10), 11–36.

Sayahi, L. (2014), *Diglossia and Language Contact: Language Variation and Change in North Africa*, Cambridge Approaches to Language Contact, Cambridge University Press.

Sheppard, C. (2018), *Genetic Algorithms with Python*, Clinton Sheppard.

Shon, S., Ali, A. & Glass, J. (2017), 'MIT-QCRI Arabic Dialect Identification System for the 2017 Multi-genre Broadcast Challenge', *arXiv preprint arXiv:1709.00387* .

Shoufan, A. & Alameri, S. (2015), Natural Language Processing for Dialectical Arabic: A Survey, *in* 'Proceedings of the Second Workshop on Arabic Natural Language Processing (ANLP)', pp. 36–48.

Silla, J., Carlos, N. & Freitas, A. A. (2011), 'A Survey of Hierarchical Classification Across Different Application Domains', *Data Mining and Knowledge Discovery* **22**(1-2), 31–72.

Skansi, S. (2018), *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence*, Undergraduate Topics in Computer Science, Springer International Publishing.

Suleiman, Y. (1994), *Nationalism and the Arabic Language: A Historical Overview*, Surrey: Curzon Press Ltd.

Taine-Cheikh, C. (2000), 'Deux macro-discriminants de la dialectologie arabe (la réalisation du qâf et des interdentales)', *Matériaux arabes et sudarabiques (GELLAS)* **9**, 11–51.

Thompson Irene (2016), 'Arabic (Egyptian Spoken)'. [Online: accessed 24/05/2018].
URL: *http://aboutworldlanguages.com/arabic-egyptian*

Torres-Carrasquillo, P. A., Gleason, T. P. & Reynolds, D. A. (2004), Dialect Identification Using Gaussian Mixture Models, *in* 'Odyssey: The Speaker and Language Recognition Workshop', Toledo, Spain, pp. 297–300.

Van Leeuwen, D. A., De Boer, M. & Orr, R. (2008), A Human Benchmark for the NIST Language Recognition Evaluation 2005., *in* 'Odyssey: The Speaker and Language Recognition Workshop', Stellenbosch, South Africa.

Versteegh, K. (1997), *The Arabic Language*, Edinburgh University Press, Cambridge.

Versteegh, K. (2014), *Arabic Language*, 2 edn, Edinburgh University Press, pp. 192–220.

Versteegh, K., Eid, M., Elgibali, A., Woidich, M. & Zaborski, A. (2006), 'Encyclopedia of Arabic Language and Linguistics', *African Studies* **8**.

Wang, A., Hoang, C. D. V. & Kan, M.-Y. (2013), 'Perspectives on Crowdsourcing Annotations for Natural Language Processing', *Language Resources and Evaluation* **47**(1), 9–31.

Wang, H., Xiao, X., Zhang, X., Zhang, J. & Yan, Y. (2009), A Hierarchical System Design for Language Identification, *in* 'International Symposium on Information Science and Engineering (ISISE)', pp. 443–446.

Wells, J. C. (1982), *Accents of English*, Vol. 2, Cambridge University Press.

White, L. & Mattys, S. L. (2007), 'Calibrating Rhythm: First Language and Second Language Studies', *Journal of Phonetics* **35**(4), 501–522.

Wong, K.-Y. E. (2004), Automatic Spoken Language Identification Utilizing Acoustic and Phonetic Speech Information, PhD thesis, Queensland University of Technology.

Wray, S. & Ali, A. (2015), Crowdsource a Little to Label a Lot: Labeling a Speech Corpus of Dialectal Arabic, *in* 'Interspeech', pp. 2824–2828.

Wray, S., Mubarak, H. & Ali, A. (2015), Best Practices for Crowdsourcing Dialectal Arabic Speech Transcription, *in* 'Proceedings of the Second Workshop on Arabic Natural Language Processing (ANLP)', pp. 99–107.

Wu, T. (2009), Feature Selection in Speech and Speaker Recognition, PhD thesis, Katholieke Universiteit Leuven, Leuven.

Yacoub, M. (2016), 'English Loanwords in the Egyptian Variety of Arabic: Morphological and Phonological Changes', *World Academy of Science, Engineering and Technology, International Journal of Cognitive and Language Sciences* **3**(6), 121–136.

Yeou, M., Embarki, M. & Al-Maqtari, S. (2007), 'Contrastive focus and F0 patterns in three Arabic dialects', *Nouveaux Cahiers de Linguistique Française* **28**, 317–326.

Yin, B., Ambikairajah, E. & Chen, F. (2007), Hierarchical Language Identification based on Automatic Language Clustering, *in* 'Interspeech', pp. 178–181.

Young, T., Hazarika, D., Poria, S. & Cambria, E. (2017), 'Recent Trends in Deep Learning Based Natural Language Processing', *CoRR* **abs/1708.02709**.
URL: *http://arxiv.org/abs/1708.02709*

Zaghouani, W. (2014), Critical Survey of the Freely Available Arabic Corpora, *in* 'LREC'14 Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT)', Reykjavik, Iceland, pp. 1–8.

Zaghouani, W. & Dukes, K. (2014), Can Crowdsourcing be Used for Effective Annotation of Arabic?, *in* 'Language Resources and Evaluation Conference (LREC)', pp. 224–228.

Zaidan, O. & Callison-Burch, C. (2014), 'Arabic Dialect Identification', *Computational Linguistics* **40**(1), 171–202.

Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y. & Aepli, N. (2017), Findings of the VarDial Evaluation Campaign 2017, *in* 'Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)', Valencia, Spain, pp. 1–15.

Zissman, M. A., Gleason, T. P., Rekart, D. M. & Losiewicz, B. L. (1996), Automatic Dialect Identification of Extemporaneous Conversational, Latin American Spanish Speech, *in* 'ICASSP', Atlanta, Georgia, pp. 777–780.

Zitouni, I. (2016), *Natural Language Processing of Semitic Languages*, Springer.

Appendix A

Note on Transcription

In our thesis, we adopt for transliteration of sentences the following transcript of Arabic characters.

TABLE A.1: Transcription of Arabic Characters (Jabbari 2013, Versteegh, K. 2014).

Name	Arabic script	Transliteration	Arabic Example	Meaning	IPA sign
’alif (hamza)	ا	’	انا /’ana/	I	[ʔ]
bā’	ب	b	بحر /baḥr/	sea	[b]
tā’	ت	t	تمر /tamr/	dates	[t]
t̲ā’	ث	t̲	ثلاجه /t̲allaja/	refrigerator	[θ]
jīm	ج	j	جمل /jamal/	camel	[ʤ]
ḥā’	ح	ḥ	حبيب /ḥabīb/	friend	[ħ]
ẖā’	خ	ẖ	خال /ẖāl/	uncle	[x]
dā’	د	d	درس /dars/	lesson	[d]
d̲ā’	ذ	d̲	ذلك /d̲ālika/	that	[ð]
rā’	ر	r	روح /rūh/	soul	[r]
zāy	ز	z	زهر /zahr/	bloom	[z]
sīn	س	s	سيارة /sayāra/	car	[s]
šīn	ش	š	شيئ /šay’/	thing	[ʃ]
ṣād	ص	ṣ	صباح /ṣabāḥ/	morning	[sˤ]

Continued on next page

Name	Arabic script	Transliteration	Arabic Example	Meaning	IPA sign
ḍād	ض	ḍ	ضيف /ḍayf/	guest	[dˤ]
ṭā’	ط	ṭ	طالب /ṭālib/	student	[tˤ]
ḍ̇ā’	ظ	ḍ̇	ظرف /ḍ̇arf/	envelope	[ðˤ]
ʿayn	ع	ʿ	عين /ʿayn/	eye	[ʕ]
ġayn	غ	ġ	غدا /ġadan/	tomorrow	[ɣ]
fā’	ف	f	فرنسا /faransā/	France	[f]
qāf	ق	q	قريب /qarīb/	relative	[q]
kāf	ك	k	كتاب /kitāb/	book	[k]
lām	ل	l	لك /laka/	for you	[l]
mīm	م	m	من /man/	who	[m]
nūn	ن	n	ناس /nās/	people	[n]
hā’	ه	h	هذا /hāḏā/	this	[h]
wāw	و	w	وقت /waqt/	time	[w]
yā’	ي	y	اليمن /alyaman/	Yemen	[j]

Trandcription sign	IPA sign
g	[g]
ž	[ʒ]
ǵ	[dz]
ǧ	[dʒ]

Table A.2: Additional Sign Used in Arabic Dialects Used in Transcription (Versteegh, K. 2014).

Vowel		Arabic letter	Arabic Example	Meaning	IPA sign
Short	/a/	فتحة	نَحن /nahnu/	we	[a]
	/i/	كسرة	مِن /min/	of, from	[i]
	/u/	ضمة	غُرفة /ġurfa/	room	[u]
Long	/ā/	ا	باب /bāb/	door	[a:]
	/ū/	و	صابون /ṣābūn/	soap	[u:]
	/ī/	ي	في /fī/	in, at	[i:]

Table A.3: Arabic Vowels (Jabbari 2013).

Diphthong	Arabic Example	Meaning
/aw/	يَوم /yawm/	day
/ay/	ضَيف /ḍayf/	guest

Table A.4: Arabic Diphthongs (Jabbari 2013).

Appendix B

Arabic Dialects

B.1 Maghrebi Dialect

Maghrebi (Western or North African) Arabic is the colloquial language spoken in the Maghreb region. Speakers of Maghrebi Arabic call their dialect Derja, Derija, or Darija ("الدارِجَة"). It includes the dialects of Morocco, Algeria, Tunisia, Libya, and Mauritania (Ḥassāniya) (Versteegh, K. 2014).

Historically, the origin of Maghreb populations are from Berber origin. The arabization process of this region is related to the Muslim expansion from the east. It took place in two waves: first in the 7^{th} century, and then in the 11^{th} century, which is a marked separation in time between both arabization processes compared to other areas of the Arabophone world.

From a linguistic perspective, Maghrebi dialects may be divided into two main groups: the Pre-Hilālī and Bedouin dialects. Pre-Hilālī dialect is called sedentary dialect. It is spoken in areas that are affected by the expansion of Islam in the second half of 7^{th} century. The only affected regions by the Arabic language are the urban centres and in some cases in newly established military camps, for that, the greater part of this region remained entirely Berber-speaking. Bedouin dialect is spoken in areas which are influenced by the Arab immigration in the 11^{th} century of Arab tribes. Banū Hilālī is the biggest tribe among them. During this process, the Arabic language reached the countryside and the nomadic areas of North Africa, but it never replaces the Berber language completely (Pereira 2011). Figure B.1 illustrates the Berber-speaking areas in North Africa.

The -Maghrebi dialects shared a certain common features (Pereira 2011, Sayahi 2014, Versteegh, K. 2014), such as:

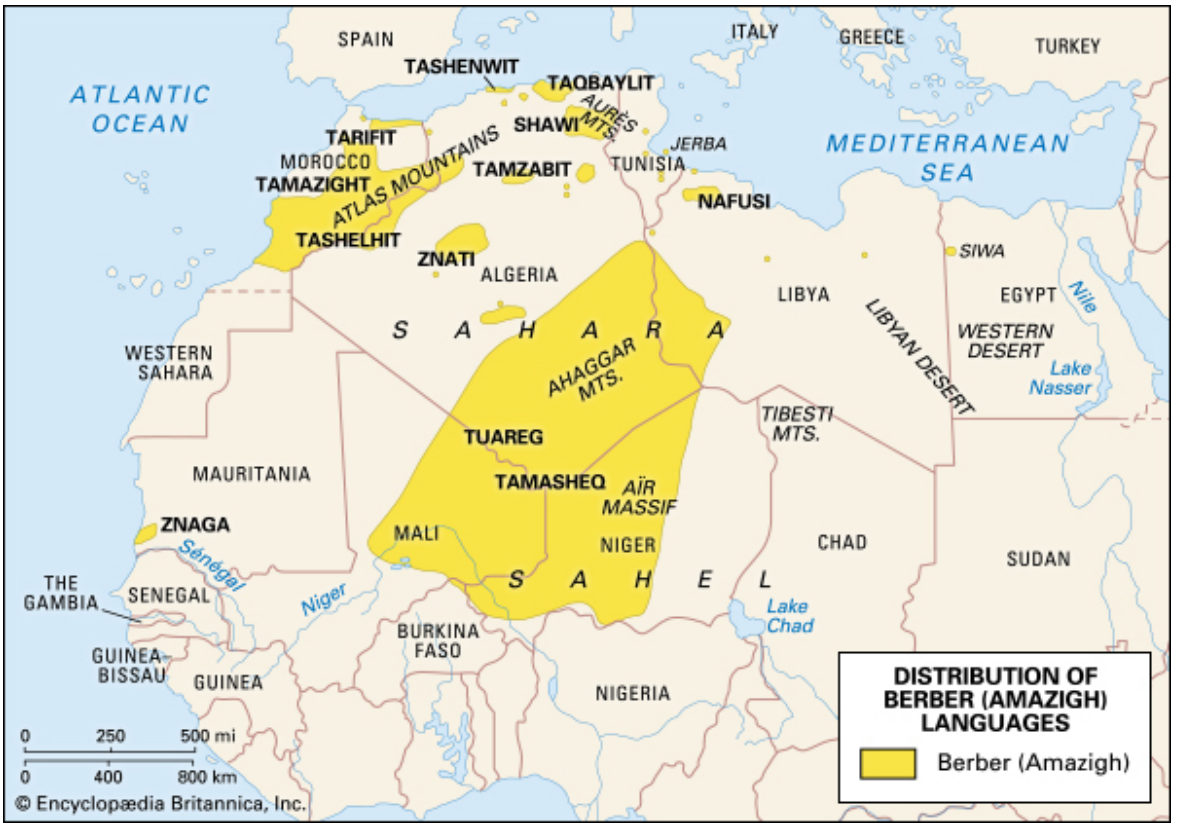


FIGURE B.1: Berber-speaking Areas in North Africa.

- A very simple vowel system: Only two short vowels /ə/ (merge /a/ and /i/) and /u/, and three long vowels: /ā/, /ī/, /ū/.
- Morphological feature in the verbal system: In imperfect verb, the prefix $n-$ for the 1^{st} person singular (e.g. نلعب /nəlʕəb/ 'I play') and $n-$ plus suffix $-u$ for the plural (e.g. نلعبو /nəlʕbu/ 'we play').
- The stress: In words of the form فَعَل /fa'al/, the original primary stress in MSA was on the penultimate, which is contrary unstressed short vowel for these dialects, such as: žbəl ('mountain', MSA: جَبَل /jabal/), عْرَب 'rəb ('Arabs' MSA: عَرَب /'arab/),
- Syllable Structure: The type CVCC was changed to CCVC, where C is consonant and V is vowel , for instance: سْقَف /sqəf/ ('roof', MSA: سَقْف/saqf/)

Maghrebi Arabic is unintelligible by speakers from other regions in the Middle East because it is heavily influenced by the French and Berber languages (Zaidan & Callison-Burch 2014). It is affected by other languages that were introduced into the region, such as: Turkish, Spanish, Italian, and to a greater degree French (Sayahi 2014). For that, it uses frequently borrow

words from these languages and conjugate them according to the rules of their dialects. In addition, the French borrow words are used in Algeria, Morocco and Tunisia, the Spanish borrow words in Morocco, and the Italian borrow words in Libya and Tunisia. For instance, the Algerian dialect uses French borrowing words, such as: موتور /mūtūr/('moteur' in French, motor), لاطوسيون /laṭūsyūn/ ('la tension', blood pressure), شَارجَا /šārjā/('charge', to load), and قَارَا /gārā/ ('garer', to park) (Harrat et al. 2016).

B.2 Egyptian Dialect

Egyptian Arabic (or Masri) is the colloquial language spoken in Egypt, which is the largest population country in the Arab world. For that, its dialect is the most widely spoken Arabic dialect. In addition, it is the most widely understood dialect due to the influence of Egyptian television and cinema throughout the Arabic-speaking world (Haeri 2003, Thompson Irene 2016).

In the deserts of eastern Egypt, there are Arabic-speaking Bedouins in pre-Islamic times (Holes 2004). In the other region, the Egyptians speak Coptic before the Arab expansion of Egypt in the 7^{th} century. With the Arab expansions, Arabic progressively displaced Coptic. Its dialects have common features with the dialects of the Levant than with the dialects of the Maghreb (Khan et al. 2012). Some Egyptian Arabic keywords are: ازيك /'azayak/ ('how are you ?'), دلوقتي /dlwaqtī/ ('now'), and اتخمد /'iṯ̱amed/ ('sleep') (Al-Walaie & Khan 2017). Egyptian Arabic borrowings words from Greek, Italian, French, and English. At present, the primary source of borrowing is English. Some examples of borrowing are as follows: انترفيو /'interfiew/ ('interview'), جروب /grūb/ ('group'), جراش /garaš/ ('garage'), and ميزد كول /mizzd kaul/ ('missed call') (Thompson Irene 2016, Yacoub 2016).

The Egyptian Arabic can be divided to the following dialect (Versteegh, K. 2014):

- The dialects of the Delta; these are subdivided into two groups: the eastern Delta dialects in the Šarqiyya and the western Delta dialects.
- The dialect of Cairo.
- The Middle Egyptian dialects (from Gizeh to Asyūṭ).
- The Upper Egyptian dialects (from Asyūṭ to the south); these are subdivided into four groups: the dialects between Asyūṭ and Nag Hammadi; the dialects between Nag Hammadi and Qēna; the dialects between Qēna and Luxor; and the dialects between Luxor and Esna.

The Egyptian Arabic dialects shared certain common features, which distinguish its dialects from other dialect groups (Versteegh, K. 2014), such as:

- Short Vowel: Three short vowels of Classical Arabic, but /i/ and /u/ are elided in open and unstressed syllables.
- Long Vowels: there are five (/ā/, /ī/, /ū/, /ē/, /ō/), which are shortened in unstressed position.
- The position of the demonstratives: The demonstratives always occur in postposition. The demonstratives of Cairene are da, di, dōl, for instance: الراجل دا /ir-rāgil da/ (this man) and الفلاحين دول /il-fellaḥīn dōl/ (these peasants).
- The position of the interrogatives: The position of the interrogatives is remarkable at the end, while the most Arabic dialects front the interrogatives. For example, شوفت مين؟ /šufti mīn/ ('Whom did you see?').

B.3 Levantine Dialect

Levantine (or the Levant), al-mašriq in Arabic, Arabic dialect is spoken across a vast area covering Syria, Lebanon, Palestine, Jordan. The term 'Levant' meaning 'east, orient', it refers to the east region of the Mediterranean, or to the eastern part of the Arab world (Khan et al. 2012). In pre-Islamic times, the Arabic-speaking is presented in the Syrian desert and in some of the sedentary areas. The arabization of the Levantine area began during the first campaigns of the expansions in the 7^{th} century (Versteegh, K. 2014). Its dialects have distinctive features from the west or in the Arabian Peninsula (Khan et al. 2012). The borrowed words in Levantine Arabic dialect, are either borrowed from Persian and Turkish (neighboring languages) or influenced by European languages (English and French). Some examples of borrowing are: ريموت كونترول /rīmūt kuntrūl/ (remote control), باشا /bāšā/ (sir) is borrowed from Turkish, and بأشيش /ba'šīš/ (tip) is borrowed from Persian (Jabbari 2013).

The Levantine Arabic dialect can be distinguishes three groups (Versteegh, K. 2014):

- Lebanese and Central Syrian dialects: consisting of Lebanese (e.g., the dialect of Beirut) and Central Syrian (e.g., the dialect of Damascus and Druzes).
- North Syrian dialects (e.g., the dialect of Aleppo).

- Palestinian and Jordanian dialects: consisting of the Palestinian town dialects, the central Palestinian village dialects, and the south Palestinian/Jordanian dialects.

The Levantine Arabic dialects shared a certain common features (Jabbari 2013, Versteegh, K. 2014), such as:

- Long vowels: Presence of the three long vowels /ā/, /ī/, and /ū/.
- Initial two-consonant clusters are frequently formed in Levantine dialect, such as: بيَكتب /**by**akyub/ (He writes, MSA: يكتب /yaktubu/), منين /**mn**īn/ (Where ... from? , MSA: من اين /min'ayn/).
- Loss of gender distinction: In the second- and third-person plural of pronouns and verbs, for example:

MSA	**Levantine**	**Meaning**
كتبتم /katab**tum**/	كتبو /katabu:/	You (pl. mas.) wrote
كتبتن /katab**tunna**/	كتبو/katabu:/	You (pl. fem.) wrote
كتبوا /katab**u**/	كتبو/katabu:/	They (pl. mas.)
كتبن /katab**na**/	كتبو/katabu:/	They (pl. fem.)

B.4 Gulf Dialect

Gulf (or the Arabian peninsula) is the least-known dialect area of the Arab world. Old Arabic developed in the north of the Peninsula, which is the homeland of the Arab tribes (Versteegh, K. 2014). It includes the dialects of Saudi Arabia, Yemen, Oman, the United Arab Emirates, Bahrain, Qatar, and Kuwait (Khan et al. 2012). The borrowed words from various languages throughout different historical periods, for example: بشتك /bištak/ (cloak) is borrowed from Fares, التجوري /altijurī/ (storage) is borrowed from India, دلاغات /dlaġāt/ (socks) is borrowed from Turkey, and ابجورات /'abajūrāt/ (Rolling shutters) is borrowed from France (Abuata & Al-Omari 2015).

Linguistically, the Peninsula can be divided into four groups (Palva 1991):

- North-east Arabian dialects: these are the dialects of the Najd, in particular, those of the large tribes 'Aniza and Šammar. This group is divided into three subgroups. First, *the 'Anazī dialects* include the dialects of Kuwait, Bahrain and the Gulf states.

Second, the *Šammar dialects* include some of the Bedouin dialects in Iraq. Finally, the *Syro-Mesopotamian Bedouin dialects* include the Bedouin dialects of North Israel and Jordan, and the dialect of the Dawāġrah.

- South(-west) Arabian dialects: refers to the dialects of Yemen, Hadramaut and Aden, and the Shi'ite Baḥārna in Bahrain.
- Ḥijāzī (West Arabian) dialects: to this group belong the Bedouin dialects of the Ḥijāz and the Tihāma.
- North-west Arabian dialects: It comprises the dialects of southern Jordan, the eastern coast of the Gulf of 'Aqaba, and some regions in north-western Saudi Arabia.

B.5 Iraqi Dialect

The Mesopotamian (or Iraqi) dialect encompass the Arabic dialects spoken in Iraq, north-eastern Syria, south-eastern Turkey, and Iranian Khuzestan (Khan et al. 2012). The Arabicisation of this area, it took place in two stages. First, the Arab expansions of urban varieties sprang up around the military centres, such as Baṣra and Kūfa. Later, the migration of Bedouin tribes from the peninsula are affected the same area of the first layer of urban dialects (Versteegh, K. 2014).

It includes two major dialects: qəltu and gilit dialects, the names are deriving from the equivalent for the MSA word قلت /qultu/ ('I said') in the two dialect types. Qəltu dialect is spoken by non-Muslims and sedentary Muslims in the northern part of the area. While, gilit dialect is spoken by the Muslims of Bedouin origin and of the southern cities. Table B.1 reports the distribution of these dialects (Khan et al. 2012).

	Muslims		non-Muslims
	non-sedentary	sedentary	
Lower Iraq	gilit	gilit	qəltu
upper Iraq	gilit	qəltu	qəltu
Anatolia	gilit	qəltu	qəltu

Table B.1: Geographical Distribution of Gilit and Qəltu Dialects in Iraq.

It used a borrowed words from Persian, Turkish, or English. These words are also found in the Gulf but not always with the same meanings. For example some borrowed word in Mesopotamian dialect such as: bazzūna 'cat', xōš 'good', and temman 'rice' Postgate (2007).

Appendix C

Acoustic Features

Acoustic Features are one of the most primitive information which can be extracted directly from the raw speech when analyzed in physical terms, e.g. fundamental frequency, amplitude, harmonic structure (Crystal 2008, Wong 2004). The acoustic information are typically got by extracting the spectral features from segments of speech Navrátil (2006). The most widely used acoustic features are Mel Frequency Cepstral Coefficient (MFCC) (Davis & Mermelstein 1980), Perceptual Linear Prediction (PLP) and Linear Prediction Cepstral Coefficient (LPCC).

MFCC is determined from speech using the following steps Rao et al. (2015):

1. pre-emphasize the speech signal, which refers to filtering that emphasizes the higher frequencies.
2. Divide the speech signal into sequence of frames with a frame size of 20 ms, because the speech signal is a slowly time-varying or quasi-stationary signal, and a shift of 10 ms. Then, apply the Hamming window over each of the frames to enhance the harmonics, smooth the edges and to reduce the edge effect while taking the Discrete Fourier Transform (DFT) from the signal.
3. Compute magnitude spectrum for each windowed frame by applying DFT.
4. Mel-Spectrum is computed by passing the Fourier transformed signal through a set of band-pass filters known as mel-filter bank.
5. Discrete Cosine Transform (DCT): Since the vocal tract is smooth, the energy levels in adjacent bands tend to be correlated. The DCT is applied to the transformed mel frequency coefficients then it produces a set of cepstral coefficients.