Corpus of the Algerian dialect CDalg: linguistic rules of decision

Abdel Hakim Benali [1], Mohamed Hédi Maaloul [2], Lamia Hadrich Belguith[1]

[1] Faculty of Economics and Management of Sfax

[2] Hail University, Kingdom of Saudi Arabia
a.hakim.benali,mohamedmaaloul,lamia.belguith@gmail.com

Abstract. In this article, we present a method for the automatic detection of the Algerian dialect, basing on the data dictionary and the morphology of the word Thus, we intend to present in this article, and in the first place, the method adopted for the collection of data. Second, we will aim to explore the detection rules extracted from the data corpus. Thirdly, we will explain our method of identifying the Algerian dialect

Keywords: Algerian dialect, Collection, Transcription, dialect identification, decision rules.

Resumé. Dans cet article, nous présentons une méthode de détection automatique du dialecte algérien, basée sur le dictionnaire de données et la morphologie du mot. Nous avons donc l'intention de présenter dans cet article, et en premier lieu, la méthode adoptée pour la collections des données. Deuxièmement, nous aimerons explorer les règles de détection du corpus de données. Troisièmement, nous expliquerons notre méthode d'identification du dialecte algérien

Mots-clés: Dialecte algérien, collection donnée, transcription, identification de dialecte, règles des décisions.

1 Introduction

The identification of any spoken language is instantaneous. When the speaker speaks, we will identify the language used without difficulty. This strategy, although at first glance easy, is of particular interest to us. Our ambition is to adapt, exploit and share it with automatic techniques [1].

Language detection automation is one of the most interesting areas for TAL researchers. The researchers have launched a revolution in the detection of languages, they don't stop here and they embark on the automatic recognition of dialects.

The Arabic language has recently seen some work on automatic recognition. However, the speech processing of Arabic dialects (AD) is confronted with many challenges [2]. On the one hand, the oral translation of the Arabic dialect and the absence of orthographic norms give rise to the difficulty of its automatic processing. On the other hand, a number of dialects with linguistic differences on phonological, morphological, syntactic and lexical levels amplify the challenges of constructing tools and resources for all Arabic dialects [3].
We quote the work of Alfardy et al. [4]. A work based, for the automatic identification and glossing of the Arabic dialect, on a corpus of Egyptian dialect, we also quote the work of Saadane et al. Who proposed a hybrid method of automatic identification of dialectal origin for written Arabic language based on data dictionaries and a statistical method [5].
Among the Arabic dialects, we find the Algerian dialect; this dialect is characterized by the multitude of Arabic sub-dialect and the non-Arabic, which we call the Amazigh dialect.
The diversity of dialects in Algeria goes back to the colonial era and the large area of the country and the sharing of lexicon with other neighboring countries. According to Bougrine et al [6], the Algerian Arabic dialect is composed of seven arabic sub-dialects (Saharan, Tellian, High-plains, Ma'quelian, Sulaymite, Algerian-Blanks, Sahel-Tell).
The objective of this article is to present our corpus CDalg and to propose a method for the detection of the Algerian dialect based on the data dictionary and the morphology of the word also the confirmation of the decision by rules of correction.
This article is divided into three parts; the first part will expose the linguistic corpus. The second part will present the rules of decision and correction extracted after an in-depth study of the corpus. The third part will focus on the proposed method for the detection of the Algerian dialect.

2 Presentation of the CDalg corpus

For the collections of oral utterances, two methodologies were used, the first one consists of the direct recording of the speakers around various subjects as the case with. Graja et al in the TudiCol [7]. and the second is the extraction of the audios from channel radio and TV online (streaming) and from youtube's videos as the example of Zribi et al in the corpus STAC corpus [8],

To better situate ourselves in this context, we have chosen to build our linguistic corpus CDalg while taking into account that there is currently little corpus dedicated to Algerian dialectal Arabic varieties. We recall to this effect some: ALGASD [9], ALG-DARIDJAH [10], AMCASC [11], KalamDZ [5] and PADIC. In this last, there are three dialects of the Maghreb, two from Algeria, and one from Tunisia and two from the Middle East who are concerned by this study. [12]

The second stage of construction of the oral corpus is the orthographic transcription, for that we chose to use the praat tool because it is free software [13] and tested on STAC corpus [8] also it provide the segmentation of the speech.

As we have seen previously, the Algerian dialect is varied to several sub-dialects so we decided to take the sub-dialect (Algerian-blanks which is the dialect of the capital Algiers) as a reference of the Algerian dialect; we can justifier our choice by the multitude of Algerian dialects and the high number of practitioners of the Algerier-Blanks dialect The current size of our corpus is more than 150 thousand utterances of different dialects with a duration of 2 hours and 09 minutes of recording for the Algerian dialect, these utterances have undergone a segmentation step in sentences, this segmentation has given an average of 4211 sentences, the average of each sentence is 7 words.
The table below gives the details.

Dialect type	MSA	ALG	TUN	PAL	SYR	MAR
Number of words	35983	38965	34369	38281	38547	39568
Heure d'enregistrement/ hour	-	2,09	1,17	2,06	1,24	2,11

Table 1 Current size of the CDalg corpus

3 Study of the corpus and identification of decision rules

After the analysis of our working corpus, we have been able to draw some rules of distinction from the Algerian dialect of the remains of the dialects treated. These rules are formed of linguistic signals and observed heuristics, which are mainly independent markers. They have important values distinguishing them from the DZ.

These rules can be divided into two types: linguistic indices and morphological labels.

3.1 Decision rules based on linguistic indices

These indices are extracted generally from the data dictionary

Dz	Buckwalter	Arabe	English
يا خو	yA xw	أخي	My brother
بزاف	bzAf	كثيرا	A lot
قاع	vAç	كل	All
دجوز	djwoz	أدخل	Enter
شحال	$HAl	كم	How many

Table 2 example of linguistic index according to the dictionary of the Algerian dialect

يا خو عندي مشاكل بزاف في راسي

yA xw Endy m$Akl bzAf fy rAsy

Brother, I have many problems in my head

In the example above, the clues that identify the Algerian dialect according to the data dictionary are: ((brother) [xw] "خو") and ((much) [bzAf] "بزاف"). So we can conclude that this sentence is of the Algerian dialect type.

3.2 Decision rules based on morphological labels

The question with the "What"
Indice [واش / What] + Conjugated verb
واش كليت؟ / What did you eat ?/ wAšklyt?

واش راك خو لاباس؟ عندي بزاف ماشفتكش

wA$ rAk xw lAbAs ? Endy bzAf mAftk

How are you my brother? I haven't seen you for a long time.

In the example above, we can distinguish a linguistic clue, it is the clue of negation "ش" ,"ما", and also a morphological label, the pronoun What + a conjugated verb, " واش راك".

An order
The verb "to go" to the present (go) "أمشي" + a verb to the imperative => replace the first A "أ" with a T "ت".

Verb	Verb to the imperative	Go + Verb to the imperative	Buckwalter
أكل	أكل	أمشي تكل	Âmšytkl
لعب	ألعب	أمشي تلعب	Âmšytlϛb

Table 3 an example of morphological label based decision rules for an order

3.3 Correction rules

Our analysis of the corpus allowed us to identify different compositions of the decision correction rules. These can be applied during conflict in the decision rules.
A Algorithme 1 présente, ainsi, un exemple d'une règle de correction de type étiquette-indice.

Sample algorithm of a correction rule of the label-clue type

1: ***Let*** *context C: a sentence*
  Let *X: a morphological label*
  Let *Y: a linguistic clue*
2: ***if*** *(X detects DZ* ***AND*** *detects Other)* ***then***
3: ***if*** *(in C* ***there is*** *a Y detected as DZ)* ***then***
4: *C: is a sentence of type DZ*
5: ***else*** *C: is an ambiguous sentence*

Algorithm 1 Example of a label-index correction rule

أمينة بغيت نسقسيك, واش قريتي لبارح فلمسيد ؟

Omynp bgyt nsqsyk, wA$ qryty lbArH flmsyd ?

Amina, I wanted to ask you: what did you study yesterday at school?

- *Rule based on linguistic clue:* بغيت, مسيد
- *Rule based on morphological labels:* واش قريتي

In this case, it may be that "بغيت" is a specific a linguistic clue for the Algerian dialect and for other dialects. In this case, we see the verb after it, and as "نسقسيك" belongs to the rules based on morphological labels for the Algerian dialect and " واش قريتي " is a morphological labels for DZ, we easily conclude that this sentence relates to DZ.

4 Proposed method

Our proposed method for the identification of the Algerian dialect is based on the dictionary of Algerian dialect and also on the rules of distinction extracted from the study corpus.

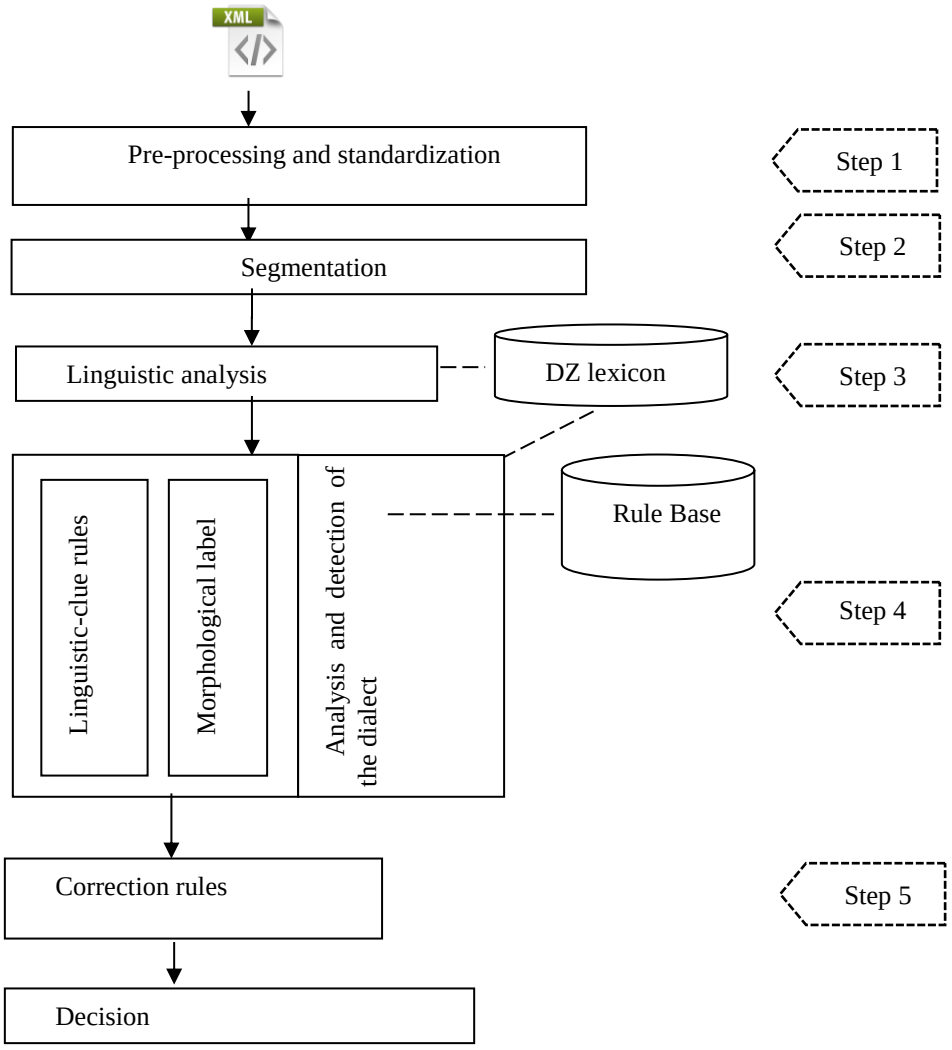

Figure 1 proposed method

5 System realization

The realization of our system followed several steps, the first one was conceived for the proposition of a tool for the textual segmentation STGAlg which unscrews the text in textual unit then the second stage was conceived for the proposal of an automatic standardization tool following the orthographic standardization convention of the Algerian dialect ALG-CODA, then this step was the adaptation of the Stanford parser for the Algerian dialect.

Our system has been realized based on a corpus of small size and as perspective; we aim to evaluate it on a large corpus

6 Conclusion

In this article, we have proposed a method for the automatic identification of Algerian dialect. This method is based essentially on the linguistic rules of the Algerian language. First, we created a corpus of the Algerian dialect. Second, we presented some examples of decision rules. Our future work is oriented towards the evaluation of our proposed method on a large corpus, and also moves towards the constitution of a tool for the morphological and linguistic analysis of the Algerian dialect. Despite the importance of the rules of correction, they remain, from time to time, ineffective for the choice of a decision in case of contradiction between the rules based on linguistic index and rules based on morphological labels. For this very reason, we plan to supplement this method with a numerical step in order to resolve the discrepancy between the results and also to settle the cases of no decision.

Références

1. Martine Adda-Decker, Automatic Language Identification, Spoken Language Processing, JJ Mariani ed., Wiley-ISTE, 2009
2. Houda Saadane, Hosni Seffih, Christian Fluhr, Khalid Choukri, Nasredine Semmar: Automatic Identification of Maghreb Dialects Using a Dictionary-Based Approach. LREC2018
3. Ines Zribi, Traitement automatique du dialecte tunisien : construction de ressources linguistiques, Thèse doctorat à l'université de Sfax, 2016.

4. Saâdane, H.; Nouvel, D.; Seffih, H. & Fluhr, C.Une approche linguistique pour la détection des dialectes arabes Actes de TALN 2017, volume 2 : articles courts, 2017
5. S. BOUGRINE, H. CHERROUN, D. ZIADI, A. LAKHDARI, A. CHORANA, Toward a Rich Arabic Speech Parallel Corpus for Algerian, 2017
6. Sub-Dialects, The 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media, page 2-10,2016.
7. M. Graja, M. Jaoua and L.H Belguith. "Discriminative Framework for Spoken Tunisian Dialect Understanding". International Conference on Statistical Language and Speech Processing (SLSP 2013), Tarragona Spain, July 29-31.2013.
8. Ines Zribi, Mariem Ellouze, Lamia Hadrich Belguith, and Philippe Blache, Spoken Tunisian Arabic Corpus "STAC ": Transcription and Annotation, pp. 123–135; rec. 2015-01-25; acc. 2015-02-27; Research in Computing Science 90 (2015).
9. Droua-hamdani, G.; Selouani, S. A. & Malika, B., Algerian Arabic speech database, (ALGASD): corpus design and automatic speech recognition application, Arabian Journal for Science and Engineering, 2010, 35, 158
10. Soumia BOUGRINE, Hadda CHERROUN, Djelloul ZIADI, Abdallah LAKHDARI, Aicha CHORANA, Toward a Rich Arabic Speech Parallel Corpus for Algerian Sub-Dialects. The 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media, page 2-10, 2016.
11. Mourad Djellab, Abderrahmane, Amrouche, Ahmed bouridane and, Noureddine Mechallegue, Algerian Modern Colloquial Arbic Speech Corpus (AMCASC): regional accents recognition within complex socio-linguistic environments, lang resource & evaluation, 2016.
12. Meftouh, K.; Harrat, S.; Jamoussi, S.; Abbas, M. & Smaili, K. Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus. The 29th Pacific Asia Conference on Language, Information and Computation, 2015.
13. Thierry Bazillon. Transcription et traitement manuel de la parole spontanée pour sa reconnaissance automatique. PhD thesis, Université du Maine, 2011.