Morphological disambiguation of Tunisian dialect

Inès Zribi, Mariem Ellouze, Lamia Hadrich Belguith, Philippe Blache

▶ **To cite this version:**

Inès Zribi, Mariem Ellouze, Lamia Hadrich Belguith, Philippe Blache. Morphological disambiguation of Tunisian dialect. Journal of King Saud University - Computer and Information Sciences, 2017, 29 (2), pp.147-155. 10.1016/j.jksuci.2017.01.004 . hal-02869843

HAL Id: hal-02869843
https://hal.science/hal-02869843

Submitted on 30 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Journal of King Saud University – Computer and Information Sciences 29 (2017) 147–155

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Morphological disambiguation of Tunisian dialect

Inès Zribi [a,*], Mariem Ellouze [a], Lamia Hadrich Belguith [a], Philippe Blache [b]

[a] ANLP Research group, MIRACL, University of Sfax, Tunisia
[b] LPL, CNRS, Aix-Marseille University, France

ARTICLE INFO

Article history:
Received 14 June 2016
Revised 9 January 2017
Accepted 17 January 2017
Available online 29 January 2017

Keywords:
Tunisian dialect
Spoken language
Morphological analysis
Morphological disambiguation

ABSTRACT

In this paper, we propose a method to disambiguate the output of a morphological analyzer of the Tunisian dialect. We test three machine-learning techniques that classify the morphological analysis of each word token into two classes: *true* and *false*. The class label is assigned to each analysis according to the context of the corresponding word in a sentence. In failure cases, we combine the results of the proposed techniques with a bigram classifier to choose only one analysis for a given word. We disambiguate the result of the morphological analyzer of the Tunisian Dialect *Al-Khalil-TUN* (Zribi et al., 2013b). We use the Spoken Tunisian Arabic Corpus *STAC* (Zribi et al., 2015) to train and test our method. The evaluation shows that the proposed method has achieved an accuracy performance of 87.32%.
© 2017 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Morphological analysis (MA) is a crucial stage in a variety of natural language processing (NLP) applications (information retrieval, question answering, etc.). The analysis of languages with complex and rich morphology handicaps the performance of these applications due to the large number of analyses produced for each word independently of the context in which the word occurs. Therefore, a morphological disambiguation module is required.

Morphological disambiguation (MD) (also called Part-of-Speech (POS) tagging) consists in determining one correct POS tag among a set of POS tags that are assigned to a word, by taking into account the word's context.

In the literature, many techniques/systems have been developed for POS tagging modern standard Arabic (MSA). They follow two principal approaches to developing a tagger: a handcrafted rule-based approach, and a statistical approach. The handcrafted rule-based approach may be a practicable solution, but it requires a considerable investment of human effort. The most referenced works are done by Al-Taani and Al-Rub (2009) and Tlili-Guiassa (2006). Statistical approaches prove to be able to learn tagging from tagged data on the basis of a sufficient quantity of tagged documents. The most referenced works are carried out by Diab et al. (2004), Habash and Rambow (2005), Khoja (2001).

Contrariwise, Arabic dialects have not received much attention due to the scarcity of resources (corpus and lexicon) and tools (morphological analyzers, tokenizers, etc.). In addition, Arabic dialects are a spoken variety. Tagging a spoken language is typically harder than tagging a written one, due to the effect of disfluencies, incomplete sentences, etc. (Duh and Kirchhoff, 2005).

In this paper, we present the Tunisian Arabic Morphological DisAmbiguation System (TAMDAS). This system uses the output of a Tunisian Dialect (TD) morphological analyzer (Al-Khalil-TUN) (Zribi et al., 2013b) and a TD corpus (the STAC corpus) (Zribi et al., 2015), to morphologically disambiguate TD annotated transcriptions.

TAMDAS tests three different classifiers and combines their results with a bigram module in failure cases. We build a classifier based on feature vectors, which are generated from the morphologically annotated corpus, and then use it to classify the possible analyses of each word into *correct* and *false* classes.

This paper has seven main sections. Section 2 presents an overview of previous works that studied TD and the POS tagging of dialectal Arabic. Section 3 presents the characteristics of TD. In Section 4, we present the challenge of tagging a spoken language, especially in the case of TD. In Section 5, we describe the TD resources, then, in Section 6, we present our method. Finally, we give the results of the system evaluation, and discuss some errors.

* Corresponding author.
E-mail addresses: ineszribi@gmail.com (I. Zribi), Mariem.ellouze@planet.tn (M. Ellouze), l.belguith@fsegs.rnu.tn (L.H. Belguith), philippe.blache@lpl-aix.fr (P. Blache).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2017.01.004
1319-1578/© 2017 The Authors. Production and hosting by Elsevier B.V. on behalf of King Saud University.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

2. Related works

Some studies have been conducted in the field of Dialectal Arabic (DA) processing with a variety of approaches and at different degrees of linguistic depth. Most of the approaches tend to develop dialectal data (Al-Badrashiny et al., 2014; Al-Shargi et al., 2016; Khalifa et al., 2016; Maamouri et al., 2014; Samih and Maier, 2016), and tools (Darwish, 2014; Habash et al., 2012b; Habash and Rambow, 2006; Salloum and Habash, 2014, 2011) to treat a specific Arabic dialect. The most referenced works are carried out by Habash et al. (2013, 2005), Habash and Rambow (2005), Rambow et al. (2006).

In fact, few research studies treated the POS tagging task of Arabic dialects. Most of them dealt with Levantine and Egyptian Arabic. They treated these dialects as written varieties of Arabic languages (no characteristic of speech are considered). However, the automatic processing of Tunisian Dialect (TD) and its spoken varieties has not received much attention.

The DA POS tagging techniques follow two principal approaches. The ***first approach*** suggests using MSA resources and a few DA resources to create a POS tagger. In this context, (Duh and Kirchhoff, 2005) used the Buckwalter Morphological Analyzer (Buckwalter, 2004) designed for MSA, the LDC MSA Treebank corpus and some dialectal resources (the CallHome Egyptian Colloquial Arabic corpus, the LDC Levantine Arabic corpus) in combination with unsupervised learning algorithms in order to develop a POS tagger for Egyptian Arabic. The authors proposed to bootstrap the HMM tagger using POS information from the morphological analyzer. They improved the tagger by integrating additional data from other dialects (Duh and Kirchhoff, 2005). They reported a POS accuracy of 70.9%. Likewise, Rambow et al. (2006) explored MSA data and resources to develop a POS tagger for Levantine dialect. They adapted an MSA POS-tagger for Levantine data. They suggested that leveraging the existing resources is a viable option. Rambow et al. (2006) developed a bilingual small lexicon MSA/Levantine dialect. Combining information from this lexicon and a parameter renormalization strategy based on minimal linguistic knowledge, Rambow et al. (2006) noticed the biggest improvement in the tagger. Moreover, Habash et al. (2013) developed a morphological analysis and disambiguation for Egyptian Arabic based on an existing tool for MSA (the MADA tool, Habash and Rambow (2005) and Roth et al. (2008). MADA uses an existing morphological analyzer of MSA and applies a set of models (support vector machines and N-gram language models) to produce a per word in-context prediction. A ranking component computes the scores of the analysis produced by the morphological analyzer using a tuned weighted sum of matches with the predicted features (Habash et al., 2013). The top-scoring analysis is chosen as the best prediction of the tool (Habash et al., 2013).

The ***second approach*** of POS tagging DA intends to start from scratch. No MSA resources are used in this approach. Al-Sabbagh and Girju (2012) implemented Brill's Transformation-Based tagging algorithm (Brill, 1994) for the task of POS tagging Egyptian Arabic. For training, they used the manually annotated Twitter-based corpus. They reported an 87.6% accuracy on POS tagging.

Only two studies dealt with the POS tagging of the Tunisian dialect. They adopted the first approach. Boujelbane et al. (2014)) have retrained the MSA Stanford POS tagger (Toutanova and Manning, 2000). To retrain their system, they used a corpus derived from a translation of the MSA Treebank into TD. An accuracy of 78.5% in POS tagging of Tunisian transcribed texts was reported. Hamdi et al. (2015) proposed three steps for POS tagging TD. Their method is based on MSA resources. They convert a TD sentence into a MSA lattice, which is disambiguated to produce MSA target sentences. Finally, the MSA tagger assigns to each word its POS tag. This system achieved an accuracy of 89%.

3. Tunisian dialect

The Tunisian Dialect (TD) is the dialect of the Arabic language spoken in Tunisia. It is considered as a low variety given that it is neither codified nor standardized, even though it is the mother tongue daily spoken by everyone (Saidi, 2007). The regional varieties of TD are the Tunis dialect (Capital), the Sahel dialect, the Sfax dialect, the Northwestern Tunisian dialect, the Southwestern Tunisian dialect and the South-Eastern Tunisian dialect (Gibson, 1998; Talmoudi, 1980).

There are a lot of different and similar points between TD and MSA (Zribi et al., 2013a). In order to compare these two varieties of Arabic language, we focus on four levels: namely the phonological level, the morphological level, the lexical level and the syntactic level.

3.1. The phonological level

The vocalic system of TD is reduced (Tilmatine, 1999). Some short vowels are not overtly rendered, especially if they are located at the end of the word (Mejri et al., 2009). The MSA verb [1]شَرِبَ <šariba> /šariba/[2] 'he drank' is pronounced شْرِبْ /šrib/ (note the deletion of the vowels located at the beginning and the end of the verb). Moreover, TD has a long vowel/e:/ which does not exist in MSA (Zribi et al., 2014).

The consonant system also includes some phonetic differences (Mejri et al., 2009). In some cases, the Arabic consonant ق <q> /q/ is pronounced /g/. The MSA word بَقَرَة <baqaraħ> /baqara/ 'cow' is pronounced in TD /bagra/. In addition, some consonants in TD have multiple pronunciations. For example, the consonant غ <γ> /γ/ and ج <j> /j/ can also be pronounced respectively as /x/ and /z/.

3.2. The morphological level

The main difference between MSA and TD is on the affix level. We can notice the presence of new dialectal affixes and the deletion of others. Dual suffixes ان <An> and ين <yn> are generally absent. They are replaced by the numeral زوز <zwz> 'two' located after or before the plural form of the noun. However, some words in TD can be agglutinated to the suffix ين <yn> to express duality. In verb conjugation, TD is characterized by the absence of the dual (feminine and masculine) and the feminine in the plural. It has witnessed many simplifications in its affixation system (Ouerhani, 2009). Indeed, new affixes appeared. The first one is the negation clitic ش <š>. It is agglutinated to the end of the verb that must be preceded by the negation particle ما <mA>(*e.g.*, ما كليتش <mA klyt-š> 'I don't eat') (Mejri et al., 2009). The interrogation prefix of MSA أ <Â> is transformed in TD into the suffix شي <-šy> (*e.g.*, خرجشي, <xrj-šy>, 'Did he go out?'). Likewise, the future prefix سـ <s-> is replaced by the particle باش <bAš> 'will'. In addition, we note the absence of the dual clitics in TD.

3.3. The lexical level

Historical events have made the linguistic situation in Tunisia rather complex. The prolonged Ottoman Turkish political domination of North Africa roughly from the mid-fifteenth to the late nineteenth century and the French colonization from 1830 had an impact on the absorption of foreign vocabulary into the lexicon of local Arabic dialects (Holes, 2004).

In addition to Turkish and French, we find many examples of European language lexical elements in TD. We can identify a signif-

[1] We follow the CODA-TUN convention (Zribi et al., 2014) when writing examples of words in TD.

[2] Transliteration is coded following Buckwalter transliteration. For more details about it, see (Habash et al., 2007).

icant number of Spanish words and Italian, even Maltese words. The Arabic spoken dialects in Tunisia contain many foreign items of vocabulary such as: قطوس <qTws> 'cat', of Maltese origin; كوجينة <kwjynħ> 'kitchen', براكة <brAkħ> 'shed', فيشطة <fyšTħ> 'party', of Italian origin; and بوسطة <bwsTħ> 'post office', بلاصة <blASħ> 'place', باكو <bakw> 'package', all derived from French.

Borrowings are frequent in Tunisian dialects. They have been fully morphonologically assimilated to Arabic phonology or to Berber phonology. For instance, بريكية <brykyħ> 'lighter', ترينو <trynw> 'train' and كروسة <krwsħ> 'carriage' are derived respectively from the French words "*briquet*", "*train*" and "*carrosse*".

Code switching between Arabic and French changes the TD lexicon (Ouerhani, 2009). It allows the introduction of new words (nouns and verbs) derived from foreign languages. Tunisians easily and frequently switch between MSA, TD and the French language in a conversation (Zribi et al., 2013a). We can cite as examples: "*ça va*?" 'Okay?', "*désolé*" 'sorry', "*rendez-vous*" 'appointment', etc. All these expressions and words are used without being adapted phonologically.

3.4. The syntactic level

The principal syntactic differences between MSA and TD are minor. The MSA word order is generally VSO (Verb Subject Object) especially in verbal sentences, whereas, in TD, the preferred word order is SVO (Mahfoudhi, 2002). The VSO and VOS orders are also used in TD.

4. Challenges in tagging Tunisian dialect

4.1. Absence of TD resources

The POS tagging task is an important step for many NLP applications. To be successful, many resources and tools should be used. Among the most used ones, we can cite a big annotated corpus and a lexicon. The creation and presence of such resources for spoken languages, such as TD, which is an under-resourced language, represent the most challenging precondition for POS tagging. TD has neither a standard orthography nor large collections of written text. The few TD resources created over the last five years are still in their infancy. The size of annotated corpora is relatively small compared to MSA.

4.2. Detection of sentence boundaries

In written languages, the sentence boundaries are easily delimited. A sentence begins with a capital letter (especially in the Indo-European languages) when a simple dot or even the discourse markers indicate the end of the sentence (Dister et al., 2009). In spoken language, there is none of these phenomena. Furthermore, incomplete sentences, dialogue, conversation, overlapping statements, disfluencies, etc. make the definition of sentence boundaries in speech very difficult. In addition, these phenomena affect the syntactic structure of sentences in speech conversations.

4.3. Lexical particularities of speech

The transcribed speech corpus contains many para-linguistic elements, such as onomatopoeia, truncated words, laughter marks, mouth sounds, breathing sounds, etc. These elements should be treated in a specific way.

4.4. Non-canonical syntactic structures in TD

Sentences in speech do not follow in certain cases well-formed, canonical syntactic structures due to their spontaneous character and some characteristics of TD. Indeed, TD is a spoken form of Arabic from which Tunisians easily and frequently switch to MSA and French.

Let us take the simple English sentence: 'Did you succeed the exams, Mohamed?'. This sentence can be translated into the following phrase: Ça va الامتحان محمد?, <Ça va AlAmtiHAn mHmd ?>. The translated phrase is composed of the French phrase "*ça va*" followed by two TD words.

The Part-of-Speech tagging of this sentence, which is very frequent in daily speech, is very difficult. To conclude, the presence of many foreign words in TD speech and code switching phenomena increase the difficulties of automatic tagging of TD.

4.5. Ambiguity

The lack of short vowels causes morphosyntactic ambiguity for MSA (Habash, 2010). TD shares this characteristic with MSA. Even with the presence of short vowels, morphosyntactic ambiguity occurs. In many cases, a TD word has different morphological analyse that share the same POS tag, but their root or gender and number are different. For instance, the TD word خرجت <xrjt> is an ambiguous verb form. It has two possible diacritization forms. The first one is خَرجِت <x**a**r**j**i**t**> 'she went out' which is the third person feminine singular verb in the past tense. The second form is خرَجِت <xr**aj**i**t**>. It can be the second person (feminine or masculine) singular verb in the past tense 'you went out', or the first person singular in the past tense 'I went out'.

5. Tunisian dialect resources

5.1. Tunisian corpus

5.1.1. STAC presentation

To train and test the performance of our method, we used the STAC (Spoken Tunisian Arabic Corpus) corpus (Zribi et al., 2015). STAC is a speech corpus that contains additional information other than text. It contains multiple annotations such as sentence boundaries, disfluencies, named entities, etc. The STAC corpus consists of 4 hours and 50 minutes of speech (some radio and TV broadcasts and conversations recorded in railway station) that are recorded and manually transcribed using the transcription tool Praat[3]. The corpus relates to various fields: politics, health, social and religious issues, and others.

Transcribing and annotating STAC are based on the annotation conventions of OTTA (Zribi et al., 2013a) in conjunction with rules defined by the CODA-TUN (Zribi et al., 2014), an extension of the convention CODA (Conventional Orthography for Dialectal Arabic, Habash et al., 2012a), to TD. CODA is designed to develop computational models of Arabic dialects. First, it is defined for Egyptian Arabic and then extended to other Arabic dialects, such as Algerian (Saadane and Habash, 2015) and Palestinian (Jarrar et al., 2014).

The STAC corpus is composed of about 42388 words. STAC was analyzed with the Al-Khalil-TUN morphological analyzer (Zribi et al., 2013b) and a unique correct analysis is marked for each word in a sentence. The choice of a correct analysis is performed by an expert (Zribi, 2016). The annotation provided by STAC is used as a gold standard to compare the results of the different developed systems and assess their accuracy.

5.1.2. Data preparation

Preparing the training data is an important task for any classification task. These raw data cannot be used without a prior work of segmentation and annotation. The STAC corpus is a spoken transcribed corpus that incorporates the transcription of many conver-

[3] http://www.fon.hum.uva.nl/praat/.

sations between two and sometimes more than two people, and one monologue. In spontaneous speech, there are various types of sentences from which we distinguish the following four ones:

- S1: a sentence is started by a speaker and completed by another.
- S2: a sentence is started and not completed (incomplete sentence).
- S3: a well formed sentence (the speaker starts and completes the sentence).
- S4: a sentence contains disfluencies (hesitations, repetitions, onomatopoeia and other phenomena related to spontaneous speech).

Table 1 presents some examples of sentences. We note that (Zribi et al., 2015) consider an intervention of a speaker as a paragraph. They segment it by defining a sentence as a semantically meaningful unit.

The STAC corpus is orthographically transcribed and enriched with many annotation marks, such as the hesitation marks, named entities, language marks, non-linguistic words, etc. Some of these marks are eliminated and others need special treatment to get a homogeneous and usable corpus to tag them. We kept some marks, which are useful in the task of tagging, such as the mark of named entities and the language mark.

5.2. Tunisian dialect morphological analyzer (Al-Khalil-TUN)

Only a few tools have been developed to analyze the morphology of DA. Generally, proposed methods for DA morphology focus on extending MSA tools to cover DA phenomena (Habash et al., 2012b; Habash and Rambow, 2006; Salloum and Habash, 2014). In comparison, only two works (Hamdi, 2015; Zribi et al., 2013b) focus on TD. In the present work, we utilized Al-Khalil-TUN (Zribi et al., 2013b) released under a license that made it free software, unlike the analyzer proposed by Hamdi (2015). Al-Khalil-TUN (Zribi et al., 2013b) is an adapted version of the MSA morphological analyzer Al-Khalil (Boudlal et al., 2010) which is a "root-pattern" based morphological analyzer. To adapt this analyzer, (Zribi et al., 2013b) created a TD lexicon composed of roots and patterns related to their morphological characteristics (Zribi et al., 2013b).

Given that there is no "root-pattern" lexicon for TD, Zribi et al. (2013b) exploited the points of similarity between TD and MSA for lexicon development. The transformation of MSA patterns into TD patterns and the extraction of TD specific roots and patterns represent the main steps in the creation of a TD lexicon. These steps are based on a MSA lexicon, which is composed of roots, patterns, affix, and function words.

The first step consists in determining the corresponding patterns in TD from a set of MSA patterns. Zribi et al. (2013b) derived a set of TD patterns from the MSA lexicon while preserving the MSA roots.

The generated lexicon is then used to extract TD roots and patterns. Zribi et al. (2013b)) started with a lexicon composed of a set of TD roots and TD derivation patterns, and a set of unknown words. If a set of conditions is satisfied, then, Zribi et al. (2013b) added these roots and patterns to the lexicon.

The lexicon is, then, improved by adding a list of TD clitics and function words. To obtain this list, Zribi et al. (2013b) translated all MSA function words and clitics and extracted some others from the STAC corpus.

The generated lexicon was integrated in the process of morphological analysis of Al-Khalil (Boudlal et al., 2010). Moreover, they added new rules to the process of word tokenization.

Zribi et al. (2013b) used a part of the STAC corpus for training and testing the TD version of Al-Khalil. The evaluation results of the system are good, since they have achieved an F-measure of 88.86%.

Table 1
Example of sentences.

Sentence Type	Examples	
S3	Speaker A	أنا السنة مانيش باش نمشي للبحر <ĀnA Alsnħ mAnyš bAš nmšy ll-bHr> 'This year, I won't to go to the beach'
S1	Speaker B	خاطر خايف <xATr xAyf> 'because I'm frightened'
S2	Speaker A	علاش ما. . . <ʕlAš mA> 'Why not. . . '
S3	Speaker A	يتفرج عليه. <ytfrj ʕly-h> 'He watches him'
S4	Speaker A	الحاجة الحاجة الحاجة ال- الحاجة اللي فهمتها. <Al-HAjħ Al-HAjħ Al-HAjħ Al- AilHAjħ Ally fhmt-hA> 'The thing, the thing, the thing, the the thing that I understood'

6. Method overview

The aim of this work is to build a MD system for TD. Our starting point is the output of the morphological TD analyzer Al-Khalil-TUN (Zribi et al., 2013b). We propose to extend the analyzer by developing and integrating a disambiguation module. Fig. 1 presents the architecture of the TAMDAS system. The main steps of our system for tagging TD are as follows:

- *Automatic sentence boundary detection of transcribed TD.* We aimed at the integration of an automatic model for identifying sentence boundaries. We integrated the system developed by Zribi et al. (2016), which can detect boundaries of transcribed oral sentences. This system uses three different methods to detect the sentence boundary (rule-based, statistical and hybrid methods). We have applied the statistical method that gives the best evaluation results.
- *Morphological analysis.* We then morphologically analyzed the words of the sentences. A set of analyses is assigned to each word.
- *Morphological disambiguation.* We propose to develop a technique that can choose one correct analysis among a set of analyses given to a word, while considering the context.

Most POS tagging and MD algorithms are either rule-based or stochastic. Handcrafting a set of rules for MD of an agglutinative language may not be an applicable solution, and requires a considerable effort. Furthermore, stochastic taggers (Hidden Markov Models (HMM) (Marshall, 1987), Transformation-Based Learning (Brill, 1994), etc.) or classification-based taggers (Support Vector Machine (SVM), Conditional Random Field (CRF), etc.) use an annotated corpus to generate models which can be applied to non-annotated data. In this work, our starting point is the output of a morphological analyzer, Al-Khalil-TUN (Zribi et al., 2013b) where each word has different morphological analyses and possible tokenizations. We propose to explore a rule-based classification method for the task of MD of TD. We aim to automatically extract a set of tagging rules from an annotated corpus. The classification rules will classify each analysis given by the morphological analyzer into two classes: *true* and *false*. These rules are based on the values of different components of the analysis and the position of the word in the sentence.

The idea of exploring a rule-based classifier for the task of POS tagging has been adopted by few works. Piasecki and Wardyński

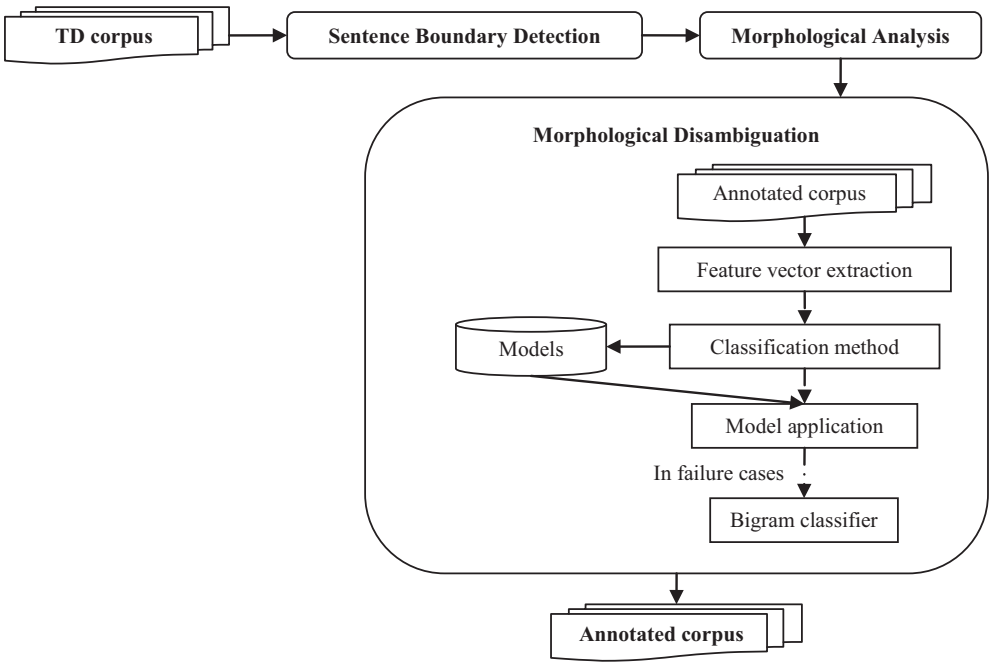

Figure 1. Architecture of TAMDAS system.

(2006) combined the result of two rule-based classifiers with handcrafted rules for the POS tagging of a small corpus of Polish, which is a relatively free word order language. Habash and Rambow (2005) used a rule-based algorithm to choose the correct analysis in the morphological disambiguation of MSA–MADA. Roth et al. (2008) have extended MADA by applying an automatic tuning of classifier parameters to choose only one correct morphological analysis.

Our work is similar to the one developed by Habash and Rambow (2005), Pasha et al. (2014) and Roth et al. (2008) for MSA text tagging. Nonetheless, our proposed method is different in many respects. It is simpler. It consists in training only one classifier for all morphological features to clarify the morphological results. In contrast, the method of Habash and Rambow (2005), Pasha et al. (2014), Roth et al. (2008) proposes to train a classifier for each morphological features. Moreover, the step of choosing only one correct result in failure cases is different from ours.

6.1. Analysis classification

We used the STAC corpus for training and testing our method. The STAC annotation consists simply in selecting the correct analysis produced by the analyzer, or an indication that no such analysis exists.

We did experiments using two classification methods belonging to the rule-based classifier included in the WEKA machine learning tool (Hall et al., 2009), including PART (Mohamed et al., 2013) and RIPPER (Cohen, 1995). We also tested a SVM classifier (Vapnik, 1995) in order to compare the effect of the use of three different classifiers on TD morphological disambiguation.

6.1.1. The classifiers

6.1.1.1. RIPPER. A propositional rule learner was implemented by (Cohen, 1995). The RIPPER hypothesis is expressed as a set of "if-then" rules. It consists of two main phases: the first phase constructs an initial rule set using a rule induction algorithm, while the second one optimizes initially obtained rule sets. The training dataset is randomly divided into two subsets: a growing set, which usually consists of 2/3 of the examples, and a pruning set, which consists of the remaining 1/3. The growing set is used for the initial rule construction (the rule growth phase) and the pruning set is used for pruning (the rule pruning phase). A minimum Description Length (MDL) based heuristic is used as a criterion to stop the rule construction process (Cohen, 1995; Mohamed et al., 2013).

6.1.1.2. PART. It is a partial decision tree algorithm, which is the developed version of C4.5 and RIPPER algorithms (Mohamed et al., 2013). The most important feature of the PART algorithm is that it does not need to perform a global optimization like C4.5 and RIPPER to generate accurate rules, but it follows a "separate-and-conquer" strategy. For example, it builds a rule and removes the instances. It, also, covers, and continues to create a recursive rule for the remaining instances until there are no instances. PART builds a partial C4.5 decision tree in every iteration and makes the "best" leaf into a rule (Mohamed et al., 2013).

6.1.1.3. Support Vector Machines. SVM (Vapnik, 1995) are a widely used technique in solving classification and regression problems. They are a generalization of the most popular linear classifiers. SVM are robust for noisy data and have a powerful ability of generalization, especially in the presence of a large number of features. They are insensitive to the number of examples of the training data (positive or negative). SVM have been successfully used in many NLP research and for the POS tagging task.

6.1.2. Features

Feature choice is of utmost importance in the overall classifier design. The basic features can be classified into three classes: morphological, contextual and dynamic features.

Morphological features are defined for each word analyzed by the morphological analyzer. Table 2 presents the morphological feature vectors identified by the Al-Khalil-TUN and used for generating a classifier model. It should be noted that we used the relevant morphological features generated by the morphological analyzer. For example, we do not use "root" or "pattern" features. A relevant feature is one that helps in the MD task. For example,

Table 2
The morphological features used.

Feature	Abbreviation	Possible values
Part-of- Speech	POS	verb, noun, adv, etc.[a]
Person	Per	1 (first person), 2 (second person), 3 (third person), na (not applicable).
Number	Num	s (singular), d (dual), p (plural), u (undefined), na.
Voice	Vox	a (active), p (passive), na.
It agglutinates pronoun	Pron	Yes, no, na.
It agglutinates conjunction	Conj	Yes, no, na.
Gender	Gen	f (feminine), m (masculine), na.
It agglutinates particle	Part	Yes, no, na.
It agglutinates negation particle	Neg	Yes, no, na.
Definiteness	Def	Yes, no, na.
It contains a particle.	Part	Yes, no, na.

[a] adj, interrog_adv, adv_place, adv_temp, conj, sub_conj, fw, ind_obj_pron, noun_count, prop_noun, number, number_noun, part, part_abst, part_cond, part_fut, part_interrog, part_neg, part_restrict, part_verb, part_voc, prep, pron, dem_pron, poss_pron, rel_pron, rel_adv, sub_conj, verb.

the voice feature is relevant in disambiguating the POS tag verb. If the analyzer does not recognize a word, all the values of these features are replaced by value "*u*"[4]. We used the value "*u*" for the morphological features of the disfluent words.

We note that the POS tag set used by our system is very similar to the one used by the MADAMIRA system (Pasha et al., 2014).

The *contextual feature* is a window of +/−*n* words from the word to tag. We tested different values of *n*. We did experiment with *n = 0*, *n = 1*, and *n = 2*. We showed that *n = 2* is the best configuration for our task.

We also used the feature *position*, a *contextual feature*, which specifies the position of the word in the sentence. This feature takes three possible values: *B* is for a word located at the beginning of a sentence; *E* is for the word located at the end of a sentence, *I* for the other positions in a sentence.

Finally, we chose a *dynamic feature,* which uses the POS tags that are dynamically assigned to the two preceding words.

In all experiments presented in this section, we performed 10-fold cross-validation runs. We reported the weighted average of the 10 runs on the entire corpus. We chose a 10-fold cross-validation because the size of the STAC corpus is relatively small.

In Table 3, we presented the F-measure of different feature combinations. We noticed that the use of morphological, dynamic and contextual features gives the best performance for our three classifiers: SVM, RIPPER and PART. Best results are in bold.

Table 4 presents the value of recall, precision and F-measure for the classification results using the best feature combination. We noticed that SVM gives the best evaluation results. Best results are in bold.

6.1.3. Creation of training data

The STAC corpus is composed of a set of all the possible morphological analyses for each word, with the unique correct analysis marked. According to the analysis attributed to the preceding and the following words, a *true* or *false* class is attributed to each word analysis. Take the example of a sentence composed of three words. The first word w_{i-1} has three solutions. The first solution Sol_1 is labeled *true*. It is followed by the word w_i that has two solutions. Its second solution Sol_2 is labeled *true*. The third word has two analyses. To create training instances for the second word w_i, we combine its analysis with that of the preceding and the following words. We assign the class *true* only if the analyses of the current, the preceding and the following words are all *true*. Fig. 2 illustrates this example.

[4] Undefined.

6.2. Bigram classifier

Sometimes, the application of generated models fails to choose one correct analysis for a given word. This is due to the ambiguity that exists in TD. Therefore, we use a bigram classifier to choose one correct analysis. The bigram classifier calculates the frequencies of each bigram tag (POS_{i-1}, POS_i) based on our training corpus. The probabilities for each bigram of tags are stored in a bigram dictionary. If the classifier fails, we attribute the tag "unrecognized" to the word.

7. Evaluation and discussion

In this paper, we presented our method to create TAMDAS, a system for POS tagging TD in a spoken variety of Arabic language. To train and test TAMDAS, we divided the STAC corpus into two sets. We used 35708 words for the training of our systems (TAMDAS and the baseline) and 6680 words for the evaluation.

To test the performance of our system, we used two tag sets. The first one does not contain POS tags for oral phenomena. The second one incorporates the following tags: *onom*, *TrunW*, *interj*, *FPause* and *break* to respectively mark onomatopoeia, truncated words, interjections, filled pauses, and silent pauses.

Tables 5 and 6 present the error rates of words correctly classified with and without speech tags based on three classifiers. We ignored the step of bigram classifier when calculating these values. Best results are in bold.

The evaluation shows that the rule-based classifiers give the best results compared to a statistical classifier. We noticed that the results given by the PART classifier are the best ones.

In order to compare our system with another TD POS tagger, we developed a baseline system. The baseline method is very simple. It assigns to each word the tag most frequently attributed to the word in the training corpus. In order to achieve this, we used a lexicon composed of words and all their possible POS tags. We attributed to each such pair (word, POS tag) its frequency in the training corpus. Then, we projected this lexicon on the test corpus and gave each word the most frequent POS tag.

We compared our results to the work of Boujelbane et al. (2014) for tagging TD. For retraining the Stanford MSA POS tagger (Toutanova and Manning, 2000) with a TD corpus, Boujelbane et al. (2014) used a TD corpus that is the result of the Arabic Treebank's translation into TD. The MSA's percentage in this corpus is high. Note that the Arabic Treebank is composed of a transcribed set of TV news in MSA.

Since our system uses a POS tag set similar to MADAMIRA (Pasha et al., 2014), we compared it with MADAMIRA (Pasha et al., 2014) which can disambiguate MSA and Egyptian sentences.

Ideally, we would like to compare the performance of these systems against a TD morphosyntacically annotated gold standard. The systems of Boujelbane et al. (2014) and Pasha et al. (2014) cannot annotate speech phenomena. Therefore, we filtered all speech words (incomplete words, repeated word, filled pauses, etc.). We also noticed that the tag set of the system of Boujelbane et al. (2014) is sometimes different from ours. Therefore, we tried to reduce the differences.

Table 3
Classification results with all possible feature combinations.

Features		RIPPER	PART	LibSVM
Morphological	POS + Gen, Num, Per, Vox, Asp	0.783	0.789	0.781
	+ Pron, Neg, Intero	0.786	0.791	0.781
	+ Conj, Def, Part	0.78	0.799	0.788
Morphological + dynamic		0.78	0.773	0.766
Morphological + contextual		0.879	0.881	0.865
Morphological + dynamic + contextual		**0.910**	**0.905**	**0.914**

Table 4
Classification results with all possible values of *n*.

		0	1	2
Ripper	Recall	0.825	0.839	0.935
	Precision	0.831	0.828	0.923
	F-measure	0.828	0.807	**0.910**
Part	Recall	0.892	0.891	0.911
	Precision	0.896	0.886	0.903
	F-measure	0.892	0.886	**0.905**
SVM	Recall	0.876	0.861	0.937
	Precision	0.878	0.875	0.939
	F-measure	0.875	0.829	**0.914**

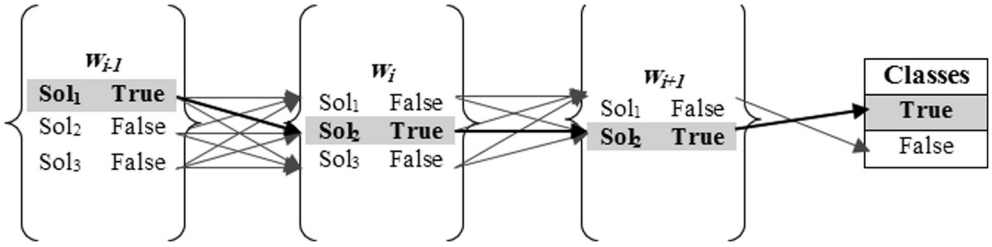

Figure 2. Assigning classes for the analysis of the word w_i according to its context.

Table 5
Error rates for words correctly classified with and without speech tags.

	Speech tag set (Spe) (%)	Ordinary tag set (Ord) (%)
RIPPER	34.83	48.49
PART	**25.66**	**34.92**
SVM	35.56	43.17

Table 7
Evaluation results.

		Accuracy (%)
Baseline		68.51
TAMDAS		**85.49**
MADAMIRA	MSA	56.63
	EGY	58.41
Stanford TD		51.82

Table 6
Error rate for some POS tags with and without speech tags.

	PART		RIPPER		SVM	
POS	Ord	Spe	Ord	Spe	Ord	Spe
adj	18.52	16.67	25.93	25.93	25.93	25.93
adv	75	65.63	87.50	75	87.50	78.13
dem_pron	20	23.33	26.67	26.67	26.67	26.67
interrog_adv	8.33	16.67	25	33.33	33.33	33.33
neg_part	55.17	68.97	82.76	79.31	89.66	82.76
noun	23.49	22.54	33.02	31.75	33.33	32.38
number_noun	28.57	25	35.71	35.71	35.71	35.71
prep	31.18	32.26	39.78	37.63	40.86	39.78
prop_noun	10.53	7.89	15.79	13.16	15.79	13.16
verb	30.37	28.80	48.17	44.50	48.17	44.50

Table 7 presents the evaluation results of the three systems (TAMDAS system, Stanford TD (Boujelbane et al., 2014) MADAMIRA (Pasha et al., 2014) and our baseline). Best results are in bold.

The comparison shows that our system has the highest accuracy. We also notice that Stanford TD (Boujelbane et al., 2014) gave an accuracy much lower than that of the TAMDAS system. The nature of the training corpus of both systems is the primary cause of the gap. Indeed, the Tunisian dialect treated by the system of Boujelbane et al. (2014) is an "intellectualized" dialect[5] which has not the same dialect nature of the STAC corpus.

The results reported by MADAMIRA MSA and the Egyptian version (EGY) are very close to the results of Boujelbane et al. (2014))'s system. The work of Jarrar et al. (2014)) on the Palestinian dialect showed that use of MADAMIRA makes a good initial baseline (78%). This can be justified by two reasons. First, the Palestinian dialect is very close to the Egyptian one as they have multiple characteristics in common. These dialects belong to Eastern dialects. In contrast, TD belongs to another group of dialects (Western dialect) that present multiple differences.

Second, the training corpus of our system is different from MADAMIRA system's. Our training corpus is based on oral transcripts but the MADAMIRA training corpus is related to the written form of Egyptian and MSA texts.

The failure cases of our system are generally due to semantic ambiguity. Let us take the TD sentence: رجعت سالمة للمدرسة<rjʕt sAlmħ ll-mdrsħ>. This sentence has two possible meanings: 'I came back to school safely' and 'Selma came back to school'. The word سالمة <sAlmħ> has two different POS tags: adjective and proper noun. Both POS are applicable in this context. In this case, the MD fails to choose the correct POS and a semantic disambiguation is needed.

8. Conclusion

In this paper, we proposed a method for building a morphological disambiguation system for the Tunisian dialect based on the output of the morphological analyzer for TD Al-Khalil-TUN (Zribi et al., 2013b). We experimented different classifiers and a bigram classifier to choose the best morphological analysis for a given word in a given context. This system showed encouraging results (accuracy = 87.32%).

As a future work, we intend to add semantic features in order to ameliorate the results of our system. We also intend to realize an extrinsic evaluation of our system on some NLP applications dealing with the spoken form of the Tunisian dialect. Finally, we aim to expand both training and test corpora in order to maximize the coverage of the TD lexicon.

References

Al-Badrashiny, M., Eskander, R., Habash, N., Rambow, O., 2014. Automatic transliteration of romanized Dialectal Arabic. In: Proceedings of the Eighteenth Conference on Computational Natural Language Learning. Baltimore, Maryland USA, pp. 30–38.

Al-Sabbagh, R., Girju, R., 2012. A supervised POS tagger for written Arabic social networking corpora. Proceedings of KONVENS 2012 (Main Track: Oral Presentations), Vienna, pp. 39–52.

Al-Shargi, F., Kaplan, A., Eskander, R., Habash, N., Rambow, O., 2016. Morphologically annotated corpora and morphological analyzers for Moroccan and Sanaani Yemeni Arabic. In: Proceedings of the Language Resources and Evaluation Conference (LREC). Portorož, Slovenia, pp. 1300–1306.

Al-Taani, A., Al-Rub, S.A., 2009. A rule-based approach for tagging non-vocalized Arabic words. Int. Arab J. Inf. Technol. 6, 320–328.

Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Ould Abdallahi Ould Bebah, M., Shoul, M., 2010. Alkhalil Morpho SYS1: a morphosyntactic analysis system for Arabic texts. International Arab Conference on Information Technology, pp. 1–6.

Boujelbane, R., 2015. Traitements linguistiques pour la reconnaissance automatique de la parole appliquée à la langue arabe : de l'arabe standard vers l'arabe dialectal. Université de Sfax, université d'Aix Marseille.

Boujelbane, R., Ellouze, M., Béchet, F., Belguith, L., 2014. De l'arabe standard vers l'arabe dialectal : projection de corpus et ressources linguistiques en vue du traitement automatique de l'oral dans les médias tunisiens. Tal 55, 73–96.

Brill, E., 1994. Some advances in transformation-based part of speech tagging. Wall Str. J. 1, 6.

Buckwalter, T., 2004. Buckwalter Arabic Morphological Analyzer Version 2.0, LDC Corpus Catalog No. LDC2004L02.

Cohen, W.W., 1995. Fast effective rule induction. Proceedings of the Twelfth International Conference on Machine Learning., 115–123 101.1.50.8204.

Darwish, K., 2014. Arabizi detection and conversion to Arabic. In: Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP). Doha, Qatar, pp. 217–224.

Diab, M., Hacioglu, K., Jurafsky, D., 2004. Automatic tagging of Arabic text : from raw text to base phrase chunks. HLT-NAACL 2004: Short Papers, pp. 149–152.

Dister, A., Constant, M., Purnelle, G., 2009. Normalizing speech transcriptions for Natural Language Processing. Proceedings of the 3rd International Conference on Spoken Communication (GSCP'09).

Duh, K., Kirchhoff, K., 2005. POS tagging of dialectal Arabic: a minimally supervised approach. Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, 55–62.

Gibson, M.L., 1998. Dialect Contact in Tunisian Arabic : Sociolinguistic and Structural Aspects. University of Reading.

Habash, N., Diab, M., Rambow, O., 2012a. Conventional orthography for dialectal Arabic. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey, pp. 711–718.

Habash, N., Eskander, R., Hawwari, A., 2012b. A morphological analyzer for Egyptian Arabic. In: Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON2012). Montréal, Canada, pp. 1–9.

Habash, N., Rambow, O., 2006. MAGEAD : a morphological analyzer and generator for the Arabic dialects. Proceedings of Coling-ACL 2006.

Habash, N., Rambow, O., 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. Proc. 43rd Annu. Meet. Assoc. Comput. Linguist. – ACL'05, pp. 573–580. http://dx.doi.org/10.3115/1219840.1219911.

Habash, N., Rambow, O., Kiraz, G., 2005. Morphological analysis and generation for Arabic dialects. Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, Ann Arbor., 17–24

Habash, N., Roth, R., Rambow, O., Eskander, R., Tomeh, N., 2013. Morphological analysis and disambiguation for dialectal Arabic. In: Proceedings of NAACL-HLT 2013. Atlanta, Georgia, pp. 426–432.

Habash, N., Soudi, A., Buckwalter, T., 2007. On arabic transliteration. In: Abdelhadi, S., Antal, van den B., Günter, N. (Eds.), Arabic Computational Morphology: Knowledge-Based and Empirical Methods.

Habash, N.Y., 2010. Introduction to Arabic natural language processing. Synth. Lect. Hum. Lang. Technol. 3, 1–187. http://dx.doi.org/10.2200/S00277ED1V01Y201008HLT010.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. ACM SIGKDD Explor. 11, 10–18. http://dx.doi.org/10.1145/1656274.1656278.

Hamdi, A., 2015. Traitement automatique du dialecte tunisien à l'aide d'outils et de ressources de l'arabe standard : application à l'étiquetage morphosyntaxique. Aix-Marseille Université.

Hamdi, A., Nasr, A., Habash, N., Gala, N., 2015. POS-tagging of Tunisian dialect using standard Arabic resources and tools. In: Proceedings of the Second Workshop on Arabic Natural Language Processing. Beijing, China, pp. 59–68.

Holes, C., 2004. Modern Arabic: Structures, Functions, and Varieties, Georgetown. ed. Washington.

Jarrar, M., Habash, N., Akra, D., Zalmout, N., Bank, W., 2014. Building a Corpus for Palestinian Arabic : a Preliminary Study 18–27.

Khalifa, S., Habash, N., Abdulrahim, D., Hassan, S., 2016. A large scale corpus of Gulf Arabic. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia, pp. 4282–4289.

Khoja, S., 2001. APT : Arabic part-of-speech tagger. Proceedings of the Student Workshop at the Second Meeting of the North American Association for Computational Linguistics (NAACL2001). Carnegie Mellon University, Pennsylvania.

Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N., Eskander, R., 2014. Developing an Egyptian Arabic treebank : impact of dialectal morphology on annotation and tool development. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland, pp. 2348–2354.

Mahfoudhi, A., 2002. Agreement lost, Agreement Regained: A Minimalist Account of Word Order and Agreement Variation in Arabic. Calif. Linguist, Notes XXVII.

Marshall, I., 1987. Tag selection using probabilistic methods. Comput. Anal. English, 42–56.

Mejri, S., Said, M., Sfar, I., 2009. Pluringuisme et diglossie en Tunisie. Synerg. Tunisie 1, 53–74.

Mohamed, W.N.H.W., Salleh, M.N.M., Omar, A.H., 2013. A comparative study of reduced error pruning method in decision tree algorithms. Proceedings – 2012

[5] Intellectualized dialect is a mixture between MSA and TD (Boujelbane, 2015).

IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2012, pp. 392–397. http://dx.doi.org/10.1109/ICCSCE.2012.6487177.
Ouerhani, B., 2009. Interférence entre le dialectal et le littéral en Tunisie : Le cas de la morphologie verbale. Synerg. Tunisie 1, 75–84.
Pasha, A., Al-Badrashiny, M., Diab, M., Kholy, A. El, Eskander, R., Habash, N., Pooleery, M., Rambow, O., Roth, R.M., 2014. MADAMIRA : a fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In: The Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1094–1101.
Piasecki, M., Wardyński, A., 2006. Multiclassifier Approach to Tagging of Polish, in: Proceedings of 1st International Symposiuim Advances in Artificial Intelligence and Applications. pp. 169–178.
Rambow, O., Chiang, D., Diab, M., Habash, N., Hwa, R., Sima'an, K., Lacey, V., Levy, R., Nichols, C., Shareef, S., 2006. Parsing Arabic Dialects.
Roth, R., Rambow, O., Habash, N., Diab, M., Rudin, C., 2008. Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature Ranking Nizar Habash – Academia.edu. Proc. 46th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol. Short Pap, pp. 117–120. http://dx.doi.org/10.3115/1557690.1557721.
Saadane, H., Habash, N., 2015. A conventional orthography for Algerian Arabic. In: Proceedings of the Second Workshop on Arabic Natural Language Processing. Beijing, China, pp. 69–79.
Saidi, D., 2007. Typology of Motion Event in Tunisian Arabic. LingO, pp. 196–203.
Salloum, W., Habash, N., 2014. ADAM: analyzer for dialectal Arabic morphology. J. King Saud. Univ. – Comput. Inf. Sci. 26, 372–378. http://dx.doi.org/10.1016/j.jksuci.2014.06.010.
Salloum, W., Habash, N., 2011. Dialectal to standard Arabic paraphrasing to improve Arabic-English statistical machine translation. in: Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK. pp. 10–21.
Samih, Y., Maier, W., 2016. An Arabic-Moroccan darija code-switched corpus. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia, pp. 4170–4175.
Talmoudi, F., 1980. A Morphosyntactic Study of Romance Verbs in the Arabic Dialects of Tunis, Susa, and Sfax. Göteborg Acta Univ, Gothobg.
Tilmatine, M., 1999. Substrat Et Convergences: Le Berbère Et L'arabe Nord-Africain. In: HAAK, M., JONG, R., DE VERSTEEGH, K. (Eds.), Estudios de Dialectología Norteafricana Y Andalusí.
Tlili-Guiassa, Y., 2006. Hybrid method for tagging Arabic text. J. Comput. Sci. 2, 245–248.
Toutanova, K., Manning, C.D., 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. Proc. 2000 Jt. SIGDAT Conf. Empir. methods Nat. Lang. Process. very large corpora held conjunction with 38th Annu. Meet. Assoc. Comput. Linguist., vol. 13, pp. 63–70. http://dx.doi.org/10.3115/1117794.1117802.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer. http://dx.doi.org/10.1109/TNN.1997.641482.
Zribi, I., 2016. Traitement automatique du Dialecte Tunisien : construction de ressources linguistiques. University of Sfax.
Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Belguith, L., Habash, N., 2014. A conventional orthography for Tunisian Arabic. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 26–31.
Zribi, I., Ellouze, M., Belguith, L.H., Blache, P., 2015. Spoken Tunisian Arabic Corpus " STAC ": transcription and annotation. Res. Comput. Sci. 90, 1–13.
Zribi, I., Graja, M., Khemakhem, M.E., Jaoua, M., Belguith, L.H., 2013a. Orthographic transcription for spoken Tunisian Arabic. In: Gelbukh, A. (Ed.), CICLing 2013, Part I, LNCS 7816, pp. 153–163.
Zribi, I., Kammoun, I., Ellouze, M., Belguith, L.H., Blache, P., 2016. Sentence boundary detection for transcribed Tunisian Arabic. In: KONVENS 2016. Bochum, Germany, pp. 323–331.
Zribi, I., Khemakhem, M.E., Belguith, L.H., 2013b. Morphological analysis of Tunisian dialect. In: International Joint Conference on Natural Language Processing. Nagoya, Japan, pp. 992–996.