Natural Language Processing for Dialectical Arabic: A Survey

Abdulhadi Shoufan
Khalifa University
Abu Dhabi, U.A.E.
abdulhadi.shoufan@kustar.ac.ae

Sumaya Al-Ameri
Khalifa University
Abu Dhabi, U.A.E.
sumaya.alameri@kustar.ac.ae

Abstract

This paper presents a wide literature review of natural language processing for dialectical Arabic. Four main research areas were identified and the dialect coverage in research work was outlined. The paper can be used as a quick reference to identify relevant contributions that address a specific NLP aspect for a specific dialect.

1 Introduction

The last ten years have experienced a growing interest in natural language processing for dialectical Arabic. This growth can be attributed to several factors including the wide usage of Arabic dialects in social media. The topics treated by computational linguists for Arabic dialects range from fundamental language aspects including morphology up to sophisticated solutions such as machine translation.

To have an overview of the research that has been done in this area we went through as many papers as possible and tried to specify the main contributions of each paper. We could identify four main categories, whereas each category has some subcategories. The main categories are basic language analyses, building language resources, semantic-level analysis and synthesis, and identifying Arabic dialects. Then, we mapped each paper to categories and subcategories as well as to the addressed dialect or dialects in a matrix form as given in Table 1. By this means, it can be easily identified what has been done in the Arabic NLP, by whom, and for what dialects.

The following four sections describe the related work in the four main categories. For space reasons, however, we limited the description to main aspects. The final section provides a brief discussion of the findings of this survey.

2 Basic Language Analyses

Several solutions have been proposed for the morphological analysis, syntactical analysis, and orthographic analysis and generation. The following three sections describe these solutions, respectively.

2.1 Morphological Analysis and POS Tagging

The morphology of dialectal Arabic had gained early attention by computational linguists. In (Habash & Rambow, 2006), a morphological analyzer and generator, denoted MAGED, was presented. This tool is able to analyze the Levantine dialect and to convert MSA to Levantine. In a later publication the authors detailed the morphophonemic and the orthographic rules encoded in MAGEAD (Habash & Rambow, 2007).

In (Habash, Eskander, & Hawwari, 2012), a morphological analyzer for Egyptian Arabic is proposed with further development in (Salloum & Habash, 2014).

In (Almeman & Lee, 2012), two morphological analyzers for Gulf, Levantine, Egyptian, North African, Sudani, and Iraqi dialects were presented. The first one relies on a MSA morphological analyzer. The second one applies word segmentation and uses web data as a corpus to produce statistical information about the frequency of different segment combinations. In (Zribi, Khemakhem, & Belguith, 2013), a morphological analyzer for the Tunisian dialect based on a MSA analyzer was proposed. Furthermore, a lexicon for the Tunisian dialect is built as an expansion of a MSA lexicon. An unsupervised approach for morphological segmentation was applied to improve machine translation from the Qatari dialect to English (Al-Mannai et al., 2014).

In (Duh & Kirchhoff, 2005), a part-of-speech tagger for Egyptian Arabic was proposed based on a morphological analyzer for MSA and a min-

Proceedings of the Second Workshop on Arabic Natural Language Processing, pages 36–48,
Beijing, China, July 26-31, 2015.

imally supervised approach that requires raw text data from several Arabic varieties.

In (Al-Sabbagh & Girju, 2012a), a function-based POS tagger is proposed that was trained on a manually-annotated Egyptian Arabic corpus.

In (Habash et al., 2013) a MSA morphological tagger is retargeted to Egyptian Arabic. The solution performs part-of-speech tagging, diacritization, lemmatization, and tokenization.

A rule-based stemmer for Arabic Gulf dialect was proposed in (Abuata & Al-Omari, 2015), and a fine-grained POS tagger for Tunisian dialect was presented in (Boujelbane et al., 2014).

2.2 Syntax and Parsing

The syntax of Arabic dialects was purely addressed in the context of computational linguistics. In (Brustad, 2000), the author presented a comparative study of Moroccan, Egyptian, Syrian, and Kuwaiti dialects with respect to syntax however without computational aspects.

In (Chiang et al., 2006) a parser for the Levantine Arabic is proposed. The parser doesn't rely on annotated Levantine corpus or a parallel Levantine-MSA corpus. Rather, the Levantine word is translated into a bag of MSA words that are scored and decoded relying on MSA corpus. The resulting text is then parsed using an MSA parser. Finally, the terminal nodes in the resulting parse structure are replaced with the original Levantine words.

Levantine was also the dialect treated in (Maamouri et al., 2006). In this work a pilot Levantine Arabic Treebank is developed by a morphological and syntactic annotation of 26,000 words of Levantine Arabic conversational telephone speech. The Treebank was used to develop and evaluate parsers for Levantine texts. Grammatical mapping rules were defined to provide language resources for machine translation from Tunisian dialect to MSA and other target languages in (Sadat, Mallek, et al., 2014).

2.3 Orthographic Analysis

In contrast to MSA, dialectical Arabic has no orthographic standard. The same word can be written in different forms. This poses difficulties to NLP tools. In (Dasigi & Diab, 2011), first steps towards normalizing Arabic dialects orthography for Levantine and Egyptian were made. For that, different similarity measures were employed that exploit string similarity and contextual semantic similarity.

In (Habash, Diab, & Rambow, 2012), a conventional orthography is proposed to help building computational models for Arabic dialects in general and Egyptian in particular. The rules and guidelines produced were named CODA.

Recently, a conventional orthography for Tunisian Arabic was proposed in (Zribi et al., 2014). Also, Several papers on the transliteration from Arabizi into Arabic orthography appeared (Bies et al., 2014), (Darwish, 2013), (Masmoudi et al., 2015). Arabizi is Arabic text written in Latin characters.

In (Zribi, Graja, et al., 2013), orthography guidelines for Tunisian dialect were presented for the purpose of transcribing a Tunisian speech corpora. The rules presented are based on the standard Arabic transcription conventions. This work was later used in (Zribi, Khemakhem, & Belguith, 2013) for morphological analysis presented in the Morphological Analysis and POS Tagging section.

3 Building Resources for Dialectal Arabic

The problem of the lack of language resources in dialectical Arabic is well known. Many researchers addressed this problem by creating lexicons, wordnets, corpora, and treebanks.

In (Zaghouani, 2014), a useful survey of freely available Arabic corpora including lexicons was presented. The author highlighted the huge lack of freely available dialectal corpora because only two resources could be identified (Graja et al., 2010), (Almeman & Lee, 2013)

In (Sansò, 2004), the MED-TYP project was presented which aimed at building a typological database for Mediterranean languages including MSA and Arabic dialects. While the researchers found out that the Mediterranean could not be identified as a linguistic area in the traditional sense, a number of significant contact phenomena were discovered.

3.1 Building Lexicons and Lexical Analysis

In (Graff et al., 2006), a lexicon for the Iraqi dialect was presented. The lexicon comprises words from recorded speech tagged with pronunciation data, morphology information, and part-of-speech. The annotation was performed manually with the aid of a user interface and supporting

tools.

In (Al-Sabbagh & Girju, 2010) a lexicon for Egyptian Cairene Arabic is described. Each Cairene entry was mapped to its MSA synonym and tagged with its part-of-speech. Additionally, the entry is tagged with its top-ranked meaning according to web queries.

A spelling corrector for the Iraqi dialect was presented in (Rytting et al., 2011). An orthographic density metric is used to motivate the need for a fine-grained ranking method for candidate words.

In (Graff & Maamouri, 2012), the update of three bilingual dictionaries for English-speaking learners of Moroccan, Syrian and Iraqi Arabic was presented. The original editions of the dictionaries were developed by the Linguistic Data Consortium and Georgetown University Press in the 1960's. In the updated dictionaries, both Arabic script and International Phonetic Alphabet orthographies are used. A web interface enables searching, editing, reviewing and managing the lexicon efficiently.

In (Boujelbane et al., 2013), a Tunisian dialect text corpus as well as a method for building a bilingual dictionary are described. The target is to create a language model for a speech recognition system for the Tunisian Broadcast News.

In (Duh & Kirchhoff, 2006), a Levantine lexicon was built using transductive learning through partially annotated text. For the purpose of sentiment analysis of social networks data, a dedicated lexicon for slang sentimental words and idioms was developed in (Hedar & Doss, 2013).

In (Cavalli-Sforza et al., 2013) an Iraqi WordNet is presented based on the MSA WordNet, the English WordNet, and an English-Iraqi dictionary. A Tunisian dialect WordNet was built in (Bouchlaghem & Elkhlifi, 2014) starting from a Tunisian corpus.

3.2 Building Corpora and Treebanks

In (Al-Sabbagh & Girju, 2012b), a primary work on building a multi-genre corpus for Egyptian Arabic was described. The corpus data is compiled from Twitter, blogs, forums, and online knowledge market services. The paper addresses different aspects related to building dialectal Arabic corpora such as function-based web harvesting, dialect identification, vowel-based spelling variation, linguistic hypercorrection, unsupervised part-of-speech tagging and base phrase chunking for dialectal Arabic.

Using the web as a source was also described in (Almeman & Lee, 2013), where multi-dialect Arabic corpora were built for Gulf, Levantine, Egyptian and North African dialects. The work by Boujelbaneon et al. on building a lexicon for Tunisian dialect can be recited here due to building a corpus from Tunisian broadcast news (Boujelbane et al., 2013).

In (Cotterell & Callison-Burch, 2014), a multi-dialect, multi-genre corpus for Egyptian, Gulf, Levantine, Maghrebi, and Iraqi dialects was presented. Another multi-dialecti corpus based on twitter data was built in (Mubarak & Darwish, 2014) for seven different dialects. A preliminary work on a corpus for Palestinian dialect with 43K words was presented in (Jarrar et al., 2014). A parallel corpus for Algerian dialect and MSA was proposed in (Harrat et al., 2014) for the purpose of machine translation.

In (Maamouri et al., 2006), which was cited in Section 2.2, a pilot Levantine Arabic Treebank was presented. A conversational telephone speech with about 26,000 words was annotated with morphological and syntactic data. Recently, Maamouri et al. presented a treebank for the Egyptian Dialect (Maamouri et al., 2014).

As the quality of the annotation process is essential for building accurate language resources, some researchers payed special attention to this process. In (Diab et al., 2010), multiple systems to develop NLP resources for Arabic dialects including Levantine, Egyptian, Moroccan, and Iraqi were presented. The systems utilized MAGEAD (Habash & Rambow, 2006) as well as Buckwalter morphological analyzer and generator (BAMA) (Buckwalter, 2004). The COLABA ability to process Arabic dialects was evaluated through the COLABA information retrieval system.

A web application for annotating Egyptian, Iraqi, Levantine, and Moroccan dialects was proposed in (Benajiba & Diab, 2010). The authors follow non-functional objectives including optimizing speed, accuracy, and efficiency while maintaining the security and integrity of the data. In (Zaidan & Callison-Burch, 2011), the building of a 52M-word Arabic online commentary dataset rich in dialectal content was presented. The long-term annotation effort to identify the dialect level in each sentence was also discussed. The au-

thors of (Elfardy & Diab, 2012b) presented a set of guidelines for detecting code switching in Arabic on the word and token levels. These guidelines were used to annotate a corpus that is rich in Egyptian, Levantine, and Iraqi dialects with frequent code switching to MSA. In (Habash et al., 2008a), guidelines for identifying the level of dialectalness of a certain text were presented. Three levels for dialectalness were proposed: MSA with non-standard orthography, MSA words with dialect morphology, and a Dialectal lexeme.

In (Hawwari et al., 2014), a framework for classifying and annotating Egyptian multi-word expressions in a specialized computational lexicon was proposed. A graphical tool for annotating Moroccan tweets was presented in (Tratz et al., 2013).

In (Zaghouani et al., 2014), comprehensive guidelines for annotating an Arabic corpus including Qatar dialect was proposed. The corpus is denoted Qatar Arabic Language Bank (QALB). A special attention in this work is paid to the manual correction which should provide training data for learning-based Arabic error correction tools.

4 Semantic-Level Analysis and Synthesis

Most work in this area relates to machine translation from or to Arabic dialects. Some papers treat other tasks such as information retrieval and sentiment analysis.

4.1 Machine Translation

In (Bakr et al., 2008), the authors proposed a hybrid approach to convert an Egyptian sentence into its corresponding diacritized MSA. The approach is generic, i.e., it can be extended to other Arabic dialects. Some techniques for lexical acquisition of colloquial words are developed.

In (Sawaf, 2010), a hybrid machine translation system was extended to handle Arabic dialects from 15 regions including Northern Iraq, Baghdad, Southern Iraq, Saudi-Arabia, Southern Arabic Peninsula, Egypt, Sudan, Libya, Morocco, Tunisia, Lebanon, North Syria, Damascus, Palestine and Jordan. A decoding algorithm was developed to normalize non-standard, spontaneous and dialectal Arabic into Modern Standard Arabic.

In (Salloum & Habash, 2011), the quality of Arabic-English statistical machine translation was improved to deal with Levantine and Egyptian dialects using morphological knowledge. A simple rule-based approach was used to generate MSA paraphrases for dialectal Arabic out-of-vocabulary words and low frequency words.

In (Zbib et al., 2012), crowdsourcing was applied to build Levantine-English and Egyptian-English parallel corpora, consisting of 1.1M words and 380k words, respectively. The dialectal sentences were selected from a large corpus of Arabic web text, and translated using Amazon's Mechanical Turk. The data was used to build dialectal machine translation systems.

In (Jehl et al., 2012), the authors collected bilingual sentence pairs for training statistical machine translation systems to translate microblog messages. The paper addressed the Gulf, Levantine, and Egyptian dialects as well as MSA. The technique presented was found to perform better than other methods such as techniques based on extracting phrases from similar text.

In (Al-Gaphari & Al-Yadoumi, 2012) an algorithm was proposed that normalizes Sanaáni dialect to MSA based on morphological rules. Input text was tokenized and each token was analyzed into stem and affixes. The stem and the affixes can be either dialect-specific, MSA-specific, or both. For each morphological rule the algorithm checks the possibility of applying such a rule.

In (Salloum & Habash, 2012), a rule-based approach for machine translation from Arabic dialects to MSA was presented. The approach relies on morphological analysis, morphological transfer rules and dictionaries in addition to language models to produce MSA paraphrases of dialectal sentences. The treated dialects are Levantine, Egyptian, Iraqi, and Gulf Arabic.

In (Mohamed et al., 2012), a translator from MSA to the Egyptian dialect was presented. Among others, this process helps in the annotation of the Egyptian dialect and in the translation from this dialect to English.

In (Soltau et al., 2011), a corpus-based translator from MSA to Levantine was described. The translator is trained on corpora with a mixture of Levantine dialect and MSA.

The Iraqi dialect was studied with respect to MT in two papers by Condon et al. In (Condon et al., 2010), a two-way evaluation of English-Iraqi dialog translation was performed. Four MT systems were evaluated and error types were specified. The English-Iraqi speech translation systems were evaluated using automated metrics. The study described Iraqi speech data features and the

difficulties it presents on machine translation quality evaluation.

In (Jeblee et al., 2014), domain and dialect adaptation was suggested to produce a statistical machine translation system from English to the Egyptian dialect with MSA as a pivot. A machine translation system of the Moroccan dialect into MSA based on statistical models and a rule-based approach was proposed in (Tachicart & Bouzoubaa, 2014).

4.2 Other Semantic Tasks

Sentiment and subjectivity analysis (SSA) was treated in several papers. In (Abdul-Mageed et al., 2014), the authors investigated how to treat Arabic dialects and whether genre-specific features have a measurable impact on performance of a sentiment analyzer.

In (Hedar & Doss, 2013), a classifier for Arabic slang that applies sentiment analysis to classify news and comments on Facebook was presented.

In (Mourad & Darwish, 2013), the issue of limited Arabic SSA lexicons was addressed by providing baselines that employ Arabic specific processing including stemming, POS tagging, and tweets normalization. Also, a random graph walking algorithm was employed to expand SSA lexicons. Open issues in sentiment analysis were discussed in (El-Beltagy & Ali, 2013) and a sentiment lexicon for Egyptian dialect was presented.

Recently, other sentiment analysis systems for social media data were proposed in (Duwairi et al., 2014) and (Ibrahim et al., 2015) for the Jordanian and Egyptian dialects, respectively.

In (El-Fishawy et al., 2014), a microblog summarization technique based on machine learning for Egyptian dialect was presented. The results achieved were compared to several well-known algorithms such as SumBasic, TF-IDF, PageRank, MEAD, and human summaries.

(Pasha et al., 2013) addressed the challenges of retrieving information in Arabic dialects, which have significant linguistic differences from Standard Arabic. The presented tool automatically generates dialect search terms with relevant morphological variations from English or Standard Arabic query terms.

In (Zirikly & Diab, 2014) and (Zirikly & Diab, 2015) different approaches for Named Entity Recognition in the Egyptian dialect were proposed. Named entity recognition in microblogs was also treated by Darwish and Gao, however, for MSA mainly (Darwish & Gao, 2014).

In (Darwish & Magdy, 2014), a general study of Arabic information retrieval was presented. The survey includes different domains and applications of Arabic IR systems as well as the specific challenges in this NLP area.

5 Dialect Identification and Recognition

The recognition of dialectal content in an Arabic text or speech gained a special interest in the literature.

5.1 Dialect Identification in Text

Some of the previously cited work on text annotation, e.g. (Diab et al., 2010) and (Zaidan & Callison-Burch, 2011), or machine translation, e.g., (Soltau et al., 2011), implicitly include components for dialect identification.

In (Habash et al., 2008b), standard annotation guidelines to identify a switching between MSA and an Egyptian or a Levantine dialect in written text were presented. The guidelines can be used to annotate large collections of data used for training and testing NLP tools.

In (Elfardy & Diab, 2013), a supervised approach on the sentence level is proposed to differentiate between MSA and the Egyptian dialect. Token level labels are used to derive sentence-level features that are employed with other core and meta features to train a generative classifier that predicts the correct label for each sentence in the given input text. This work was extended to the Iraqi, Levantine and Moroccan dialects by the same authors in (Elfardy & Diab, 2012a).

In (Zaidan & Callison-Burch, 2012), the authors used a large annotated dataset to train and evaluate automatic classifiers for the sake of Arabic dialect identification. Given an Arabic sentence, the task consists in determining the variety of Arabic in which it is written. The variety can be MSA, Maghrebi, Egyptian, Levantine, Iraqi, or Gulf.

Recently, a native Bayes classifier based on character bi-gram model was proposed to identify 18 different Arabic dialects (Sadat, Kazemi, & Farzindar, 2014). In (Darwish et al., 2014), the authors based their identification approach of the Egyptian dialect on lexical, morphological, as well as phonological information.

In (Zaidan & Callison-Burch, 2014), the authors created a large monolingual dataset with dialect

Table 1. Dialectical Arabic NLP- Literature Overview

	Basic Language Analyses			Building Language Resources		Dialect Identification and Recognition		Semantic Analysis	
	Morph.	**Syntax**	**Orthog.**	**Lexica**	**Corpora**	**From Text**	**From Speech**	**M. Translation**	**Others**
Gulf	(Almeman & Lee, 2012), (Abuata & Al-Omari, 2015)		(Darwish, 2013), (Masmoudi et al., 2015)		(Zaidan & Callison-Burch, 2011), (Almeman et al., 2013), (Cotterell& Callison-Burch, 2014)	(Zaidan & Callison-Burch, 2011), (Sadat, Kazemi, & Farzindar, 2014), (Zaidan & Callison-Burch, 2014)	(Belgacem et al., 2010), (Zaidan&Callison-Burch, 2012), (Zhang et al., 2013), (Biadsy et al., 2009), (Akbacak et al.,2011)	(Jehl et al., 2012), (Salloum & Habash, 2012), (Sawaf, 2010)	(Mourad & Darwish, 2013)
Kuwaiti					(Mubarak & Darwish, 2014)	(Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)		
Saudis					(Mubarak & Darwish, 2014)	(Sadat, Kazemi, & Farzindar, 2014)	(Alghamdi et al., 2008), (Iskra et al., 2004)	(Sawaf, 2010)	
UAE					(Mubarak & Darwish, 2014)		(Lei & Hansen, 2009), (Iskra et al., 2004)	(Khamis, 2007)	
Qatari					(Mubarak & Darwish, 2014), (Zaghouani et al., 2014)	(Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)	(Al- Mannai et al., 2014)	
Bahraini						(Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)		
Omani						(Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)		
S. A. Peninsula								(Sawaf, 2010)	
Yemeni					(Belgacem et al., 2010)				
Sana´ani								(Al- Gaphari & Al Yadoumi, 2012)	
North Africa	(Almeman & Lee, 2012), (Habash et al., 2013)		(Masmoudi et al., 2015), (Darwish, 2013)		(Almeman & Lee, 2013)				
Egyptian	(Duh & Kirchhoff 2005), (Habash et al., 2012), (Almeman & Lee, 2012), (Al-Sabbagh & Girju, 2012a), (Salloum &Habash, 2014)		(Dasigi & Diab, 2011), (Habash, Diab, & Rambow, 2012), (Bies et al., 2014)	(Hedar & Doss, 2013)	(Habash et al.,2008), (Diab et al., 2010), (Benajiba & Diab, 2010), (Zaidan & Callison-Burch, 2011), (Al-Sabbagh & Girju, 2012), (Elfardy& Diab, 2012b), (Elfardy& Diab,2012c), (Almeman& Lee,2013), (Mubarak& Darwish, 2014), (Cotterell& Callison-Burch,2014), (Maamouri et al., 2014), (Hawwari et al., 2014), (Maamouri et al.,2014)	(Diab et al., 2010), (Zaidan & Callison-Burch, 2011), (Elfardy & Diab, 2012), (Elfardy & Diab, 2013), (Zaidan & Callison-Burch, 2012), (Habash et al., 2008b), (Zaidan & Callison-Burch, 2014), (Darwish et al., 2014)	(Belgacem et al., 2010), (Zhang et al., 2013), (Lei & Hansen, 2009), (Biadsy et al., 2009), (Akbacak et al., 2011), (Kirchhoff & Vergyri, 2005), (Iskra et al., 2004)	(Zbib et al.,2012), (Salloum & Habash, 2011), (Jehl et al., 2012), (Bakr et al.,2008), (Salloum & Habash, 2012), (Sawaf, 2010), (Mohamed et al., 2012), (Jeblee et al., 2014)	(Pasha et al., 2013), (Hedar & Doss, 2013), (El- Fishawy et al., 2014), (Ibrahim et al., 2015), (Mourad & Darwish, 2013), (Zirikly & Diab, 2014/2015), (El-Beltagy & Ali, 2013), (Darwish & Gao, 2014)
Cairene				(Al-Sabbagh & Girju, 2010)					
Morrocan				(Graff & Maamouri , 2012)	(Benajiba & Diab, 2010), (Diab et al., 2010), (Tratz et al., 2013) , (Mubarak & Darwish,	(Sadat, Kazemi,& Farzindar, 2014)	(Elfardy & Diab, 2012a), (Belgacem et al., 2010), (Iskra et al., 2004)	(Sawaf, 2010), (Tachicart & Bouzoubaa,	

	Basic Language Analyses			Building Language Resources		Dialect Identification and Recognition		Semantic Analysis	
	Morph.	**Syntax**	**Orthog.**	**Lexica**	**Corpora**	**From Text**	**From Speech**	**M. Translation**	**Others**
					2014)			2014)	
Tunisian	(Zribi, Khemakhem, & Belguith, 2013), (Boujelbane et al., 2014)		(Zribi et al., 2013), (Zribi et al., 2014)	(Boujelbane et al., 2013)	(Boujelbane et al., 2013), (Zribi, Graja, et al., 2013)	(Sadat, Kazemi, & Farzindar, 2014)	(Belgacem et al., 2010), (Boujelbane et al., 2013), (Iskra et al., 2004)	(Sawaf,2010), (Sadat, Mallek, et al., 2014)	
Libyan				(Graja et al., 2010)		(Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)	(Sawaf, 2010)	
Sudani	(Almeman & Lee, 2012)				(Mubarak & Darwish, 2014)	(Sadat, Kazemi, & Farzindar, 2014)		(Sawaf, 2010)	
Algerian					(Harrat et al., 2014)	(Harrat et al., 2015), (Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)		
Maghrebi*					(Cotterell & Callison-Burch, 2014)	Zaidan & Callison-Burch, 2012), (Zaidan & Callison-Burch, 2014)			
Levantine	(Habash &Rambow, 2006), (Habash &Rambow,2007) (Almeman & Lee, 2012),	(Chiang et al., 2006), (Maamouri et al., 2006)	(Habash &Rambow, 2007), (Dasigi & Diab, 2011), (Darwish, 2013), (Masmoudi et al., 2015)	(Duh & Kirchhoff 2006)	(Maamouri et al., 2006), (Diab et al., 2010), (Benajiba & Diab, 2010), (Soltau et al., 2011), (Zaidan & Callison-Burch, 2011), (Elfardy& Diab, 2012b), (Almeman& Lee,2013), (Almeman et al., 2013), (Cotterell & Callison-Burch, 2014)	(Habash et al., 2008), (Habash et al., 2008b), (Diab et al., 2010), (Zaidan & Callison-Burch, 2011), (Zaidan & Callison-Burch, 2012), (Elfardy & Diab, 2012c), (Zaidan & Callison- Burch, 2014)	(Elfardy & Diab, 2012a), (Zhang et al., 2013), (Biadsy et al., 2009), (Akbacak et al., 2011), (Iskra et al., 2004)	(Zbib et al., 2012), (Salloum & Habash, 2011), (Jehl et al., 2012), (Salloum & Habash, 2012), (Soltau et al., 2011)	(Mourad & Darwish, 2013)
Syrian				(Graff & Maamouri, 2012)		(Harrat et al., 2015), (Sadat, Kazemi, & Farzindar, 2014)	(Belgacem et al., 2010), (Lei & Hansen, 2009), (Iskra et al., 2004)		
North Syrian								(Sawaf, 2010)	
Damascus								(Sawaf, 2010)	
Lebanese						(Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)	(Sawaf, 2010)	
Jordanian	(Salloum &Habash, 2014)					(Sadat, Kazemi, & Farzindar, 2014)	(Iskra et al., 2004)	(Sawaf, 2010)	(Duwairi et al., 2014)
Palestinian					(Jarrar et al., 2014)	(Harrat et al., 2015), (Sadat, Kazemi, & Farzindar, 2014)	(Lei & Hansen, 2009), (Iskra et al., 2004)	(Sawaf, 2010)	
Iraqi	(Almeman & Lee, 2012)		(Masmoudi et al., 2015), (Darwish, 2013)	(Graff et al., 2006), (Rytting et al., 2011), (Graff & Maamouri 2012), (Cavalli-Sforza et al., 2013)	(Diab et al., 2010), (Habash et al., 2008a), (Benajiba & Diab, 2010), (Elfardy & Diab, 2012b), (Cotterell& Callison-Burch, 2014)	(Zaidan & Callison-Burch, 2012), (Zaidan & Callison-Burch, 2014), (Sadat, Kazemi, & Farzindar, 2014)	(Elfardy & Diab, 2012), (Belgacem et al., 2010), (Zhang et al., 2013), (Lei & Hansen, 2009), (Biadsy et al., 2009), (Akbacak et al., 2011)	(Condon et al., 2010), (Salloum & Habash, 2012)	
South Iraqi								(Sawaf, 2010)	
North Iraqi								(Sawaf, 2010)	
Baghdadi								(Sawaf, 2010)	

* The Maghrebi overlaps with other listed dialects such as Moroccan. But we kept it because authors of related work were not specific.

annotations to identify Levantine, Gulf, Egyptian, Iraqi, and Maghrebi dialects. The identification of several Maghrebi dialects in addition to Syrian and Palestinian Arabic was an aspect in the cross-dialectical study proposed in (Harrat et al., 2015).

5.2 Dialect Recognition in Speech

In (Lei & Hansen, 2009), a factor analysis-based modeling technique was proposed to describe the composition of the supervector defined by the Gauss Mixture Model for dialect identification. The method utilizes knowledge types of information contained in the transcript file of the data. The addressed dialects in this work are the Emirati, the Egyptian, the Iraqi, the Palestinian, and the Syrian dialects.

In (Biadsy et al., 2009), the authors described a system that automatically identifies the Arabic dialect (Gulf, Iraqi, Levantine, Egyptian and MSA) of a speaker given a sample of his/her speech.

In (Akbacak et al., 2011), the authors studied the effectiveness of recently developed language recognition techniques based on speech recognition models for the discrimination of Arabic dialects.

In (Belgacem et al., 2010), an automatic recognition system for Arabic dialects was proposed. The analyzed dialects are Tunisian, Moroccan, Algerian, Egyptian, Syrian, Lebanese, Yemeni, Iraqi, and Gulf. The proportion of vocalic intervals and the standard deviation of consonantal intervals are analyzed using the platform Alize and Gaussian Mixture Models.

In (Zhang et al., 2013), the authors investigated variations to supervector pre-processing for dialect identification based on phone recognition-support vector machines. They studied the normalization of supervector dimensions in the pre-squashing stage, the impact of alternative squashing functions, and the N-gram selection for supervector dimensionality reduction. Addressed dialects include Iraqi, Gulf, Egyptian, and Levantine.

Speech recognition for Arabic dialects was addressed in (Kirchhoff & Vergyri, 2005), (Boujelbane et al., 2013), and (Alghamdi et al., 2008) for the Egyptian, Tunisian and Saudi dialects, respectively. In (Kirchhoff & Vergyri, 2005), the authors described the use of MSA acoustic data to improve the recognition of Egyptian conversational dialect. To simplify this task, the MSA data is vowelized automatically before combining it with the Egyptian conversational dialect data. The corpus building in (Boujelbane et al., 2013) was motivated by the need to create language models towards a speech recognition system for the Tunisian Broadcast News.

Recently, Ali et al. presented a system for Egyptian speech recognition that reduces word error rate using micro blog data(Ali, 2014).

In (Alghamdi et al., 2008), the authors aimed to present a speech database by native speakers across Saudi Arabia. The paper shows an approach that enables researchers to select samples from a population to produce a speech database where a dialect map is unobtainable. The resulted corpus was used to train a speech recognition system.

In (Iskra et al., 2004), the results of the OrienTel project were presented. This European project dealt with building telephony databases across Northern Africa and the Middle East.

6 Discussion

Table 1 summarizes the discussed research work on Arabic NLP. The columns represent the different research areas and the rows show the different covered dialects. Based on this table and on the discussions in the previous sections the following comments can be made.

1. By counting all published works, it can be seen that the research on computational linguistics for dialectal Arabic, as an alternative to Modern Standard Arabic, is emerging. Given that the different Arabic dialects are spoken by more than 390 million people in total, the total amount of research conducted in this area is still very limited.

2. The most treated dialect in Arabic NLP is the Egyptian Arabic. This may be attributed to the fact that Egypt is the country with the largest population in the Arabic world. However, such a population argument fails to explain why the Levantine Arabic has been paid relatively high attention, while the dialects of some population-rich countries such as Sudan, Morocco, and Algeria have been treated very poorly. The relatively high concentration on Levantine Arabic may be associated with geopolitical issues and the Middle-East conflict.

3. Most research work has been spent on building and annotating dialectical corpora due to the fact that dialectical Arabic is still a resource-poor language. Dialect identification and speech recognition were also researched intensively. Recall that these two tasks are frequently performed towards building language resources. While the morphology of dialectical Arabic was addressed in some papers, the syntactical analysis is almost ignored in research.

4. The selection of the geographic granularity level on which Arabic dialects are treated is not clear. The majority of related work that addresses Levantine, for instance, treats this variety as one dialect. Levantine, however, is spoken in Syria, Jordan, Lebanon, and Palestine. In each of these countries, furthermore, a lot of varieties can be identified.

From this discussion it is obvious that the research on Arabic dialects should be enhanced both on the dialect as well as on the topic level. A hierarchical scheme should be introduced to define the granularity of Arabic dialects so that researchers can be more specific in assigning their work to some dialect or dialects. The built language resources especially annotated corpora should be made available to accelerate the research in this area. More research on the syntactical analysis of Arabic dialects is required to improve the quality of related tools.

Acknowledgments

This work was partially funded by Lockheed Martin Company.

References

Abdul-Mageed, M., Diab, M., & Kübler, S. (2014, January). Samar: Subjectivity and sentiment analysis for arabic social media. *Comput. Speech Lang.*, *28*(1), 20–37. Retrieved from http://dx.doi.org/10.1016/j.csl.2013.03.001 doi: 10.1016/j.csl.2013.03.001

Abuata, B., & Al-Omari, A. (2015). A rule-based stemmer for arabic gulf dialect. *Journal of King Saud University-Computer and Information Sciences*.

Akbacak, M., Vergyri, D., Stolcke, A., Scheffer, N., & Mandal, A. (2011). Effective arabic dialect classification using diverse phonotactic models. In *Interspeech* (Vol. 11, pp. 737–740).

Al-Gaphari, G., & Al-Yadoumi, M. (2012). A method to convert sanaani accent to modern standard arabic. *International Journal of Information Science and Management (IJISM)*, *8*(1), 39–49.

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. (2014). Advances in dialectal arabic speech recognition: A study using twitter to improve egyptian asr. In *Proceedings of the international workshop on spoken language translation (iwslt).*

Al-Mannai, K., Sajjad, H., Khader, A., Al Obaidli, F., Nakov, P., & Vogel, S. (2014). Unsupervised word segmentation improves dialectal arabic to english machine translation. *ANLP 2014*, 207.

Almeman, K., & Lee, M. (2012). Towards developing a multi-dialect morphological analyzer for arabic. In *4th international conference on arabic language processing, rabat, morocco.*

Almeman, K., & Lee, M. (2013). Automatic building of arabic multi dialect text corpora by bootstrapping dialect words. In *Communications, signal processing, and their applications (iccspa), 2013 1st international conference on* (pp. 1–6).

Al-Sabbagh, R., & Girju, R. (2010). Mining the web for the induction of a dialectical arabic lexicon. In *Lrec.*

Al-Sabbagh, R., & Girju, R. (2012a). A supervised pos tagger for written arabic social networking corpora. In *Proceedings of konvens* (pp. 39–52).

Al-Sabbagh, R., & Girju, R. (2012b). Yadac: Yet another dialectal arabic corpus. In *Lrec* (pp. 2882–2889).

Bakr, H. A., Shaalan, K., & Ziedan, I. (2008). A hybrid approach for converting written egyptian colloquial dialect into diacritized arabic. In *The 6th international conference on informatics and systems, infos2008. cairo university.*

Belgacem, M., Antoniadis, G., & Besacier, L.

(2010). Automatic identification of arabic dialects. In *Lrec*.

Benajiba, Y., & Diab, M. (2010). A web application for dialectal arabic text annotation. In *Proceedings of the lrec workshop for language resources (lrs) and human language technologies (hlt) for semitic languages: Status, updates, and prospects.*

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* (pp. 53–61).

Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., ... Rambow, O. (2014). Transliteration of arabizi into arabic orthography: Developing a parallel annotated arabizi-arabic script sms/chat corpus. *ANLP 2014*, 93.

Bouchlaghem, R., & Elkhlifi, A. (2014). Tunisian dialect wordnet creation and enrichment using web resources and other wordnets. *ANLP 2014*, 104.

Boujelbane, R., BenAyed, S., & Belguith, L. H. (2013). Building bilingual lexicon to create dialect tunisian corpora and adapt language model. *ACL 2013*, 88.

Boujelbane, R., Mallek, M., Ellouze, M., & Belguith, L. H. (2014). Fine-grained pos tagging of spoken tunisian dialect corpora. In *Natural language processing and information systems* (pp. 59–62). Springer.

Brustad, K. (2000). *The syntax of spoken arabic: A comparative study of moroccan, egyptian, syrian, and kuwaiti dialects*. Georgetown University Press.

Buckwalter, T. (2004). *Buckwalter arabic morphological analyzer version 2.0. ldc catalog number ldc2004l02* (Tech. Rep.). ISBN 1-58563-3-0.

Cavalli-Sforza, V., Saddiki, H., Bouzoubaa, K., Abouenour, L., Maamouri, M., & Goshey, E. (2013). Bootstrapping a wordnet for an arabic dialect from other wordnets and dictionary resources. In *Computer systems and applications (aiccsa), 2013 acs international conference on* (pp. 1–8).

Chiang, D., Diab, M. T., Habash, N., Rambow, O., & Shareef, S. (2006). Parsing arabic dialects. In *Eacl*.

Condon, S., Parvaz, D., Aberdeen, J. S., Doran, C., Freeman, A., & Awad, M. (2010). Evaluation of machine translation errors in english and iraqi arabic. In *Lrec*.

Cotterell, R., & Callison-Burch, C. (2014). A multi-dialect, multi-genre corpus of informal written arabic. In *Proceedings of the language resources and evaluation conference (lrec)*.

Darwish, K. (2013). Arabizi detection and conversion to arabic. *arXiv preprint arXiv:1306.6755*.

Darwish, K., & Gao, W. (2014). Simple effective microblog named entity recognition: Arabic as an example. *Proc. of LREC, Reykjavik, Iceland*.

Darwish, K., & Magdy, W. (2014). *Arabic information retrieval*. Now Publishers.

Darwish, K., Sajjad, H., & Mubarak, H. (2014). Verifiably effective arabic dialect identification. *EMNLP-2014*.

Dasigi, P., & Diab, M. T. (2011). Codact: Towards identifying orthographic variants in dialectal arabic. In *Ijcnlp* (pp. 318–326).

Diab, M., Habash, N., Rambow, O., Altantawy, M., & Benajiba, Y. (2010). Colaba: Arabic dialect annotation and processing. In *Lrec workshop on semitic language processing* (pp. 66–74).

Duh, K., & Kirchhoff, K. (2005). Pos tagging of dialectal arabic: a minimally supervised approach. In *Proceedings of the acl workshop on computational approaches to semitic languages* (pp. 55–62).

Duh, K., & Kirchhoff, K. (2006). Lexicon acquisition for dialectal arabic using transductive learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (pp. 399–407).

Duwairi, R., Marji, R., Sha'ban, N., & Rushaidat, S. (2014). Sentiment analysis in arabic tweets. In *Information and communication systems (icics), 2014 5th international conference on* (pp. 1–6).

El-Beltagy, S. R., & Ali, A. (2013). Open issues in the sentiment analysis of arabic social media: A case study. In *Innovations in information technology (iit), 2013 9th international conference on* (pp. 215–220).

Elfardy, H., & Diab, M. (2012a). Aida: Automatic identification and glossing of dialectal arabic. In *Proceedings of the 16th eamt*

conference (project papers) (pp. 83–83).

Elfardy, H., & Diab, M. T. (2012b). Simplified guidelines for the creation of large scale dialectal arabic annotations. In *Lrec* (pp. 371–378).

Elfardy, H., & Diab, M. T. (2013). Sentence level dialect identification in arabic. In *Acl (2)* (pp. 456–461).

El-Fishawy, N., Hamouda, A., Attiya, G. M., & Atef, M. (2014). Arabic summarization in twitter social network. *Ain Shams Engineering Journal*, *5*(2), 411–420.

Graff, D., Buckwalter, T., Jin, H., & Maamouri, M. (2006). Lexicon development for varieties of spoken colloquial arabic. In *Proceedings of the fifth international conference on language resources and evaluation (lrec)* (pp. 999–1004).

Graff, D., & Maamouri, M. (2012). Developing lmf-xml bilingual dictionaries for colloquial arabic dialects. In *Lrec* (pp. 269–274).

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.*

Habash, N., Diab, M. T., & Rambow, O. (2012). Conventional orthography for dialectal arabic. In *Lrec* (pp. 711–718).

Habash, N., Eskander, R., & Hawwari, A. (2012). A morphological analyzer for egyptian arabic. In *Proceedings of the twelfth meeting of the special interest group on computational morphology and phonology* (pp. 1–9).

Habash, N., & Rambow, O. (2006). Magead: a morphological analyzer and generator for the arabic dialects. In *Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics.*

Habash, N., & Rambow, O. (2007). Morphophonemic and orthographic rules in a multi-dialectal morphological analyzer and generator for arabic verbs. In *International symposium on computer and arabic language (iscal), riyadh, saudi arabia.*

Habash, N., Rambow, O., Diab, M., & Kanjawi-Faraj, R. (2008a). Guidelines for annotation of arabic dialectness. In *Proceedings of the lrec workshop on hlt & nlp within the arabic world* (pp. 49–53).

Habash, N., Rambow, O., Diab, M., & Kanjawi-Faraj, R. (2008b). Guidelines for annotation of arabic dialectness. In *Proceedings of the lrec workshop on hlt & nlp within the arabic world* (pp. 49–53).

Habash, N., Roth, R., Rambow, O., Eskander, R., & Tomeh, N. (2013). Morphological analysis and disambiguation for dialectal arabic. In *Proceedings of naacl-hlt* (pp. 426–432).

Harrat, S., Meftouh, K., Abbas, M., Jamoussi, S., Saad, M., & Smaili, K. (2015). Cross-dialectal arabic processing. In *Computational linguistics and intelligent text processing* (pp. 620–632). Springer.

Harrat, S., Meftouh, K., Abbas, M., & Smaili, K. (2014). Building resources for algerian arabic dialects. *Corpus (sentences)*, *4000*(6415), 2415.

Hawwari, A., Attia, M., & Diab, M. (2014). A framework for the classification and annotation of multiword expressions in dialectal arabic. *ANLP 2014*, 48.

Hedar, A. R., & Doss, M. (2013). Mining social networks arabic slang comments. *IEEE Symposium on Computational Intelligence and Data Mining (CIDM).*

Ibrahim, H. S., Abdou, S. M., & Gheith, M. (2015). Sentiment analysis for modern standard arabic and colloquial. *arXiv preprint arXiv:1505.03105*.

Iskra, D. J., Siemund, R., Borno, J., Moreno, A., Emam, O., Choukri, K., ... others (2004). Orientel-telephony databases across northern africa and the middle east. In *Lrec.*

Jarrar, M., Habash, N., Akra, D., & Zalmout, N. (2014). Building a corpus for palestinian arabic: a preliminary study. *ANLP 2014*, 18.

Jeblee, S., Feely, W., Bouamor, H., Lavie, A., Habash, N., & Oflazer, K. (2014). Domain and dialect adaptation for machine translation into egyptian arabic. *ANLP 2014*, 196.

Jehl, L., Hieber, F., & Riezler, S. (2012). Twitter translation using translation-based cross-lingual retrieval. In *Proceedings of the seventh workshop on statistical machine translation* (pp. 410–421). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from `http://dl.acm.org/citation`

.cfm?id=2393015.2393074

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

Lei, Y., & Hansen, J. H. (2009). Factor analysis-based information integration for arabic dialect identification. In *Acoustics, speech and signal processing, 2009. icassp 2009. ieee international conference on* (pp. 4337–4340).

Maamouri, M., Bies, A., Buckwalter, T., Diab, M., Habash, N., Rambow, O., & Tabessi, D. (2006). Developing and using a pilot dialectal arabic treebank. In *Proceedings of the fifth international conference on language resources and evaluation, lrec06.*

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. *Proc. of LREC, Reykjavik, Iceland*.

Masmoudi, A., Habash, N., Ellouze, M., Estève, Y., & Belguith, L. H. (2015). Arabic transliteration of romanized tunisian dialect text: A preliminary investigation. In *Computational linguistics and intelligent text processing* (pp. 608–619). Springer.

Mohamed, E., Mohit, B., & Oflazer, K. (2012). Transforming standard arabic to colloquial arabic. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2* (pp. 176–180).

Mourad, A., & Darwish, K. (2013). Subjectivity and sentiment analysis of modern standard arabic and arabic microblogs. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 55–64).

Mubarak, H., & Darwish, K. (2014). Using twitter to collect a multi-dialectal corpus of arabic. *ANLP 2014*, 1.

Pasha, A., Al-Badrashiny, M., Altantawy, M., Habash, N., Pooleery, M., Rambow, O., ... Diab, M. (2013). Dira: Dialectal arabic information retrieval assistant. In *The companion volume of the proceedings of international joint conference on natural language processing (ijcnlp)* (pp. 13–16).

Rytting, C. A., Zajic, D. M., Rodrigues, P., Wayland, S. C., Hettick, C., Buckwalter, T., & Blake, C. C. (2011). Spelling correction for dialectal arabic dictionary lookup. *ACM Transactions on Asian Language Information Processing (TALIP)*, *10*(1), 3.

Sadat, F., Kazemi, F., & Farzindar, A. (2014). Automatic identification of arabic language varieties and dialects in social media. *SocialNLP 2014*, 22.

Sadat, F., Mallek, F., Sellami, R., Boudabous, M. M., & Farzindar, A. (2014). Collaboratively constructed linguistic resources for language vari-ants and their exploitation in nlp applications–the case of tunisian arabic and the social media. In *Workshop on lexical and grammatical resources for language processing* (p. 102).

Salloum, W., & Habash, N. (2011). Dialectal to standard arabic paraphrasing to improve arabic-english statistical machine translation. In *Proceedings of the first workshop on algorithms and resources for modelling of dialects and language varieties* (pp. 10–21).

Salloum, W., & Habash, N. (2012). Elissa: A dialectal to standard arabic machine translation system. In *Coling (demos)* (pp. 385–392).

Salloum, W., & Habash, N. (2014). Adam: Analyzer for dialectal arabic morphology. *Journal of King Saud University-Computer and Information Sciences*, *26*(4), 372–378.

Sansò, A. (2004). Med-typ: A typological database for mediterranean languages. In *Lrec.*

Sawaf, H. (2010). Arabic dialect handling in hybrid machine translation. In *Proceedings of the conference of the association for machine translation in the americas (amta), denver, colorado.*

Soltau, H., Mangu, L., & Biadsy, F. (2011). From modern standard arabic to levantine asr: Leveraging gale for dialects. In *Automatic speech recognition and understanding (asru), 2011 ieee workshop on* (pp. 266–271).

Tachicart, R., & Bouzoubaa, K. (2014). A hybrid approach to translate moroccan arabic dialect. In *Intelligent systems: Theories and applications (sita-14), 2014 9th inter-*

national conference on (pp. 1–5).

Tratz, S., Briesch, D., Laoudi, J., & Voss, C. (2013). Tweet conversation annotation tool with a focus on an arabic dialect, moroccan darija. *LAW VII & ID*, 135.

Zaghouani, W. (2014). Critical survey of the freely available arabic corpora. In *Proceedings of the workshop on free/open-source arabic corpora and corpora processing tools workshop programme* (p. 1).

Zaghouani, W., Habash, N., & Mohit, B. (2014). The qatar arabic language bank guidelines.

Zaidan, O., & Callison-Burch, C. (2011). The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Acl (short papers)* (pp. 37–41).

Zaidan, O., & Callison-Burch, C. (2012). Arabic dialect identification. *Computational Linguistics (submitted)*.

Zaidan, O., & Callison-Burch, C. (2014). Arabic dialect identification. *Computational Linguistics*, *40*(1), 171–202.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., ... Callison-Burch, C. (2012). Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 49–59).

Zhang, Q., Boril, H., & Hansen, J. H. (2013). Supervector pre-processing for prsvm-based chinese and arabic dialect identification. *IEEE ICASSP 2013*.

Zirikly, A., & Diab, M. (2014). Named entity recognition for dialectal arabic. *ANLP 2014*, 78.

Zirikly, A., & Diab, M. (2015). Named entity recognition for arabic social media. In *Proceedings of naacl-hlt* (pp. 176–185).

Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Belguith, L., & Habash, N. (2014). A conventional orthography for tunisian arabic. In *Proceedings of the language resources and evaluation conference (lrec), reykjavik, iceland.*

Zribi, I., Graja, M., Khmekhem, M. E., Jaoua, M., & Belguith, L. H. (2013). Orthographic transcription for spoken tunisian arabic. In *Computational linguistics and intelligent text processing* (pp. 153–163). Springer.

Zribi, I., Khemakhem, M. E., & Belguith, L. H. (2013). Morphological analysis of tunisian dialect. In *International joint conference on natural language processing* (pp. 992–996).