Critical description of TA linguistic resources

Asma Mekki, Inès Zribi, Mariem Ellouze, Lamia Hadrich Belguith

▶ **To cite this version:**

Asma Mekki, Inès Zribi, Mariem Ellouze, Lamia Hadrich Belguith. Critical description of TA linguistic resources. Procedia Computer Science, 2018, 142, pp.230-237. 10.1016/j.procs.2018.10.480 . hal-02869839

HAL Id: hal-02869839
https://hal.science/hal-02869839

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

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 142 (2018) 230–237

Procedia Computer Science

www.elsevier.com/locate/procedia

The 4th International Conference on Arabic Computational Linguistics (ACLing 2018), November 17-19 2018, Dubai, United Arab Emirates

Critical description of TA linguistic resources

Asma Mekki[a], Inès Zribi[a], Mariem Ellouze[a], Lamia Hadrich Belguith[a]

[a]*ANLP Research group, MIRACL, University of Sfax, Tunisia*

Abstract

This paper presents a critical description of natural language processing for Tunisian Arabic. Indeed, several linguistic resources were proposed for the three types of Tunisian Arabic (intellectualized dialect, spontaneous dialect and electronic dialect). We present different linguistic resources (corpora, lexicons and linguistic analysis tools). This study can be used as a quick reference for the scientific community working on natural language processing in general and more precisely those studying Tunisian Arabic.

© 2018 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/)
Peer-review under responsibility of the scientific committee of the 4th International Conference on Arabic Computational Linguistics.

Keywords: Tunisian Arabic; linguistic resources; critical description.

1. Introduction

The automatic processing of dialectal Arabic has taken off in recent years, particularly the Tunisian Arabic (TA) which has become the focus of research in Natural Language Processing (NLP). After the Tunisian revolution, TA has taken off in Tunisian broadcasts (resource for the intellectualized dialect), TV and radio programs (resource for the spontaneous dialect), social media comments and publications (resource for the written dialect). Today, many resources and tools are available that support three different forms of TA. Scientific community take account of TA forms particularities, which implies that most proposed resources are specific to a well-defined type of dialect.

In this paper, we will provide a critical description of linguistic resources dealing with TA. In section 2, we will present an overview of TA. Section 3 will be devoted to a critical study of the different linguistic resources (lexicons, corpora, treebank and ontology) for the TA. Then, we will present the various tools proposed for TA (section 4).

* Corresponding author. Tel.: +216-74-862-233 ; fax: +216-74-862-432 .
E-mail address: asma.elmekki.ec@gmail.com

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/)
Peer-review under responsibility of the scientific committee of the 4th International Conference on Arabic Computational Linguistics.
10.1016/j.procs.2018.10.480

2. Tunisian Arabic

Tunisian Arabic (TA), commonly known as "Darija" or "Tounsi", is the dialect spoken in Tunisia; it belongs to the dialects of North Africa. It is the native language of almost twelve million people and with the rise of its use in social networks, blogs, interviews, etc. the need to understand and analyze it has become progressively rigorous [1]. The vocabulary of TA is influenced by the alternation of a number of languages such as Modern Standard Arabic (MSA), Berber, French (FR) etc. [1]. We distinguish three types of dialects for TA: Intellectualized Dialect (ID), Spontaneous Dialect (SD) and Electronic Dialect (ED). We present these types below.

2.1. Intellectualized dialect

Educated people in conversations on radio and TV in the Tunisian broadcasts use intellectualized Dialect (ID). This dialect is a mixture of a very high rate of MSA words and TA. For this reason, sentences structure of ID is the closest to MSA compared to other types. MSA is regular and can be described as reliably using rules, despite the complexity of its linguistic system. However, mixing these two varieties (MSA and TA) in the non-regular ID increases the ambiguity of the linguistic analysis. We can also find words in MSA (e.g. السيارة ‹AlsyArħ›'the car') but with the addition of TA clitic (عالسيارة ‹ςAlsyArħ›'on the car').

2.2. Spontaneous dialect

Spontaneous Dialect (SD) presents the daily spoken dialect of Tunisians. It contains a large mass of purely dialect words. It is characterized by the presence of words from several languages like French, MSA, Berber, Italian, etc. Tunisians use words and even expressions of the French language without any phonological or morphological changes (e.g. "bonjour", "ça va", "mécanicien", etc.). Moreover, this alternation between French, TA and sometimes MSA is defined as the main feature of the SD. Thus, SD is characterized by the presence of disfluencies such as filled pauses (e.g. امّم Ammm), incomplete words (e.g. والاع ‹wAlAς›of the word الاعلام ‹AlAςlAm›'media'), etc.

2.3. Electronic dialect

With the rise of the internet and social networks, the Electronic Dialect (ED) has become the language of online communication. As well as the SD, ED is a combination of several languages (TA, SMS language, MSA, French, etc.). Most of the messages coming from social networks, blogs, etc. are written with the Latin alphabet known as Arabizi and specific numbers for some Arabic letters such as (3 for ع ‹ς›, 5 for خ ‹x›, 7 for ح ‹H›, etc.). Among the encountered examples, ED uses non-standard abbreviations (e.g. hmd, hamd, hamdallah, thank God), multiple meaning words (e.g. "hedi" that can be : هادي ‹hAdy›('calm' or 'Hadi') or هذي ‹hðy›'this'), onomatopoeia, repeated letters for emphasis, emoticons, etc. Table 1 presents three examples of sentences to distinguish between TA types.

Table 1. Examples of sentences for TA forms.

Sentences	Transliteration	Translation	Type of dialect
الحق في القراف مضمون	‹AlHq fy AlqrAf mDmwn›	'The right to strike is guaranteed'	ID
امّم انا étudiante في	‹Am AnA étudiante fy›	'Amm I am a student in'	SD
Bjr Hmd chna7welek enti :)	-	'Hi that's Ok; how are you?'	ED

3. Tunisian Arabic Resources

3.1. Lexicons

3.1.1. Bilingual word lists

Boujelbane et al. [2] developed a bilingual lexicon (MSA-TA) for verbs, nouns and function words. The construction of this lexicon is based on a manual translation of MSA (i.e. they exploited ATB [3]) to TA. The lexicon is composed of 1.500 verbs and 1.050 nouns. The entry of the lexicon is the triplet (lemma, schema and root) in MSA and its matches in TA. Similarly, a study of the different contexts of each function word was conducted.

Sadat and al. [4] build a bilingual lexicon of Tunisian nouns and verbs (TA-MSA) that contains around 1.600 TA words and their translations in MSA. It is encoded in XML. The size of lexicons presented above is very small. They cannot cover the all of TA vocabulary. Furthermore, Sadat et al.'s lexicon [4] does not only contain words in TA, it also contains foreign words (e.g. French, English, etc.).

3.1.2. Lexical databases

RIO Ontology[1] [5] is a domain ontology that was constructed from transcribed speech. Its construction method is based on hybrid method. A statistical method is used for extracting the concepts (14 concepts) by calculating the frequency of each term in TUDICOI corpus [6]. High frequency terms are used as domain concepts. A linguistic method is also used for the identification of semantic relations (25 semantic relations) between concepts. It is composed of only 14 concepts and 387 concept instances.

Bouchlaghem et al. [7] exploited a 32.848 words to create a TunDiaWN wordnet for TA. After collecting the corpus, they extracted TA words preserving useful data. Next, [7] applied a clustering method (using a k-mode algorithm) that aims to group extracted TA words into meaningful clusters. Finally, TA experts validate and enrich the TunDiaWN.

Karmani et al. [8] created a Tunisian WordNet "aebWordNet"[2] that covers simple TA words. They extracted from the English-TA dictionary "Dictionary of the Peace Corps"[3]" 5.133 words in order to create a first version of the WordNet. Next, they automatically enrich this lexicon by a derivational lexicon based on 1.507 roots. Finally, Karmani et al. generate sets of cognitive synonyms (synsets) by projecting Princeton Wordnet PWN [4] synsets to TA. It counts 18.209 synsets. The main weakness of these resources lies in their size, which is relatively small. Likewise, we have found some mistakes that are present in [8] lexicon (e.g. errors in some grammatical categories annotation).

Table 2. TA lexicon list.

Lexicon	Words	Availability	Normalization
Boujelbane et al. [2]	2.550 entries	*[5]	-
Sadat et al. [4]	1.600	no	-
RIO Ontology	1.438	yes	OTTA [9]
TunDiaWN	32.848	no	-
aebWordNet	8.279 lemmas	yes	-

[1] https://sites.google.com/site/marwagraja/resources
[2] https://github.com/NadiaBMKarmani/aebWordNet-Lexicon
[3] https://archive.org/details/ERIC_ED183017
[4] http://wordnet-rdf.princeton.edu/
[5] *: resource or tool is available with mail request to first-author.

3.2. Corpora

3.2.1. Spoken corpora

Raw text corpora. TUDICOI[6] (TUnisian DIalect COrpus Interlocutor) [6] is composed of 1.825 dialogues composed of 12.182 statements of 1.831 users who ask for railway information services (e.g. train schedule, train destination, tariffs etc.). It contains 52 hours of audio recordings. It follows OTTA orthographic convention [9].

TARIC [7](Tunisian Arabic Railway Interaction Corpus) [10] consists of 20 hours of transcribed speech using Transcriber[8]. It is composed of 4.662 dialogues with 18.657 statements. TARIC contains 71.684 words and it is transcribed in accordance with the transcription convention "CODA-TUN" [11]. The vocabulary used in both TUDICOI and TARIC is of a restricted domain (railway services). They present only a small part of the TA vocabulary that makes them inadequate corpora for modeling TA.

Boujelbane et al., [12] transcribed 5 hours and 20 minutes of recordings using Transcriber tool. These recordings come mainly from a Tunisian TV channel. It is composed of 37.964 words. 12.207 words are from a TV News program with 21.4% TA words and 25.757 words from political debate broadcasts. TA words represent approximately 32.1% of this corpus. This is the smallest corpus in the presented raw text corpora.

Table 3. Spoken raw text corpora list.

Corpus	Words	Availability	Normalization
TUDICOI [6]	21.682	yes	OTTA
TARIC [10]	71.684	yes	CODA-TUN
Boujelbane et al. [12]	37.964	*	CODA-TUN

Miscellaneous Annotated corpora. STAC[9] (Spoken Tunisian Arabic Corpus) [13] is a corpus for spontaneous TA. It is transcribed and annotated according to OTTA and CODA-TUN conventions. This corpus is composed of 4 hours and 50 minutes of audio recording collected from different TV channels and radio stations. It is composed of 97,20% words in TA, 0,37% in MSA and 2,43% words in French. STAC includes a morphosyntactic and disfluencies annotations. It is a limited corpus (42.388 words) that treats only a part of SD (i.e. radio broadcasts).

3.2.2. Social media corpora

Raw text corpora. Sadat et al. [14] manually collected 3.843 sentences from forums and blogs as part of a project to identify Arabic dialects (Tunisian, Algerian, Egyptian, Iraqi, Sudanese, etc.). This corpus was manually segmented to coherent sentences. An information of the source and the date is summarized to each sentence.

Masmoudi et al. [15] collected 70.861 messages in Arabizi from SMSs (108 messages, 1.645 words), Facebook (70.237 messages, 864.935 words) and YouTube (516 messages, 4.324 words). The validation of TA messages is performed with experts. They manually transliterated and normalized 3.500 Arabizi words to Arabic script.

Table 4. Social media raw text corpora list.

Corpus	Words	Type of character	Availability	Normalization
Sadat et al. [14]	18.199	Arabic and Latin	no	-
Masmoudi et al. [15]	870.904	Arabic and Latin	*	CODA-TUN

[6] https://sites.google.com/site/marwagraja/resources
[7] https://sites.google.com/site/masmoudiabir/res
[8] A tool for segmenting, labeling and transcribing speech : http://trans.sourceforge.net/
[9] https://sites.google.com/site/ineszribi/ressources/corpus

Annotated corpora. TSAC[10] (Tunisian Sentiment Analysis Corpus) [16] was collected from Facebook comments that are written on official pages of Tunisian radios and TV channels. TSAC contains 17.000 comments that are manually divided to positive (8.215 comments, 63.874 words) and negative (8.845 comments, 49.322 words) polarities.

Younes et al. [17] collected 43.222 Arabizi messages from different sources as (Facebook, Mini-project of a course, Google forms, etc.). They classified comments to dialect or non-dialect classes. Later, Younes et al. [18] have enriched the corpus. They extracted 73.024 messages.

Table 5. Social media annotated corpora list.

Corpus	Words	Type of character	Annotation type	Availability	Normalization
TSAC	113.196	Arabic and Latin	Sentiment	yes	no
Younes et al. [17]	420.897 160.418	Latin Arabic	Dialect / not dialect	*	no

3.2.3. Intellectualized

Syntactically Annotated corpora. Mekki et al. [19] have syntactically annotated the Tunisian constitution [20], which contains 12.378 words. They have been following the CODA-TUN convention [11] for normalizing its orthography. Then, they segmented some long sentences and tokenized the words. The preprocessed corpus was analyzed syntactically by the Stanford parser of MSA [21]. The final step in creating TTB[11] (Tunisian Treebank) is to fix annotation errors made by the MSA Stanford parser and validate it by experts. Mekki et al. [19] have followed these steps to syntactically annotate 1.072 sentences of STAC corpus.

Multi dialect type. McNeil et al. [22] have created a corpus for TA collected from different sources (e.g. Web, TV drama, internet forums, literature, conversation, etc.). Currently, this corpus regroups 2.006 texts (885.017 words). It contains sentences and texts, which are integrally in MSA and/or in French.

Bouchlaghem et al. [7] built MultiTD (Multi-source Tunisian Dialect corpus) collected from several resources: social media (comments and statuses on Twitter (10.249 words), Facebook (7.470 words), and TripAdvisor (3.258 words) and other sources (e.g. theater writings, (11.871 words)). This corpus was used to build the wordnet TunDiaWN. These multi dialect corpora are not normalized, as they does not contain any type of annotation.

Table 6. Multi dialect type corpora list.

Corpus	Words	Availability	Normalization
Tunisiya [12]	885.017	yes	no
MultiTD	32.848	no	no

4. Language Analysis Tools

4.1. Orthographic Analysis

4.1.1. Normalization

Boujelbane et al. [23] have developed COTA (Conventionalized Tunisian Arabic orthography) a normalization system based on the CODA-TUN convention [11]. This system is based on an hybrid approach. Words with more than one spilling variation will be normalized using a KNN model; otherwise, a linguistic approach will be applied.

[10] https://github.com/fbougares/TSAC
[11] https://sites.google.com/site/asmamekkisite/ressources
[12] http://www.tunisiya.org/

The statistical method determines for each character the action to do: change, deletion, or addition of another character. This approach is composed of two main techniques: (1) lexicon lookup and (2) patterns application. COTA achieves an 86,6% of accuracy, but it does not treat ED, which contains, generally, the highest number of orthographic errors.

4.1.2. Transliteration

Masmoudi et al. [15] applied a rule-based method for transliterating ED. For every Arabizi word, all possible transliterations are proposed and it is up to the user to choose the correct answer. [15] provides an output that follows the spelling convention CODA-TUN. If none of the proposals is correct, the user must choose the closest proposition. This system presents 92% of agreement for words of Arabic origin and 89% of agreement for foreign words.

Younes et al. [24] proposed a method for the automatic transliteration of Arabizi into Arabic script. The proposed method is based on a first-order Hidden Markov Model. The training corpus used consists of a lexicon comprising 19.763 distinct entries [17]. It consists of a list of manually labeled Arabizi words (i.e. an Arabic letter is assigned to each Latin letter). The evaluation results are 85% of correctly transliterated characters and 53% of fully transliterated words. The output of [24] system is not normalized. It does not follow any orthographic convention.

TACA (Tunisian Arabic Chat Alphabet) machine transliteration [25] is essentially based on two parts: Fistly, they use a training corpus to align Latin letters and Arabic letters graphemes that will use, afterwards, to generate transliteration rules. Secondly, they identify the possible graphemes of each input word. [25] generate all possible transliterations. Thereafter, to validate the right transliteration, Karmani et al. apply aebWordNet [8] and the morphological analyzer [26]. The evaluation results present 81.99% of character accuracy and 82.8% of word accuracy.

The training corpus contains an absolutely miner number of words (500 words). In addition, aebWordNet has a limited size and it is not normalized, which complicates the transliteration task. Different evaluation metrics are presented for each transliteration system. So, we could not compare the presented results.

4.2. Sentence boundary detection

Zribi et al. [27] proposed three different methods for sentence boundary detection. The first method uses 23 contextual rules built manually that are essentially based on punctuation, lexical elements (e.g. conjunctions) and two simple prosodic features. The second method (based on PART classifier) aims to classify words into four classes according to their position in a sentence. The third method mixes the results of the linguistic and statistical methods. The statistical method gives the best results with F-measure equal to 82.1%. This tool is dedicated to a specific part of SD.

4.3. Morphological Analysis and POS Tagging

4.3.1. Morphology

Hamdi et al. [28] adapted the rule-based morphological analyzer MAGEAD [29] to TA. For the first step, they defined, for each Tunisian scheme, a new Morphological Behavior Class (MBC) in the hierarchy. Then they define new abstract morphemes (e.g. enclitic negation), as well as concrete morphemes that correspond to them. The third stage concerns phonological and orthographic rules specific to the TA. MAGEAD-DT [28] treats only verbs.

Zribi et al. [30] proposed a method based on five steps to create "Al-Khalil-TUN" morphological analyzer. First, they created a lexicon for TA. Then, Zribi et al. [30] integrated it into "Al-Khalil" morphological analyzer for MSA. The third step is the update of word segmentation rules as well as the list of affixes and clitics. They end up by making a set of modifications that making "Al-Khalil-TUN" able to resolve the different specificities of SD with 88.86% of F-measure.

Karmani et al. [26] developed a linguistic tool for morphological analysis of TA. It is composed of 22 rules, lexical dictionary and aebWordNet [8]. In fact, since the system does not process the standardized TA, it does not only increase the ambiguity of the system but it also increases the number of facts enormously. In addition, this system was evaluated (97,8% decision, 63,29% precision for morphemes decomposition and 86,19% precision for labelling) with a test corpus composed of only 1.000 words which is a limited number.

4.3.2. POS Tagging

Boujelbane et al. [31] retrained the POS tagger Stanford . They used the translated corpus from MSA to TA which is generated from an automatic translation of ATB. For the test, they used a transcript of the political debates for

TA. This system offers an accuracy of 78.5%. Boujelbane et al. [31] have used a translated corpus from MSA to TA. Although, the used resources will affect the proposed systems.

Hamdi et al. [32] proposed to exploit MSA resources for tagging TA. They convert each TA sentence into an MSA lattice by analyzing them morphologically with "MAGEAD-DT" [28] to generate for each word much analysis. The proposed root and pattern for a given word which are then translated into their MSA equivalents. Afterwards, this lattice passes through a disambiguation step, which makes it possible to transform the words into pseudo-MSA. This step identifies the best path in the lattice to be analyzed by the POS Tagger. This system presents an accuracy of 89%. The morphological analyzer MAGEAD-DT deals only with verbal forms which is a disadvantage of this method.

Zribi et al. [33] proposed TAMDAS (Tunisian Arabic Morphological DisAmbiguation System) based on the output of "Al-Khalil-TUN". They tested three different techniques (RIPPER, PART and SVM). RIPPER and PART are two classifiers that use the technique of induction of predictive rules. The third technique SVM represents linear classifiers. [33] proposed to use a Bigram classifier which aims to find a single correct analysis for a given word. The proposed method was also used by [32]. However, the solution that has the highest frequency of appearance in the training corpus is not always the right one. [33] offers an accuracy of 87.32%.

4.4. Syntax and Parsing

Mekki et al. [19] adapted Stanford parser in order to create a syntactic analyzer for the TA. They parameterized the MSA attributes in favor of TA. [19] integrated TTB [19] into "Stanford-TUN". It generates a statistical model capable of parsing sentences in two dialects types (ID and SD). The F-measure result reaches 74.6%. This system does not use a specific grammar for TA. Thus, TTB is not rich enough and it cannot cover all the possible examples.

Table 7. TA linguistic analysis tools.

Analysis	**Tool**	**Type of dialect**	**Availability**
Normalization	COTA	ID and SD	*
Transliteration	Masmoudi et al.	ED	no
	Younes et al.	ED	no
	TACA	ED	no
SBD	STAr-TUN	SD	*
POS tagger	Boujelbane et al.	ID	*
	Hamdi et al.	ID	no
	TAMDAS	SD	yes
Morphological analysis	MAGEAD-DT	ID	no
	Al-Khalil-TUN	ID and SD	yes
	Karmani et al.	ID	no
Syntactic parsing	Stanford-TUN	ID and SD	yes

5. Conclusion

For the last several years, there are increasing efforts to create resources and tools for TA. These resources can only be in favor of the automatic processing of TA. In this survey, we sought to examine and criticize the different resources and tools proposed for TA. By analyzing the various published works, we checked out that the research efforts dealing with TA treat specially the basic language analyses. In this paper, we itemized 16 resources (2 lexicons [2, 4], 1 ontology [5], 2 wordnets [7, 8], 12 corpora [6, 7, 10, 12, 13, 14, 15, 16, 17, 18, 20, 22] and 1 treebank [19]) with 13 available resources. Among them, we find 4 resources [7, 8, 10, 13] morphologically annotated and POS tagged. [19] contains morphosyntactic and syntactic information. We detailed also 12 tools (orthographic analysis [23, 15, 24, 25], sentence boundary detection [27], morphological analysis [28, 30, 26], POS tagging [31, 32, 33], syntactic analysis [19]). As presented above, resources and tools can be categorized according to dialect type (ID, SD and ED). In contrast, the performance of these developed resources can decrease considerably when tested with a corpus dealing with another form of TA. Finally, we should mention that the proposed critical evaluation does not affect the value of the offered works.

References

[1] S. Mejri, M. Said, I. Sfar, Pluringuisme et diglossie en Tunisie, Synergies Tunisie 1 (2009) 53–74.

[2] R. Boujelbane, M. E. Khemekhem, S. BenAyed, L. H. Belguith, Building bilingual lexicon to create Dialect Tunisian corpora and adapt language model, in: Proceedings of the Second Workshop on Hybrid Approaches to Translation., 2013.

[3] M. Maamouri, A. Bies, T. Buckwalter, The Penn Arabic Treebank: Building a large scale annotated Arabic corpus, in: In NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt, 2004.

[4] F. Sadat, F. Mallek, R. Sallemi, M. M. Boudabous, A. Farzindar, Collaboratively Constructed Linguistic Resources for Language Variants and their Exploitation in NLP Applications : the case of Tunisian Arabic and the Social Media, in: the Workshop on Lexical and Grammatical Resources for Language Processing, 2014, pp. 102–110.

[5] J. Karoui, M. Graja, M. M. Boudabous, L. Hadrich Belguith, Semi-automatic Domain Ontology Construction from Spoken Corpus in Tunisian Dialect: Railway Request Information, International Journal of Recent Contributions from Engineering, Science & IT (iJES) 1 (2013) 35–38.

[6] M. Graja, M. Jaoua, L. H. Belguith, Discriminative Framework for Spoken Tunisian Dialect Understanding, in: Proceedings of SLSP 2013, Vol. 7978 of Lecture Notes in Computer Science, Springer, Tarragona, Spain, 2013, pp. 102–110.

[7] R. Bouchlaghem, A. Elkhlifi, R. Faiz, Tunisian dialect Wordnet creation and enrichment using web resources and other Wordnets, in: EMNLP Workshop on Arabic Natural Langauge Processing, 2014, pp. 104–113.

[8] N. K. Moussa, H. Soussou, M. A. Alimi, Tunisian Arabic aeb Wordnet: Current State and Future Extensions, in: First International Conference on Arabic Computational Linguistics (ACLING), 2015, pp. 3–8.

[9] I. Zribi, M. Graja, M. E. Khemakhem, M. Jaoua, L. H. Belguith, Orthographic Transcription for Spoken Tunisian Arabic, in: A. Gelbukh (Ed.): CICLing 2013, Part I, LNCS 7816., 2013, pp. 153–163.

[10] A. Masmoudi, M. E. Khmekhem, Y. Esteve, L. H. Belguith, N. Habash, A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition, in: Proceedings of LREC'14, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014.

[11] I. Zribi, R. Boujelbane, A. Masmoudi, M. Ellouze, L. H. Belguith, N. Habash, A Conventional Orthography for Tunisian Arabic, in: Proceedings of LREC'2014, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 2355–2361.

[12] R. Boujelbane, M. Ellouze, F. Béchet, L. Belguith, 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. 2. Traitement automatique du langage parlé 55 (2014) 73–96.

[13] I. Zribi, M. Ellouze, L. H. Belguith, P. Blache, Spoken Tunisian Arabic Corpus "STAC": Transcription and Annotation, Research in computing science 90.

[14] F. Sadat, F. Kazemi, A. Farzindar, Automatic identification of arabic dialects in social media, in: Proceedings of the First International Workshop on Social Media Retrieval and Analysis, SoMeRA '14, ACM, New York, NY, USA, 2014, pp. 35–40.

[15] A. Masmoudi, N. Habash, M. Ellouze, Y. Estève, L. H. Belguith, Arabic Transliteration of Romanized Tunisian Dialect Text: A Preliminary Investigation, in: Proceedings of CICLing 2015, Part I, Cairo, Egypt, 2015, pp. 608–619.

[16] S. Mdhaffar, F. Bougares, Y. Eve, L. Hadrich-Belguith, Sentiment Analysis of Tunisian Dialect: Linguistic Resources and Experiments, in: Proceedings ofThe Third Arabic Natural Language Processing Workshop (WANLP), Valencia, Spain, 2017, pp. 55–61.

[17] J. Younes, E. Souissi, A quantitative view of Tunisian dialect electronic writing, in: 5th International Conference on Arabic Language Processing, Oujda, Morocco, 2014.

[18] J. Younes, H. Achour, E. Souissi, Constructing Linguistic Resources for the Tunisian Dialect Using Textual User-Generated Contents on the Social Web, in: Current Trends in Web Engineering - 15th International Conference, ICWE 2015 Rotterdam, The Netherlands, 2015, pp. 3–14.

[19] A. Mekki, I. Zribi, M. Ellouze, L. H. Belguith, Syntactic Analysis of the Tunisian Arabic, in: LPKM 2017, Kerkennah (Sfax), Tunisia, 2017.

[20] S. Klibi, S. Hamraoui, S. Ben Abda, C. Gaddes, F. Horcheni, A. Maalla, La constitution Tunisienne., Tunisia., 2014.

[21] S. Green, C. Manning, Better Arabic parsing: Baselines, evaluations, and analysis, in: COLING '10 , 2010, pp. 394–402.

[22] K. McNeil, M. Faiza, Tunisian Arabic Corpus: Creating a Written Corpus of an Unwritten Language, in: Workshop on Arabic Corpus Linguistics (WACL) Lancaster University, 2011.

[23] R. Boujelbane, I. Zribi, S. Kharroubi, M. Ellouze, An Automatic Process for Tunisian Arabic Orthography Normalization, in: HrTAL, 2016.

[24] J. Younes, A. Hadhemi, E. Souissi, A Hidden Markov Model for the Automatic Transliteration of Romanized Tunisian Dialect, in: 2nd International Conference on Arabic Computational Linguistics, 2016.

[25] N. Karmani, H. Soussou, A. M. Alimi, Tunisian Arabic Chat Alphabet Transliteration Using Probabilistic Finite State Transducers, The International Arab Journal of Information Technology (IAJIT) 16.

[26] N. B. Karmani, H. Soussou, A. M. Alimi, Intelligent Tunisian Arabic morphological analyzer, 2017.

[27] I. Zribi, I. Kammoun, M. Ellouze, L. H. Belguith, P. Blache, Sentence boundary detection for transcribed Tunisian Arabic, in: Proceedings of the 12th Edition of the Konvens Conference, Bochum, Germany, 2016.

[28] A. Hamdi, R. Boujelbane, N. Habash, A. Nasr, The Effects of Factorizing Root and Pattern Mapping in Bidirectional Tunisian - Standard Arabic Machine Translation, in: MT Summit 2013, France, 2013.

[29] N. Habash, O. Rambow, MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects, in: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, Australia, 2006, pp. 681–688.

[30] I. Zribi, M. E. Khemakhem, L. H. Belguith, Morphological Analysis of Tunisian Dialect, in: IJCNLP 2013, Nagoya, Japan, 2013, pp. 992–996.

[31] R. Boujelbane, M. Mallek, M. Ellouze, L. H. Belguith, Fine-Grained POS Tagging of Spoken Tunisian Dialect Corpora, in: NLDB, 2014.

[32] A. Hamdi, A. Nasr, N. Habash, N. Gala, POS-tagging of Tunisian Dialect Using Standard Arabic Resources and Tools, in: Proceedings of the Second Workshop on Arabic Natural Language Processing, Beijing, China, 2015, pp. 59–68.

[33] I. Zribi, M. Ellouze, L. H. Belguith, P. Blache, Morphological disambiguation of Tunisian dialect, Journal of King Saud University - Computer and Information Sciences 29 (2) (2017) 147 – 155, Arabic Natural Language Processing: Models, Systems and Applications.