Contributions to the automatic processing of the user-generated Tunisian dialect on the social web

Jihene Younes* and Hadhemi Achour

ISGT,
Université de Tunis,
LR99ES04 BESTMOD,
2000, Le Bardo, Tunisia
Email: jihene.younes@gmail.com
Email: Hadhemi_Achour@yahoo.fr
*Corresponding author

Emna Souissi

ENSIT,
Université de Tunis,
1008, Montfleury, Tunisia
Email: emna.souissi@ensit.rnu.tn

Ahmed Ferchichi

ISGT,
Université de Tunis,
LR99ES04 BESTMOD,
2000, Le Bardo, Tunisia
Email: Ahmed.Ferchichi@gmail.com

Abstract: With the growing use of social media in the Arab world, Arabic dialects are rapidly spreading on the web, leading to a growing interest from NLP researchers. These dialects are however, still under-resourced languages which is a major obstacle to their study and processing. In this paper, we focus on the automatic processing of the user-generated Tunisian dialect (TD) on the social web and propose an approach that aids to automatically generate TD language resources. This approach exploits the large amounts of textual productions on the social web to extract and generate dialectal content. It is based on two main NLP components, namely the TD identification and the TD transliteration. A machine learning approach using conditional random fields is proposed for implementing these two components and reached an accuracy of 87.45 for the TD identification and 90.49 for the automatic generation of dialectal contents by transliteration.

Keywords: Tunisian dialect; TD; language resources; LR; corpora; lexica; identification; transliteration; natural language processing; NLP; machine learning.

Reference to this paper should be made as follows: Younes, J., Achour, H., Souissi, E. and Ferchichi, A. (2020) 'Contributions to the automatic processing of the user-generated Tunisian dialect on the social web', *Int. J. Computational Intelligence Studies*, Vol. 9, Nos. 1/2, pp.33–51.

Copyright © 2020 Inderscience Enterprises Ltd.

Biographical notes: Jihene Younes is a PhD student at the ISG Tunis, University of Tunis, Tunisia. She received her Master's in Computer Science, from the ENSIT, University of Tunis, Tunisia. Her current research interests include the automatic processing of the Tunisian dialect.

Hadhemi Achour is an Assistant Professor and teaching Computer Science at the ISG Tunis, University of Tunis, Tunisia. She received her PhD in Computer Science at the University of Paris 7 in France. Her doctoral research was conducted at the France's National Scientific Research Centre (CNRS). Her main research interests are related to natural language processing and its applications, including Arabic language processing. She participated in several European projects and in ALECSO coordinated studies and research projects. She also the co-chaired and was part of international conferences program committees.

Emna Souissi is an Assistant Professor and teaching Computer Science at the ENSIT, University of Tunis, Tunisia. She holds a PhD in Computer Science from the University of Paris 7, France. Her research interests are mainly related to the field of natural language processing (NLP) and its applications, with a focus on the Arabic NLP. Her PhD research was conducted within the CNRS. In this context, she has participated in several European and Canadian projects. She is currently conducting research on the treatment of Arabic dialects and mainly Tunisian. She has also been interested in e-learning and participated in projects led by Virtual University of Tunis. In this context, she was a member of the national steering group of a Master's on E-Learning, which was part of the Swiss COSELEARN Program.

Ahmed Ferchichi is a Computer Science teacher at the ISG Tunis (University of Tunis), where he held the position of Director of Academic Affairs during the period 2000–2003. Since 2012, he is an Associate Professor at the FSJEG of Jendouba (University of Jendouba). In 2018, he is appointed a member of the board of the University of Jendouba. His research interests focus on improving the teaching of programming and software engineering. Currently, he is interested in the specification of so-called smart applications. He co-founded and chaired several international conferences.

1 Introduction

In recent years, we witnessed a swift development of communication technology. Written communication thrived off from advancing technology, and nowadays, we are well aware that social media websites are very widely used means of communication on the internet. This evolution has led to a new form of informal language which rarely conforms to grammatical and orthographical rules (Habash et al., 2012). Practitioners of this new language find it faster and easier to use abbreviations, acronyms, digits, etc., in their language productions on the social web.

In Arabic countries, this electronic language is not practiced necessarily using the Modern Standard Arabic (MSA). The written communication of Arab social web users is indeed, largely based on the various Arabic dialects (ADs) that are naturally spoken by Arab populations depending on their regions and countries (Zaidan and Callison-Burch, 2014). They are in fact, informal variants of the Arabic language that differ from the MSA in several aspects and which are increasingly being used on the social web.

Consequently, the language used by Arab social web users merges the linguistic characteristics of the ADs and the informality of the electronic language at the same time. This is why many researches in the natural language processing (NLP) field have been carried out in recent years, in order to deal with the ADs, especially when they are written on the social web. Some of them focused on constructing corresponding language resources (LRs) (Diab et al., 2000; Zaidan and Callison-Burch, 2011; Meftouh et al., 2012; Boujelbane et al., 2013a; Bouchlaghem et al., 2014; Younes and Souissi, 2014), others tackled their automatic processing (Al-Sabbagh and Girjuh, 2012; Chalabi and Gergers, 2012; Darwish, 2014; Masmoudi et al., 2015; Younes et al., 2016; Aridhi et al., 2017).

In this work, we focus on the written form of the Tunisian dialect (TD) that has emerged and spread on the social web. Like most of ADs, the TD lacks available consistent LRs useful for its automatic processing and remains an under-resourced language. LRs, as defined by the European Language Resource Association (ELRA[1]), refer to any computer readable speech or language data and descriptions that are useful for constructing or evaluating NLP algorithms and systems. LRs include written and spoken corpora, lexica, dictionaries, ontologies, etc., and their availability is thus, essential for language studies and NLP researches (El-Haj et al., 2014). As TD is constantly gaining ground in the Tunisian users' social exchanges and currently spreading on the web (blogs, forums, etc.), the development of various TD LRs and NLP tools (such as morphological analysers, POS taggers, information extraction tools, etc.) becomes necessary and essential for several NLP applications.

The main purpose of our work is to propose an approach aiming to aid the automatic construction of TD LRs, namely TD corpora, lexica and dictionaries, from the informal textual productions that are generated on the social web, useful for any NLP research work dealing with this dialect. This approach is based on two essential NLP components which are the TD Identification and the TD transliteration. Besides their usefulness for automatically creating TD LRs, the TD identification and transliteration can be crucial steps for many NLP applications involving the TD language. Language identification indeed, may be very useful for multiple applications such as machine translation, information retrieval, text-to-speech applications, question answering, etc. It may in fact be considered as a pre-processing step of some larger NLP process (Tromp and Pechenizkiy, 2011). As for the TD transliteration task, it would also be very useful for developing 'multi-form' (Latin and Arabic writings) information retrieval systems when dealing with various TD documents (Younes et al., 2016).

The remainder of this paper is organised as follows: in Section 2, we present the TD language. A brief review of works related to the TD automatic processing is presented in Section 3. Section 4 is devoted to the presentation of our approach to automatically generate TD LRs, as well as the two main NLP components needed to implement the proposed approach. In Section 5, we present the various conducted experiments and the main obtained results.

2 The TD

The TD is the mother tongue spoken by Tunisians. Today, it is widely used in electronic written form on the web, such as social networks, forums and blogs.

The swift development of communication technology engendered the spread of the TD on the internet. Encompassing almost the same characteristics of the SMS language and that of the internet, TD is nowadays marked by an ample richness and incorporates words from several origins. It may be written using both Latin and Arabic transcription. Generally, Latin script is the most widely used on the social web, according to the study of Younes and Souissi (2014) who showed that more than 60% of a built corpus from social media is written in Latin script, and to that of Younes et al. (2015) who found that 81% of the TD productions on the social web are written using the Latin alphabet.

The TD's alphabetical system competes with a more 'phonetic' type of writing, traces of a 'consonant' orthography (deleted vowels), and the unconventional use of digits (Kobus et al., 2008). This variability is also the result of a communication style, which allows many deviations from orthographic and grammatical prescriptions. Indeed, the TD is mainly, a spoken language that doesn't conform to any conventions or rules when written. The majority of the social media users choose to communicate with their mother tongue and favour writing without orthographical dependencies (Younes et al., 2015). This margin of freedom increases the number of possible transcriptions for a given word whether it's written with the Latin alphabet (LTD: Latin Tunisian dialect) or the Arabic alphabet (ATD: Arabic Tunisian TD). Table 1 shows some examples of word transcriptions based on our observation of their use by Tunisians on the social web.

Table 1 TD word transcriptions

LTD word: kteb	*ATD word:* شنوة
Meanings: *he wrote*|*book*	Meanings: *what*|*chinois Latin*
Arabic transcriptions: كتاب/كتب	Transcriptions: *chnoua*|*chnouwa*|*chnowa*|*chenwa*|*chinwa*

3 Related work

In the past few years, a growing interest has been granted to the TD automatic processing. To construct corresponding LRs, some researchers started from recordings of dialogues, conversations, radio and TV broadcasts, such as Graja et al. (2010), Masmoudi et al. (2014a, 2014b, 2014c) and proceeded to their transcription into written corpora.

Other researchers used the web and social media to extract dialectal data and build various types of resources. We can cite in this context, McNeil and Faiza (2011) who built a TD corpus as part of a project to create a TD-English dictionary. The corpus was then organised in a web application allowing basic linguistic processing (McNeil, 2015). Younes and Souissi (2014) and Younes et al. (2015) also used the social web to build various resources for TD.

Several works have been carried out to create parallel and annotated corpora, notably by Graja et al. (2013) and Zribi et al. (2015). The work of Graja et al. (2013) focused on the semantic annotation of spoken TD, using a discriminative model based on conditional random fields (CRFs). As for Zribi et al. (2015) they developed two corpora, namely a written corpus segmented in sentences, whose words are segmented and annotated by lemmas, gender, number, person, voice, labels, etc., as well as an oral corpus annotated

by the different types of disfluencies. Mdhaffar et al. (2017) collected a corpus, called Tunisian sentiment analysis corpus (TSAC) from Facebook, which they manually annotated with positive and negative polarities.

Regarding the existing TD lexicons and dictionaries, we cite the work of Boujelbane et al. (2013a, 2013b, 2014) and Boujelbane (2013) who resorted to the construction of TD-MSA bilingual lexicons, with the aim of developing a translation system for the TD into MSA. They used the Penn Arabic Treebank (Maamouri and Bies, 2004) to create this bilingual lexicon, based on the differences between TD and MSA. Hamdi et al. (2014) built a lexicon for TD deverbal nouns, with the aim of integrating it into a TD morphological and syntactic parser. Masmoudi et al. (2014a, 2014b, 2014c) built a corpus named Tunisian Arabic Railway Interaction Corpus (TARIC) by manually transcribing audio recordings, and automatically generated a dictionary of TD pronunciation named TunDPDic, using a rule-based method.

Other researchers worked on domain ontologies for the TD processing in dialogue systems (Graja et al., 2011a, 2011b; Karoui et al., 2013a, 2013b; Graja et al., 2015). Graja et al. (2011a, 2011b) used ontologies to cover the lexicon used in train stations. They used the TERMINAE methodology (Biébow and Szulman, 1999) and resorted to the TuDiCoI corpus built by Graja et al. (2010). As for Karoui et al. (2013a, 2013b), they proposed a hybrid method, combining a statistical approach for the extraction of terms and concepts and a linguistic approach for the extraction of semantic relations, for the semi-automatic construction of a domain ontology. The latter was called railway information ontology (RIO), and was constructed from the spoken TD corpus TuDiCoI (Graja et al., 2010), in order to semantically label TD statements.

Work on the construction of TD ontologies includes the Wordnet 'TunDiaWN' proposed by Bouchlaghem et al. (2014). The Wordnet was constructed from a corpus named multi-source Tunisian dialect corpus (MultiTD), collected from various sources (social networks, written plays, dictionaries, transcription of speeches, etc.). Another ontology named 'aebWordnet Synset' was proposed by Ben Moussa et al. (2014, 2015), Ben Moussa and Alimi (2015), and was modelled from the English-Tunisian Arabic bilingual dictionary 'peace corpus dictionary' of Ben Abdelkader (1977).

A Tunisian Arab Treebank was created by Mekki et al. (2017) for the syntactic analysis of the TD. They used as a corpus, the version of the Tunisian constitution written in TD.

Works on the TD identification include that of Aridhi et al. (2017) who worked on the TD transcribed in Latin alphabet and experimented using two approaches. The first was based on the N-Gram cumulative sum of internal frequencies (N-Gram CSIF) method (Ahmed et al., 2004), and the second was based on support vector machines (SVM) classification.

The transliteration task was tackled by Masmoudi et al. (2015) who resorted to the CODA standard for the TD (Zribi et al., 2014), and by Younes et al. (2016) who used a machine learning approach, based on hidden Markov models (HMM). Both works focused on the transliteration of the Latin form of the TD into Arabic script.

This brief overview of the main work carried out on building TD LRs and tools shows clearly that there is a growing interest among the NLP community in the automatic processing of the TD. We present in the Table 2 a summary of the cited works, in terms of built TD LRs.

Table 2 Constructed TD LRs: a summary of the main related work

Authors	*Year*	*Script*	*L*	*Availability*
Graja et al.	2010	A	TuDiCoI corpus: 127 dialogues; 3,403 words	√
McNeil and Faiza	2011	A	TAC corpus 2011: 400K words	√
McNeil	2015		TAC corpus 2015: 820K words	
Graja et al.	2011a, 2011b	A	Ontology of 15 concepts	
Karoui et al.	2013a 2013b	A	RIO ontology: 14 concepts, 25 relations, 387 instances	
Graja et al.	2013	A	TuDiCoI corpus: 1,825 dialogues; 21,682 words	√
Boujelbane et al.	2013a, 2013b, 2014	A	Corpus: 5 h 20 min of speech; 37,964 words	
Boujelbane	2013		Corpus: 12K words	
			Lexicon: bilingual MSA-TD	
Younes and Souissi	2014	L	Corpus: 43,222 messages	
			Lexicon: 19,763 words	
Hamdi et al.	2014	A	Lexicon: bilingual MSA-TD, 39,793 entries	
Masmoudi et al.	2014a, 2014b 2014c	A	TARIC corpus: 20 hours of speech; 71,684 words	√
			TunDPDic phonetic dictionary: 18K words.	
Bouchlaghem et al.	2014	AL	MultiTD corpus: 32,848 words	
			WordNet TunDiaWN	
Ben Moussa et al.	2014, 2015	AL	Wordnet (aebWordnet)/Synset: 18,209 entries	√
Ben Moussa and Alimi	2015			
Zribi et al.	2015	A	STAC corpus: 42,388 words – 7,788 sentences	√
Younes et al.	2015	AL	Corpus: Latin: 31,158 messages – 420,897 words	
			Corpus: Arabic: 7,145 messages – 160,418 words	
			Lexicon: Latin TD → Arabic TD: 19,763 entries	
Graja et al.	2015	A	Ontology: 18 concepts	
Masmoudi et al.	2015	L	Corpus: 70,861 messages – 870,904 words	
Younes et al.	2016	L	Dictionary; Latin-Arabic: 19,763 words	
Aridhi et al.	2017	L	Corpus: 86,940 words (annotated: TD/non-TD)	
Mdhaffar et al.	2017	AL	TSAC corpus: 17K comments (annotated: positive / negative)	√
Mekki et al.	2017	A	Corpus: 12K words – 492 sentences Treebank: 928 syntactic trees	

Table 2 shows that several TD languages resources were built as part of works on the TD processing. We should however note that the majority of these resources still have limits in terms of size, content and availability. Some of the built corpora have indeed, a limited coverage to a specific domain, such as the TuDiCoI corpus (Graja et al., 2010) and the ontologies of Karoui et al. (2013a, 2013b) that are limited to the railway station interactions between clients and staff and, therefore do not cover a broad vocabulary. TD corpora constructed by Graja et al. (2010), Zribi et al. (2013) and Masmoudi et al. (2014c) were based on a speech transcription. The followed approach results in a unique form of each transcribed word and therefore cannot cover the written form of TD, especially the one that is daily produced on the social web, which is, as shown in Section 2, rich, varied and does not conform to specific rules or standards. Other TD resources were built using existing MSA resources (Boujelbane et al., 2013b; Bouchlaghem et al., 2014). Although benefiting from MSA corpora and tools helps importantly overcoming the lack of LRs, this approach does not consider TD language productions that are not MSA derived, especially when dealing with the TD used on the social web. A lot of TD words are indeed, originated from other languages such as Berber, French, English, etc. The TD word 'rivez' is, for example, originally a French word meaning 'to revise'. Practitioners of TD changed its form and use it in their social exchanges. Therefore, dealing with the TD as an MSA derivative form only, cannot be sufficient to cover the TD language accurately. We need indeed to consider the word borrowing phenomenon and the continual evolution of the TD through continuously emerging new words.

We can also see from Table 2, that the majority of these works dealt rather with the Arabic transcription of the TD. However, the written TD using the Latin alphabet is widespread on Tunisian social networks. It even seems that TD users prefer using the Latin script for their TD textual productions as they find it is easier to access and to use (Younes et al., 2015). This form of writing, also referred to as 'Arabizi' and 'Romanised', is marked by several phenomena such as the use of digits to replace the Arabic letters with no equivalents in the Latin alphabet, abbreviations, acronyms, the use of vowels, etc., presenting thus, additional challenges and difficulties to overcome. All though this form of TD writing is more present on the social web, only few works focused on constructing LRs for the Romanised TD, namely Masmoudi et al. (2015) and Younes et al. (2015, 2016).

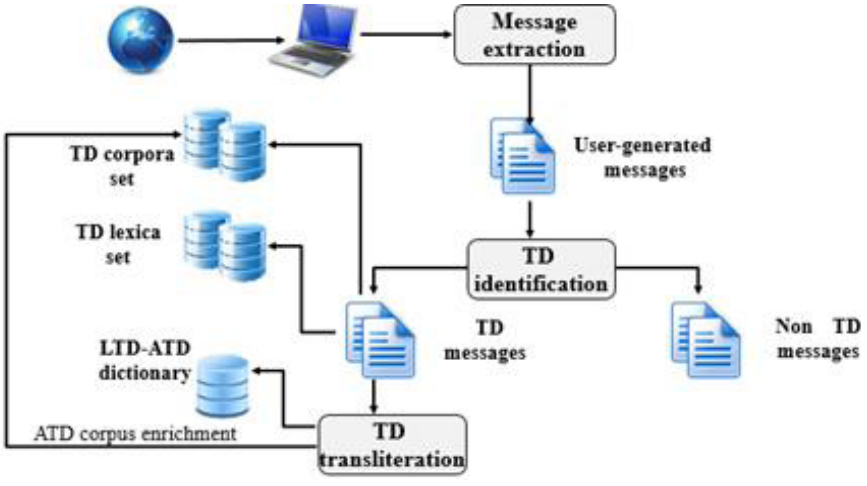
In addition, we note that, despite the efforts made in several works for the construction of various LRs for TD such as corpora, lexica, dictionaries, etc., these are still relatively limited in size (ranging from 3 to 800K words for corpora and 2 to 40K entries for lexica) and in number.

Finally, we can conclude that, despite the growing interest from the NLP community in the TD automatic processing in the last few years, the TD language still suffers, as shown in Table 2, from a lack of freely available corpora and lexica and remains an under-resourced language especially in its Latin form. This was the main motivation for the present work, which proposes an approach to large TD LRs automatic construction in both Arabic and Latin scripts.

4 Automatic generation of TD LRs

We propose in this work, a structured approach for the automatic generation of a set of LRs (TD corpora, lexica and dictionaries), that are useful for the study of the TD language and for any NLP research work dealing with the TD. This approach is based on exploiting the large amount of textual user-generated content on the social web, as well as NLP components based on appropriate NLP techniques for extracting and generating dialectal content. Figure 1 illustrates the proposed approach.

Figure 1 Proposed approach for the automatic generation of TD LRs (see online version for colours)



Considering the social web content as an initial input, informal textual productions are first extracted (mainly from Tunisian Facebook pages' posts and comments regarding this work). Since the two transcriptions (Latin and Arabic) of the TD coexist on the social web, we propose to treat the Latin transcription of the dialect, in order to propose an aid to TD LRs construction (mainly, a Latin TD corpus and lexica including both the Latin and Arabic written form of the TD). As mentioned in the previous section, the Latin transcription of the TD is indeed predominant on the web (Younes and Soussi, 2014; Younes et al., 2015).

The next step consists in identifying the TD words in these Latin textual productions which are highly multilingual, in order to construct TD corpora rich in TD words as well as TD lexica.

To generate corresponding LRs transcribed in Arabic, a phase of transliteration is included in the proposed approach, which will essentially concern the transliteration of the Romanised TD to the Arabic TD. The objective of this stage is twofold: to construct an Arabic TD lexicon, since the Arabic transcription of the TD is rather rare on the internet and on the other hand, to build a bilingual Latin Tunisian Dialect (LTD) ↔ Arabic Tunisian Dialect (ATD) dictionary.

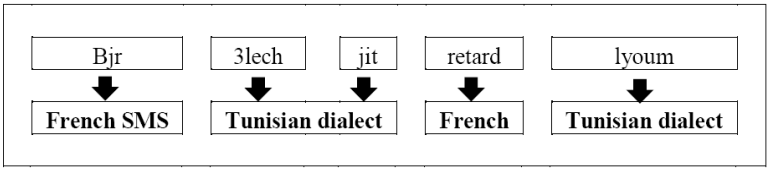
Thus, the proposed approach for the automatic generation of TD LRs includes two main steps that represent important NLP tasks: TD identification, and TD transliteration. Each of these tasks includes several difficulties that differ according to the TD transcription system (Latin or Arabic).

4.1 TD identification

The result of the message extraction step is a set of language productions that include both TD and non-TD messages. Hence, we aim in this phase to automatically identify, from the messages produced on social networks, those containing TD. The result of this stage is a set of messages allowing us to construct a corpus, composed of a large set of TD language productions, as well as a TD lexicon.

The TD identification is not a trivial task. In fact, whether in the LTD or the ATD, multilingualism is one of the most observed phenomena. Practitioners of this form of writing can introduce words from several languages (mainly French and English for the Latin script and MSA for the Arabic script), in their standard or SMS form (Younes et al., 2015). Figure 2 shows an example of a TD message transcribed in the Latin alphabet.

Figure 2 Example of a TD message



This TD message (Figure 2) is composed of five words: The first is a French word written in SMS language format. It is the abbreviation of the word ‘bonjour’ which means ‘good morning’. Followed by two TD words ‘3lech’ and ‘jit’ which mean respectively ‘why’ and ‘you came’. Then a standard French word ‘retard’, which means ‘late’. The last, is another TD word ‘lyoum’ which means ‘today’. The translation of this message into English is “good morning, why did you come late today?”

Although the multilingualism phenomenon reveals the richness of the TD, it poses, in return, a problem in the language ambiguity that complicates the process of automatic identification of the TD language (Table 3).

Table 3 Examples of TD ambiguous words

TD word	خارط		*Bard*		*Flous*	
Meaning	ATD	MSA	LTD	English	LTD	French
	Because	spirit	Cold	Poet	Money	Fuzzy

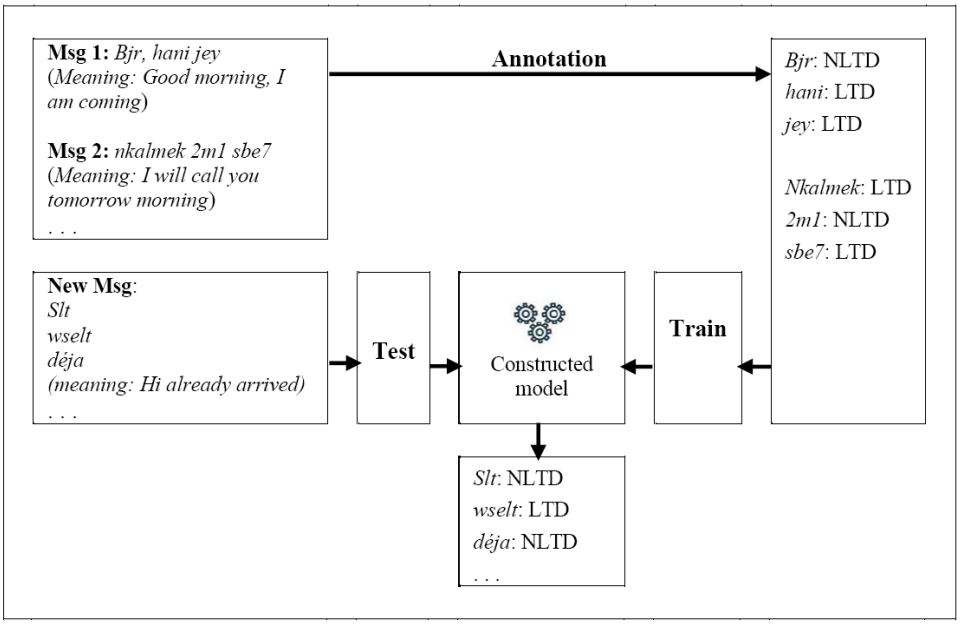
Source: Younes et al. (2015)

The difficulty lies in the automatic language identification of the extracted messages and in the decision to make if they contain ambiguous words. That is to say, how can we classify them into TD messages and non-TD messages, if they contain ambiguous words? And how can we know, for example, if the word ‘bard’ (Table 3) belongs to the LTD or the English language?

To overcome this problem, we address, in this work, the LTD identification as a sequential labelling task and propose to develop a solution based on a machine learning approach. Our objective is to automatically annotate the words composing a message as LTD or non-LTD (NLTD) words. Messages containing at least one LTD word are thus considered as textual LRs rich in dialectal content and used to construct an LTD corpus and an LTD lexicon.

As illustrated in Figure 3, the proposed word-level identification is based on a supervised learning technique that uses CRF. CRF consists of a framework for building probabilistic models to segment and label sequence data, introduced by Lafferty et al. (2001) and known to perform well with the sequence labelling tasks (Sutton and McCallum, 2012).

Figure 3 The LTD word-level identification process



A CRF defines conditional probability $P(Y \mid X)$ of label sequence Y (tags to be assigned to input words), given an input sequence X (a message as a sequence of words). A CRF on (X, Y) is specified by two vectors (Lafferty et al., 2001), 'F' and 'W'. The vector 'F' is for local features and 'W' corresponds to the weight. Each local feature can be a state: $s(y, x, i)$ or a transition: $t(y, y', x, i)$, where 'y', 'y'' are labels representing the identification tags (LTD/NLTD), 'x' an input representing an LTD word and 'i' an input position. The conditional probability distribution is defined as follows:

$$P_w(Y|X) = \frac{\exp W.F(Y, X)}{Z_W(X)} \tag{1}$$

Features depend on the inputs around the given position (Sha and Pereira, 2003). The feature vector for the input sequence 'X', and the label sequence 'Y' is presented by:

$$F(Y, X) = \sum_i f(y, x, i) \tag{2}$$

$Z_W(X)$ is defined as:

$$Z_w(X) = \sum_y \exp W.F(Y, X) \quad (3)$$

The proposed CRF model is built from a training set consisting in a set of LTD messages that are tokenised and manually annotated. The identification approach is summarised in Figure 3.

4.2 TD Transliteration

In this work, we dealt with a word-level Latin to Arabic transliteration that consists in automatically transforming a word written in Romanised TD to a word written in Arabic TD while preserving the word's pronunciation. For example, the LTD word '3asslama' which means 'hello' will be transliterated into TD Arabic script as the word 'عسلامة' □□

The difficulty of the transliteration task mainly comes from the ambiguity of the characters composing a given LTD word. In fact, many used Latin characters have two or more Arabic equivalents as well, depending on the user's preferences, or the word context. For example, the Latin consonant 't' corresponds to the Arabic letters 'ت' in 'mchit – مشيت' (I went), and 'ط' in 'tayara – طَيَّارَة' (plane). The sequence of letters 'dh' can be used to represent the Arabic letter 'ذ' in 'thkey – ذكي' (smart), the Arabic letter 'ض' in 'dh7ak – ضحك' (he laughed) or also to represent the Arabic sequences 'ده' in 'se3edhom – سا عد هم' (help them), etc.

Hence, one of the solutions to the LTD transliteration is to consider all the possible Arabic transcriptions for a Latin character, then generate the transliteration results and choose the right transcription according to the human expertise. This solution reveals the nature of the transliteration task, which can be considered as a combinatorial problem. Its automatic processing may cost a lot of time and generate irrelevant results. We thus propose solutions based on machine learning techniques, which allow us to predict the most probable transliteration for a given LTD word. As proposed in Younes et al. (2016), we consider the LTD transliteration as a sequential labelling task that consists of assigning to each Latin character composing a given Romanised TD word, a tag. This tag consists of its corresponding Arabic character, so that we can generate the equivalent Arabic TD word. An illustrative example is given in Table 4.

Table 4 Example of an LTD word transliteration ('3asslama' → '')

Characters	3	a	s	s	l	a	m	a
Tags	ع	ø	س	ø	ل	ا	م	ة

In their work, Younes et al. (2016) used and evaluated a transliteration model based on HMMs. In this work we propose a discriminative probabilistic model using CRF and show that it gives significantly better transliteration performance.

The CRF conditional probability $P(Y \mid X)$ of label sequence Y (Y is an Arabic character sequence), given an input sequence X (X is an LTD word in the form of an input Latin character sequence) is defined as stated in 4.1, where 'y', 'y'' are labels representing potential Arabic characters, 'x' an input representing a Latin character and 'i' an input position. This model is trained using a dataset consisting of a set of LTD

words that are tokenised into characters and manually annotated by assigning to each character, a tag consisting of its corresponding Arabic character.

Regarding HMM, if we represent an entered Latin text, by $W = (wi)\ 1 \le i \le n$ and a sequence of tags (Arabic letters in our case) by $T = (ti)\ 1 \le i \le n$, we compute:

$$\arg\max P(t|w) \tag{4}$$

The equation can be transformed by applying the Bayesian rule and eliminating the constant part:

$$P(t|w) = \frac{P(t)P(w|t)}{P(w)} = \frac{P(w,t)}{P(w)} \tag{5}$$

The probability of transition from one tag to the other is represented by $P(t)$ and can be computed as follows:

$$P(T = t_1 t_2 \ldots t_n) = \prod_{i=1}^{n} P\left(t_i | t_{i-n} \ldots t_{i-2} | t_{i-1}\right) \tag{6}$$

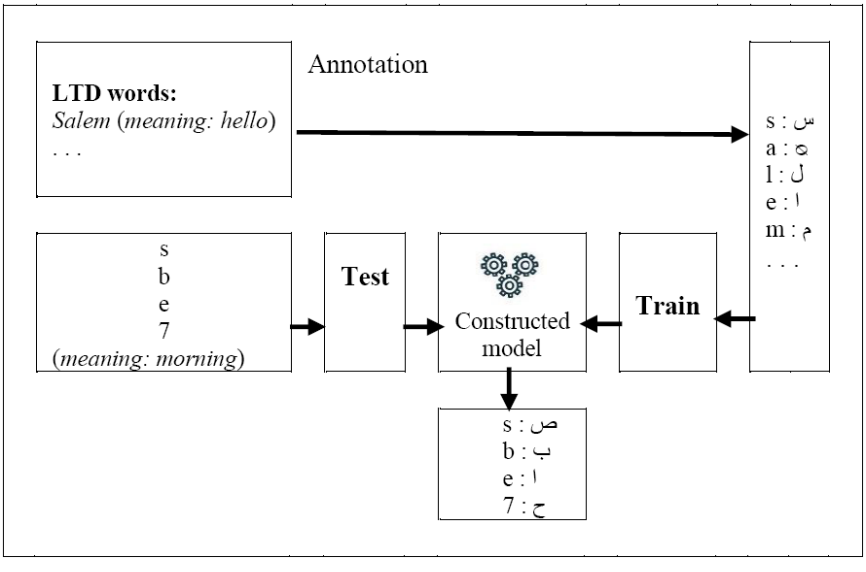
The emission probability $P(W \mid T)$ which represents the likelihood of a Latin character associated with a certain Arabic character is computed as follows:

$$P(W|T) = \prod_{i=1}^{n} P\left(w_i | t_i\right) \tag{7}$$

The HMM parameters consist of the transition and the emission probabilities. These probabilities are subsequently used to predict the best sequence of Arabic characters given a sequence of input Latin characters.

Figure 4 illustrates the transliteration process.

Figure 4 The LTD transliteration process (see online version for colours)



5 Experiments and results

5.1 Used datasets

In order to build the CRF model for the word-level TD identification, we used the corpus constructed by Younes et al. (2015), from social web. It consists of 6,079 TD messages written in the Latin script, including 60,066 words. We proceeded to the manual tokenisation and annotation of the messages since there were no available TD corpora dedicated to the identification task and to ensure the quality of the data. To each word of the message, we assigned a tag to indicate whether it belongs to TD (tag: LTD) or not (tag: NLTD). The manually annotated messages were thus used in our model, in order to subsequently be able to assign the appropriate tags to new data. Details about the Annotated corpus are shown in Table 5.

Table 5 Details of the identification TD corpus

	#messages	*#words*		
		#TD words	*#NTD words*	*Total #words*
Total corpus (100%)	6,079	45,632	14,434	60,066
Training corpus (80%)	4,863	36,974	10,818	47,792
Test corpus (20%)	1,216	8,658	3,616	12,274

As for the transliteration CRF model, it was built using the corpus of Younes et al. (2016). The corpus was composed of 19,763 LTD words, where each word was manually segmented into single characters, to which are assigned their equivalent Arabic transliterations. Details about the annotated transliteration corpus are presented in Table 6.

Table 6 Details of the transliteration TD corpus

	#LTD	*#characters*
Total corpus (100%)	19,763	129,367
Training corpus (80%)	15,810	104,202
Test Corpus (20%)	3,953	25,165

The annotation process was based on human expertise. The tag list includes the following Arabic characters: "Ø, ش, س, ز, ل, م, ن, ه, و, ؤ, ى, ي, ئ, ء, ا, إ, أ, آ, ب, ة, ت, ث, ج, ح, خ, د, ذ, ر, ف, غ, ع, ظ, ط, ض, ص, ق, ك, كز", where the null character 'ø' is used to transliterate a Latin character with no Arabic equivalent in the tag list, namely a vowel (◌َ, ◌ِ, ◌ُ), a double character (example: '3*ass*lema'/*meaning: hello*), or a character succession (example: '*ch*rit'/*meaning: I bought*). The words of the annotated corpus are then used to train the model.

5.2 Evaluation results

The results obtained from the experiments carried out on the TD identification and TD transliteration are given in Table 7.

Table 7 Evaluation results of the identification and transliteration approaches

	Accuracy	*Precision*	*Recall*	*F-measure*
TD identification	87.45	87.44	91.40	89.38
TD transliteration		*Character-level results*		
	90.49	90.49	91.10	90.80
		Word-level results		
	% correct words		78.40	

As shown in Table 7, the proposed word-level identification approach allowed us to correctly identify the language of 87% of the words composing the test set, by correctly determining if they were LTD or N LTD words. The F-score reached a value of 89.38. The observed errors are mainly related to some multilingual sequence of words. We can site as examples:

- The French coordination conjunction ‘et’ is usually used between two French words. However, in some LTD messages, the word ‘et’ can be preceded or succeeded by an LTD word. Example: ‘nemchi et narja3’ (meaning: I go and return). Although the words ‘nemchi’ and ‘narja3’ belong to the LTD, CRF does not identify them as such.
- the word ‘l’ is considered by the CRF classifier as the French preposition and identified as NLTD along with the next neighbouring word, although in some cases it refers to the abbreviation of the determinant ‘al’ in LTD. Example ‘l mra’ (meaning: the woman).

Some other errors are also due to the fact that:

- most of the words beginning with a capital letter are considered as NLTD
- most of the long words containing more than 6 letters are considered as LTD.

These errors can be reduced by introducing some features into our CRF model, such as capitalisation, word sizes, nature of some foreign words (preposition, conjunctions, etc.).

As for the tested transliteration approach, the obtained results show that 90.49% of the characters composing the test were assigned the appropriate Arabic character, which corresponds to a rate of 78.40% of words that are correctly transliterated.

In addition, and in order to compare the proposed CRF model with the bigram HMM-based approach proposed by Younes et al. (2016), we conducted additional experiments with a bigram HMM and a trigram HMM models, using the same datasets and compared them to our proposed CRF model. The obtained results presented in Table 8, show the significant improvement in performance when using CRF. The main HMM limits are basically due to the lack of dependency between the current observation and the past and future ones. This is why, resorting to discriminative models such as CRF represents an alternative, that offers much more character-level dependency than a trigram HMM, and which indeed, proved to be much more performant with the Latin to Arabic transcription of TD as we showed through the performed experiments.

As for the main transliteration errors generated by the proposed CRF model, they are mainly related to some cases of Latin character ambiguities. We can cite as examples, the letter ‘s’ that can be transliterated as the characters ‘س’ and ‘ص’, or the characters ‘t’ that can be transliterated as ‘ت’ and ‘ط’. These letters can therefore present ambiguity cases

which are challenging to solve. The transliteration depends indeed, not only on the context of the character within the word, but also on the meaning of the word within the message in which it occurs. Moreover, some TD words composed of infrequent Arabic characters like the characters 'ئ', 'إ', 'ؤ' or 'ء' generate erroneous transliterations. These characters, representing the different types of Hamza in Arabic, can rather be found in MSA written language productions. In MSA, the formal writing of the Hamza follows well-defined rules in the Arabic language, but they are very rarely used in informal dialectal writings.

Table 8 CRF vs. HMM-based TD transliteration

Transliteration model	*Accuracy*	*Precision*	*Recall*	*F-measure*
Bigram HMM	81.88	81.88	84.96	83.40
Trigram HMM	86.96	86.95	88.47	87.71
CRF	90.49	90.49	91.10	90.80

On the other hand, most of the correctly transliterated words usually include unambiguous consonants such as the characters 'b', 'l', and 'm' which have unique equivalents in the Arabic alphabet (respectively, the letters 'ب', 'ل' and 'م').

5.3 TD LR generation

In order to test the feasibility and performance of the proposed aid approach to TD LR construction, we used the two previously described components developed for the TD identification and transliteration to generate a set of TD LRs from a small sample of 500 messages, including 5,591 words and extracted from various Tunisian pages on Facebook. The results obtained are described in the following Figure 5 and Table 9.

Figure 5 Automatically generated TD LRs

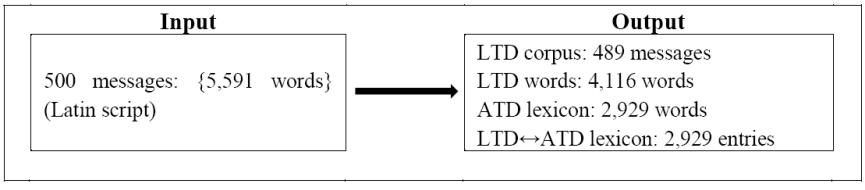


Table 9 Generated TD LRs evaluation

	Generated	*Precision*
LTD corpus	Identified LTD messages	100%
	Identified LTD words	89.77%
LTD lexicon		76.35%
LTD ↔ ATD dictionary		64.80%

The extracted LTD corpus is composed of 489 messages that are correctly identified as messages containing dialectal content. The CRF-based word-level identification also allowed us to collect 4,116 LTD words, of which 89.77% are correctly recognised as LTD words. However, the precision rate decreased to 76.35% when we deleted the

redundant words to generate the LTD lexicon. This decrease is explained by the relatively high frequency of LTD words that are correctly recognised. As for the obtained double LTD ↔ ATD dictionary that was automatically generated by transliterating the LTD lexicon, precision is of 64.80%. The error rate represents the obtained invalid pairs of ATD ↔ ATD words. We should however note that these errors are very often due to the identification step errors. Indeed, the incorrectly identified LTD words in the LTD lexicon (which exceed 20% of the total number of words), lead to invalid transliterated words (i.e., Arabic transcriptions that do not correspond to correct ATD words).

As we can see, the proposed approach in this work provides significant assistance in constructing TD LRs. However, the generated LRs quality closely depends on the efficiency of the identification and the transliteration components. It is therefore, important to further study and improve the two approaches in order to mitigate the identification and the transliteration errors before proceeding to a generation of large size TD LRs.

6 Conclusions

We presented in this paper, a set of contributions to the automatic processing of the TD textual productions on the social web. Indeed, we proposed an approach that will help us automatically generate large scale TD LRs (corpora, lexica and dictionaries), which we propose to extract automatically from the informal textual productions generated on the social web and which may constitute a starting point for any NLP application involving the TD language. Through the proposed approach, we tackled two important problems: the TD identification, and the Latin to Arabic TD transliteration. Each of these NLP problems presents a challenge given the informality and the specificities of the language we are dealing with. It is also, important to note that the two main NLP components of the proposed approach are not only useful for the construction of TD LRs, but they may also be used as crucial steps in many NLP applications treating the TD language.

We presented the results obtained following our first experiments of the proposed framework implementation, regarding TD identification and TD transliteration. In our future work, we aim to focus on improving the obtained results. We intend to induce our CRF model with more features in order to reduce the identification error rate and to propose other statistical approaches based on n-grams. As for the transliteration, we aim to explore the use of deep learning approaches such as Bi-LSTM. Once the identification and the transliteration rates are improved and stabilised, the approach can be used to generate sizable and good quality LRs for the user-generated TD.

References

Ahmed, B., Cha, S.H. and Tappert, C. (2004) ‘Language identification from text using Ngram based cumulative frequency addition’, in *Proceedings of Student/Faculty Research Day*, CSIS, Pace University, New York, USA.

Al-Sabbagh, R. and Girjuh, R. (2012) ‘Yet another dialectal Arabic corpus’, in *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

Aridhi, C., Achour, H., Souissi, E. and Younes, J. (2017) ‘Word-level identification of Romanized Tunisian dialect’, in *Proceedings of the 22nd International Conference on Applications of Natural Language to Information Systems (NLDB)*, Liege, Belgium.

Ben Abdelkader, R. (1977) *Peace Corps English-Tunisian Arabic Dictionary*, p.575, ERIC Clearinghouse, Washington, D.C.

Ben Moussa, N.K. and Alimi, A.M. (2015) 'Construction d'un Wordnet standard pour l'Arabe tunisien', in *Proceedings of Colloque pour les Étudiants Chercheurs en Traitement Automatique du Langage naturel et ses applications*, Sousse, Tunisia.

Ben Moussa, N.K., Soussou, H. and Alimi, A.M. (2014) 'Building a standardized Wordnet in the ISO LMF for aeb language', in *Proceedings of the 7th Global Wordnet Conference (GWC 2014), Association for Computational Linguistics*, Tartu-Estonia, pp.71–77.

Ben Moussa, N.K., Soussou, H. and Alimi, A.M. (2015) 'Tunisian Arabic aebWordnet: current state and future extensions', in *Proceedings of the First International Conference on Arabic Computational Linguistics*, Cairo, Egypt.

Biébow, B. and Szulman, S. (1999) 'TERMINAE: a linguistics-based tool for the building of a domain ontology', in *Proceedings of the 11th European Workshop on Knowledge Engineering and Knowledge Management*, Dagstuhl Castle, Germany.

Bouchlaghem, R., Elkhlifi, A. and Faiz, R. (2014) 'Tunisian dialect Wordnet creation and enrichment using web resources and other Wordnets', in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, Doha, Qatar.

Boujelbane, R. (2013) 'Génération de corpus en dialecte tunisien pour l'adaptation de modèles de langage', in *Proceedings of Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, Les Sables d'Olonne, France.

Boujelbane, R., Khemekhem, M.E., Ben Ayed, S. and Belguith, L.H. (2013a) 'Building bilingual lexicon to create dialect Tunisian corpora and adapt language model', in *Proceedings of the 2nd Workshop on Hybrid Approaches to Translation*, ACL 2013, Sofia, Bulgaria.

Boujelbane, R., Khemekhem, M.E. and Belguith, L.H. (2013b) 'Mapping rules for building a Tunisian dialect lexicon and generating corpora', in *Proceedings of the International Joint Conference on Natural Language Processing*, Nagoya, Japan.

Boujelbane, R., Mallek, M., Khemekhem, M.E. and Belguith, L.H. (2014) 'Fine-grained POS tagging of spoken Tunisian dialect corpora', in *Proceedings of the 19th International Conference on Application of Natural Language to Information Systems*, Montpellier, France, pp.59–62.

Chalabi, A. and Gergers, H. (2012) 'Romanized Arabic transliteration', in *Proceedings of the 24th International on Computational Linguistics*, Mumbai, India.

Darwish, K. (2014) 'Arabizi detection and conversion to Arabic', in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, Doha, Qatar.

Diab, M., Habash, N., Rambow, O., Altantawy, M. and Benajiba, Y. (2000) 'COLABA: Arabic dialect annotation and processing', in *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valletta, Malta.

El-Haj, M., Kruschwitz, U. and Fox, C. (2014) 'Creating language resources for under resourced languages: methodologies, and experiments with Arabic', *Language Resources and Evaluation*, Vol. 49, No. 3, pp.549–580.

Graja, M., Jaoua, M. and Belguith, L.H. (2010) 'Lexical study of a spoken dialogue corpus in Tunisian dialect', in *Proceedings of the International Arab Conference on Information Technology (ACIT 2010)*, Benghazi, Libya.

Graja, M., Jaoua, M. and Belguith, L.H. (2011a) 'Building ontologies to understand spoken Tunisian dialect', *International Journal of Computer Science, Engineering and Applications (IJCSEA)*, Vol. 1, No. 4, pp.23–32.

Graja, M., Jaoua, M. and Belguith, L.H. (2011b) 'Towards understanding spoken Tunisian dialect', in *Proceedings of the 18th International Conference (ICONIP 2011)*, Shanghai, China.

Graja, M., Jaoua, M. and Belguith, L.H. (2013) 'Discriminative framework for spoken Tunisian dialect understanding', in *Proceedings of the First International Conference on Statistical Language and Speech Processing (SLSP 2013)*, Tarragona, Spain.

Graja, M., Jaoua, M. and Belguith, L.H. (2015) 'Statistical framework with knowledge base integration for robust speech understanding of the Tunisian dialect', *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 23, No. 12, pp.2311–2321.

Habash, N., Diab, M. and Rambow, O. (2012) 'Conventional orthography for dialectal Arabic', in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Hamdi, A., Gala, N. and Nasr, A. (2014) 'Automatically building a Tunisian lexicon for deverbal nouns', in *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, Dublin, Ireland, pp.95–102.

Karoui, J., Graja, M., Boudabous, M.M. and Belguith, L.H. (2013a) 'Domain ontology construction from a Tunisian spoken dialogue corpus', in *Proceedings of the International Conference on Web and Information Technologies (ICWIT 2013)*, Hammamet, Tunisia.

Karoui, J., Graja, M., Boudabous, M.M. and Belguith, L.H. (2013b) 'Semi-automatic domain ontology construction from spoken corpus in Tunisian dialect: railway request information', *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, Vol. 1, No. 1, pp.35–38.

Kobus, C., Yvon, F. and Damnati, G. (2008) 'Normalizing SMS: are two metaphors better than one?', in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling)*, Manchester, England.

Lafferty, J., McCallum, A. and Peireira, F.C. (2001) 'Conditional random fields: probabilistic models for segmentation and labeling sequence data', in *Proceedings of the 18th International Conference on Machine Learning*, San Francisco, USA, pp.282–289.

Maamouri, M. and Bies, A. (2004) 'Developing an Arabic Treebank: methods, guidelines, procedures, and tools', in *Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, Switzerland.

Masmoudi, A., Habash, N., Khemakhem, M.E., Estve, Y. and Belguith, L.H. (2015) 'Arabic transliteration of Romanized Tunisian dialect text: a preliminary investigation', in *Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics*, Cairo, Egypt.

Masmoudi, A., Khmekhem, M.E., Esteve, Y., Bougares, F. and Belguith, L.H. (2014a) 'Phonetic tool for the Tunisian Arabic', in *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, Petersburg, Russia.

Masmoudi, A., Khemakhem, M.E., Estève, Y., Bougares, F., Dabbar, S. and Belguith, L.H. (2014b) 'Phonétisation automatique du dialecte tunisien', in *Proceedings of JEP 2014*, Le Mans, France.

Masmoudi, A., Khemakhem, M.E., Estève, Y., Belguith, L.H. and Habash, N. (2014c) 'A corpus and phonetic dictionary for Tunisian Arabic speech recognition', in *Proceedings of the 9th Edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland.

McNeil, K. (2015) 'Tunisian Arabic corpus: a written corpus of an 'unwritten' language', in *International Symposium on Tunisian and Libyan Arabic Dialects*, University of Vienna.

McNeil, K. and Faiza, M. (2011) 'Tunisian Arabic corpus: creating a written corpus of an 'unwritten' language', in *Proceedings of the Workshop on Arabic Corpus Linguistics*, Lancaster University, UK.

Mdhaffar, S., Bougares, F., Estève, Y. and Belguith, L.H. (2017) 'Sentiment analysis of Tunisian dialect: linguistic resources and experiments', in *Proceedings of the 3rd Arabic Natural Language Processing Workshop*, pp.55–61, Valencia, Spain.

Meftouh, K., Bouchemal, N. and Smaili, K. (2012) 'A study of a non-resourced language: an Algerian dialect', in *Proceedings of the 3rd International Workshop on Spoken Languages Technologies for Under-Resourced Languages*, South Africa.

Mekki, A., Zribi, I., Khemekhem, M.E. and Belguith, L.H. (2017) 'Syntactic analysis of the Tunisian Arabic', in *Proceedings of the International Workshop on Language Processing and Knowledge Management*, Sfax, Tunisia.

Sha, F. and Pereira, F. (2003) 'Shallow parsing with conditional random fields', in *Proceedings of HLT-NAACL*, Edmonton, Canada.

Sutton, C. and McCallum, A. (2012) 'An introduction to conditional random fields', *Foundations and Trends in Machine Learning*, Vol. 4, No. 4, pp.267–373.

Tromp, E. and Pechenizkiy, M. (2011) 'Graph-based n-gram language identification on short texts', in *Proceedings of the 20th Machine Learning conference of Belgium and The Netherlands*, The Hague, Netherlands.

Younes, J. and Souissi, E. (2014) 'A quantitative view of Tunisian dialect electronic writing', in *Proceedings of the 5th International Conference on Arabic Language Processing (CITALA)*, Oujda, Morocco.

Younes, J., Achour, H. and Souissi, E. (2015) 'Constructing linguistic resources for the Tunisian dialect using textual user-generated contents on the social web', in Daniel, F. and Diaz, O. (Eds.): *Current Trends in Web Engineering: 15th International Conference, ICWE 2015 Workshops (NLPIT)*, Rotterdam, Netherlands.

Younes, J., Souissi, E. and Achour, H. (2016) 'A hidden Markov model for the automatic transliteration of Romanized Tunisian dialect', in *2nd International Conference on Arabic Computational Linguistics Co-located with CICLing 2016, 17th International Conference on Intelligent Text Processing and Arabic Computational Linguistics*, Konya, Turkey.

Zaidan, O.F. and Callison-Burch, C. (2011) 'The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content', in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, USA.

Zaidan, O.F. and Callison-Burch, C. (2014) 'Arabic dialect identification', *Computational Linguistics*, Vol. 40, No. 1, pp.171–202.

Zribi, I., Boujelbane, R., Masmoudi, A., Khemekhem, M.E., Belguith, L.H. and Habash, N. (2014) 'A conventional orthography for Tunisian Arabic', in *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Zribi, I., Khemakhem, M.E. and Belguith, L.H. (2013) 'Morphological analysis of Tunisian dialect', in *Proceedings of the International Joint Conference on Natural Language Processing*, Nogoya, Japan.

Zribi, I., Khemakhem, M.E., Belguith, L.H. and Blache, P. (2015) 'Spoken Tunisian Arabic corpus 'STAC': transcription and annotation', *Research in Computing Science*, Vol. 90, pp.123–135.

Notes

1 http://www.elra.info/en/about/what-language-resource/.