Tunisian dialect Wordnet creation and enrichment using web resources and other Wordnets

Rihab Bouchlaghem
LARODEC, ISG de Tunis
2000 Le Bardo, Tunisie
rihab.bouchlaghem@isg.rnu.tn

Aymen Elkhlifi
Paris-Sorbonne University,
28 Rue Serpente, Paris, France
Aymen.Elkhlifi@paris.sorbonne.fr

Rim Faiz
LARODEC, IHEC de Carthage,
2016 Carthage Présidence, Tunisie
Rim.Faiz@ihec.rnu.tn

Abstract

In this paper, we propose TunDiaWN (Tunisian dialect Wordnet) a lexical resource for the dialect language spoken in Tunisia. Our TunDiaWN construction approach is founded, in one hand, on a corpus based method to analyze and extract Tunisian dialect words. A clustering technique is adapted and applied to mine the possible relations existing between the Tunisian dialect extracted words and to group them into meaningful groups. All these suggestions are then evaluated and validated by the experts to perform the resource enrichment task. We reuse other Wordnet versions, mainly for English and Arabic language to propose a new database structure enriched by innovative features and entities.

1 Introduction

The Arabic Dialects have become increasingly used in social networks and web 2.0 (blogs, forums, newspaper, newsgroups, etc.) instead of Standard Arabic (SA).

Consequently, new kinds of texts appeared being mainly dialect-written or having a mixture between Arabic Dialects and Standard Arabic. Thus, innovative opportunities and challenges arise when we try to deal with the automatic processing of such data in order to seek out useful information and take advantages of their growing availability and popularity. The NLP approaches generally applied lexical resources for the target language. Such resources are useful in several tasks which involve a language meaning understanding like: opinion mining (Kim et al., 2004; Bouchlaghem et al. 2010), information retrieval (Valeras et al., 2005; Rosso et al., 2004), query expansion (Parapar et al., 2005), text categorization (Rosso et al., 2004; Ramakrishnan et al., 2003), and many other applications.

However, this situation poses significant difficulties in the context of dialectal data because of the huge lack of Dialect-Standard Arabic lexical resources. Building similar ones is a big challenge since spoken dialects are not officially written, don't have a standard orthography and are considered as under-resourced languages, unlike standard languages.

In this paper, we address the problem of creating a linguistic resource for an Arabic dialect. We describe our approach towards building a Wordnet for Tunisian dialect (TD). We proceed, firstly, to construct a TD corpus by collecting data from various resources (social networks, websites, TD dictionaries, etc.). We develop a clustering based method that aims to organize the TD corpus words by grouping them into clusters. The suggested organization possibilities are, then, analyzed and validated by the TD experts during the TunDiaWN enrichment process. Our proposed database structure is designed to be able to highlight the specificities of the TD lexicon. It also takes advantage of Arabic Wordnet (AWN) (Elkateb et al., 2006), the Arabic version of the widely used lexico-semantic resource Princeton WordNet (PWN) (Fellbaum, 1998). This can be justified by the assumption that Tunisian Arabic has a great resemblance with Standard Arabic.

The rest of the paper is organized as follows: we begin by presenting works related to existing wordnets and approaches focused on the auto-

Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), pages 104–113,
October 25, 2014, Doha, Qatar.

matic processing of the Tunisian dialect. We then introduce the posed challenges and the hypothesis we have assumed in building the TunDiaWN. In the next section, we proceed to explain and justify the proposed approach for developing the initial version of the Tunisian Arabic lexical resource. Firstly, we detail the TD data collect process and the MultiTD corpus construction. Secondly, we present the method developed to suggest possible organizations of TD words extracted from the corpus. Then, we describe the proposed structure of TunDiaWN, especially the new added features and entities as well as the validation task performed by the TD experts. In the following section, we perform a linguistic analysis by reporting significant observations related to TD-SA discovered during the enrichment process. Conclusion and future works are presented in section 5.

2 Related works

The first version of wordnet (Fellbaum, 1998) was developed for English at Princeton University. It's a large lexical database where words having the same part of speech (Nouns, verbs, adjectives, adverbs) are gathered in sets of cognitive synonyms (synsets), each one expressing a distinct concept. Each word can belong to one or more synsets. The resulting synsets are connected by means of conceptual-semantic and lexical relations well labeled such as hyponymy and antonymy.

The success of the Princeton WordNet has motivated the development of similar resources for other languages, such as EuroWordNet, EWN (Vossen, 1998) interlinking wordnets of several European languages (Dutch, Italian, Spanish, German, French, Czech and Estonian); Balkanet (Tufis, 2004) comprising wordnets of the Balkan languages; and recently Arabic Wordnet (AWN) (Elkateb et al., 2006).

AWN was released following methods developed for EuroWordNet. These methods revolve about the manual encoding of a set of Common Base Concepts (CBC), the most important concepts from the 12 languages in EWN and BalkaNet. Other language-specific concepts are added and translated manually to the closest synset(s) in Arabic. Such resource aims to link arabic words and synsets to english ones.

AWN is related to the Classical Arabic (or Literal Arabic) which refers to the official standard form of the Arabic language used in Arab world. Other variants of Arabic are dialects which are spoken and informal. They are the primary form of Arabic Language.

The Tunisian dialect (cf. Table 1) or 'Darija' is one of the Maghreb Dialects and is mainly spoken by Tunisian people (Baccouche, 1994).

Tunisian dialect words	Transliteration	Meaning
فْلُوسْ	foluws	money
بَرْشَا	baro$aA	many
مَالَهْ	maAlah	then

Table 1. Examples of popular TD words

Most of the works that dealt with the automatic processing of Tunisian dialect are based on spoken dialogue corpus. To mention, Graja et al. (2011) performed a lexical study of manual transcription of conversations recorded in the railway station for understanding speech. The application is domain dependant and, thus, the vocabulary is limited. Moreover, Zribi et al. (2013) introduced a lexicon for the Tunisian dialect in order to adapt an existing morphological analyzer initially designed for Standard Arabic. Although the method shows good results, the proposed lexicon is far to be complete. Boujelbane et al. (2013) presented a method that aims to construct bilingual dictionary using explicit knowledge about the relation between Tunisian dialect and Standard Arabic. This approach was limited to the verbs.

3 Challenges

In the last years, Tunisian dialect is widely used in new written media and web 2.0, especially in social networks, blogs, forums, weblogs, etc., in addition to conversational media (Diab et al., 2007).

Thinking about building a wordnet for Tunisian dialect is a big challenge. In fact, like most of dialects around the world, Tunisian Arabic is considered as spoken language with no conventional written form. Moreover, there is a lack of Tunisian dialect-Standard Arabic resources and tools.

Recently, Cavalli-Sforza et al. (2013) proposed a process for creating a basic Iraqi Dialect WordNet. This work is based on other languages wordnets as well as a bidirectional English-Iraqi Arabic dictionary. To our knowledge, no other open source Wordnet for the Standard Arabic or Arabic Dialect has been developed to date.

To deal with these difficulties, we decide to produce a TD corpus gathering texts from multiple

sources. This corpus provides a useful starting point for building a wordnet for Tunisian dialect.
We assume that Arabic Dialects can be presumed to be similar to Standard Arabic, particularly in their conceptual organization. Indeed, the Tunisian dialect has a sophisticated form which combines Standard Arabic and Tunisian dialect specific forms. It has a great resemblance to the SA and adds some variances such as foreign words borrowed from other languages. Thus, given the similarities between the TD and the SA, the resources available to SA, such as AWN, can be favorably used for creating Tunisian dialectal resources.

4 Proposed approach for TunDiaWN construction

The classical building WordNets methodologies start from the CBC, and then make changes according to the concerned language.
We propose a new corpus-based approach to create WordNet resource for Tunisian dialect, which deviates from the strategies commonly adopted.
As shows Figure 1, our approach is performed in four steps:

a. *Tunisian dialect textual data collect:* it consists in producing our **MultiTD** corpus (***Multi-source Tunisian dialect*** *corpus*) which gathers TD texts from many sources: social networks (Twitter, Facebook, etc.), written pieces of theater, dictionaries, transcriptions of spontaneous speech, etc.

b. *TD words extraction:* is to preprocess the produced corpus in order to preserve useful data and extract TD words.

c. *TD words clustering*: we propose here a clustering based method that aims to group the extracted TD words into meaningful clusters, which represent great suggestions for possible enrichments of TunDiaWN.

d. *TunDiaWN enrichment*: this step is performed by the TD experts. It includes the manual validation of the suggestions proposed by the previous step. We propose, in this stage, a new database structure for TunDiaWN. The experts have to add the necessary features values, particularly the TD specific attributes (details in section 4.4).

4.1 TD data collection and *MultiTD* corpus presentation

We set out to collect data for Tunisian dialect in order to address the general lack of resources, on the one hand, and to produce a multi source corpus, on the other.
We created the **MultiTD** corpus by gathering TD data from diverse sources.
The most practical source of TD texts is online data, which is more individual-driven and less formal, and consequently more likely to comprise dialectal contents.

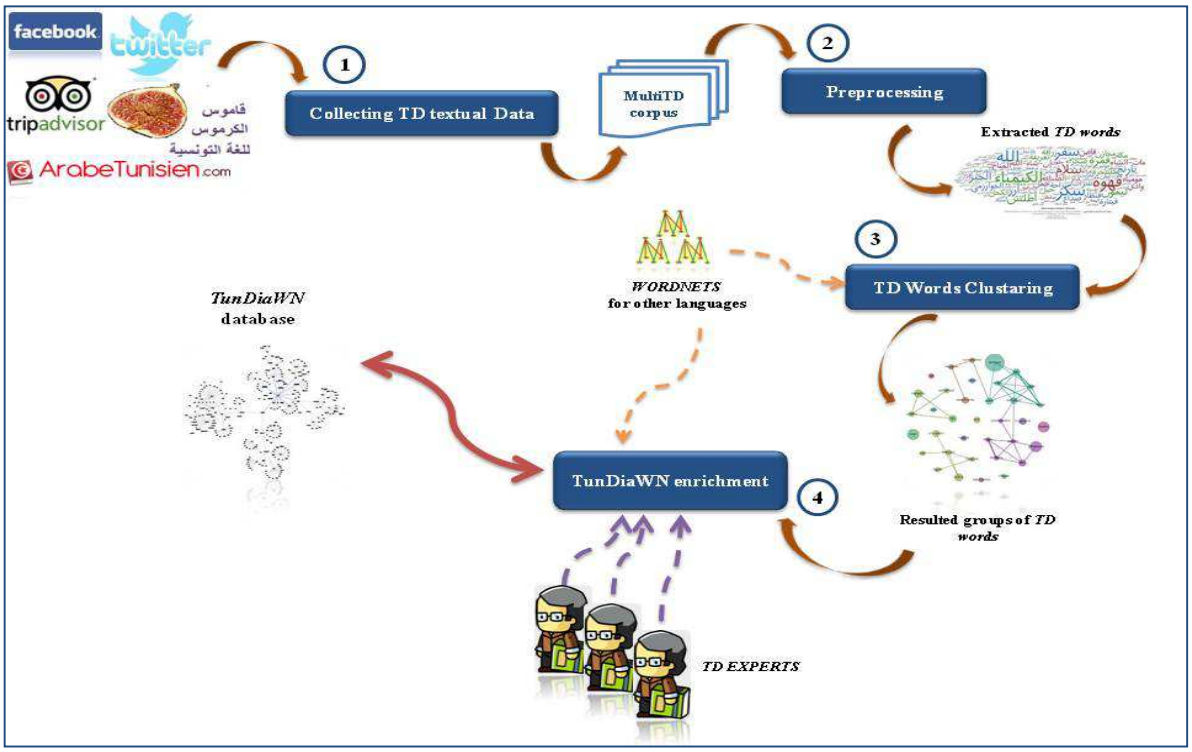

Figure 1. Proposed approach of *TunDiaWN* building

We automatically collected a great amount of TD texts from user's comments and status from *Twitter*, *Facebook* and *TripAdviser*.

We have implemented three specific modules:

- TwtterCollecter based on *Twitter4j java* api,
- FacebookAspirator using a *PHP* script and a Facebook account developer,
- TripadvisorScreen a java module to analyze Tripadvisor web pages and extract comments forms.

Manual transcriptions of TD recorded spontaneous speech are also added to the *MultiTD* corpus. Such data allows highlighting the Tunisian accent in the dialogue and, therefore, enriching the corpus by new varieties of the TD lexicon.
Other online available TD resources are used to enrich the *MultiTD* corpus. We cite notably, the *Karmous* dictionary for Tunisian Arabic[1] which comprises more than 3,800 TD words and several Tunisian proverbs and expressions organized by themes.
We use also an online TD dictionary [2] consisting of over 4,000 words and expressions; and many short TD texts[3] related to various areas: songs, theater, newspaper articles, etc.

4.2 TD words extraction

To successfully extract all TD words, the input texts must be preprocessed. In our study, the preprocessing consists, firstly, to clean the input files so as to identify the textual content. The cleaned texts are then segmented in order to extract all existing TD words.
Cleaning a raw textual source is necessary in our approach because the documents are collected from the Web. All non-textual data such as images, advertisements, scripts, etc. have to be eliminated. For this purpose, we have developed a module that removes all unwanted parts from the input documents.

The cleaned texts are then segmented into elementary textual units and the obtained TD words are extracted and stored in CSV files.

The Table 2 gives statistics about the TD words composing the *MultiTD* corpus.

		TD words count
Social netwoks	**Twitter**	10249
	Facebook	7470
	Tripadvisor	3258
TD transcripts texts		2351
Other sources (pieces of theatre, dictionaries, etc.)		9520
TOTAL		32848

Table 2. Distribution of TD words in *MultiTD* corpus, according to sources

4.3 TD words clustering using k-modes algorithm

The TunDiaWN construction is based on a semi-automatic process in which the validation tasks performed by experts are crucial.
As Table 2 Shows, the MultiTD corpus includes a huge number of TD words. The manual analysis and organization of such large data looks wasteful and time consuming.
In order to support experts in the organization and validation tasks and guide them during the construction process, we propose a clustering-based method to automatically arrange the TD words set into groups. The method aims to suggest possible organizations of the given TD words by gathering them into meaningful clusters.
To enhance similarities and meanings into the produced groups, we propose to cluster the TD words according to their TD roots. We rely here on the derivational morphology that characterizes the Tunisian dialect as well as the Standard Arabic.
In fact, many SA words having a common root [4] can be derived from a base verbal form and have related meanings. An example of such a field for the root **درس**, 'to study,' is shown in Table 3.

Arabic words	**Part of speach**	**Meaning**
دَرَسَ	verb	study
دَرَّسَ	verb	teach
تَدْرِيس	noun	teaching

Table 3. Some derivatives of Arabic root "درس" (Elkateb et al., 2006)

In the same context, the TD morphology is derivational too (cf. Table 4).

Taking advantage of this central characteristic, the set of TD words can be organized into distinct semantic groups according to the TD roots from which they are derived. The list of TD roots

[1] The dictionary can be obtained from : *http://www.fichier pdf.fr/2010/08/31/m14401m/dico-karmous.pdf*
[2] Link : *http://www.arabetunisien.com/*
[3] Download link: *http://www.langue-arabe.fr/spip.php? article25*
[4] جِذْر *in Arabic:* a sequence of typically three consonants.

we have used was obtained by translating the SA roots provided by AWN.

TD words	Transliteration	Part of speach	Meaning
قْرَى	qoraY	verb	study
قَرَّى	qar~aY	verb	teach
قْرَايَة	qoraAyap	noun	teaching

Table 4. Some derivatives of TD root "قرى"

We don't search here to automatically enrich the TunDiaWN structure by attaching new TD words, but we rather suggest new attachments and enrichment possibilities which can help the experts.

Our aim at this step is to group words having the same root. To do this task, we apply and adapt the K-modes clustering algorithm (Huang, 1997). The K-modes algorithm extends K-means (Forgy, 1965; MacQueen, 1967) paradigm to cluster categorical data by removing the numeric data limitation. Indeed, the K-modes algorithm introduces a new simple matching dissimilarity measure for categorical data objects. The algorithm replaces means of clusters with modes, and uses a frequency based method to update modes in the clustering process.

The choice of K-modes clustering algorithm is mainly motivated because of its widely use in real world applications due to its efficiency in dealing with large categorical database (He et al., 2011). K-modes algorithm is also faster than other clustering algorithms (mainly k-means) since it needs less iteration to produce a stable distribution. .

The K-modes algorithm requires a similarity measurement to be used between the objects. In our case, we propose to use the N-Gram similarity measurement between words. N-Gram is language independent in nature and doesn't require specific resources to be applied. Therefore, N-gram model seems suitable for dealing with a Tunisian dialect context. We applied the N-Gram distance proposed by Kondrak (2005) and we used the implementation provided by Apache Lucene spellchecking API[5].

The K-modes algorithm consists of the following steps:

a) Select K initial modes, one for each of the cluster.

b) Allocate data object to the cluster whose mode is nearest to it, according to the simple matching dissimilarity

c) Compute new modes of all clusters.

d) Repeat step b to c until no data object has changed cluster membership.

The classical K-modes algorithm assumes that the number of clusters, K, is known in advance and the clusters' modes are randomly initialized. The K-modes algorithm is very sensitive to these choices and an improper choice may then yield highly undesirable cluster structures. (Khan et al., 2013).

In order to deal with these drawbacks and, thereafter, maximize the performance of the algorithm, we propose a new initialization strategy for the k-modes algorithm.

Indeed, since our goal is to cluster words according to their roots, the TD roots are assigned to clusters modes in the initialization step instead of random initialization. The number of clusters (K) will, thus, take the cardinality of the target TD roots set. Therefore, the K-modes algorithm starts with k clusters each having as mode one root among the TD roots list initially translated.

We have also adopted a new strategy based on the N-Gram similarity measurement to update clusters' modes. The modes update is performed at the end of each iteration. For each cluster, the item qualified as new cluster mode must maximize the similarity sum with the rest of cluster objects.

The K-modes algorithm adapted for our purpose performs as following:

a. Initialization
   K = |set of *TD roots*|
   Initial modes = TD roots, one for each of the cluster.

b. Allocate each word (itm_i) of TD words set to the cluster $Cluster_s$ whose mode $ModeCL_s$ is nearest to it according to the equation (1) .

$$ModeCL_s = \underset{j}{argmin}^{k} (1 - simNGram(itm_i, ModeCL_s)) \quad \textbf{(1)}$$

c. Update modes of all clusters :

$$\forall Cluster_s, s = 1 \rightarrow K$$

c.1. Similarity computing

$$\forall itm_i \in Cluster_s, i = 1 \rightarrow |Cluster_s|$$

[5] The project can be freely obtained from: http://lucene.apache.org/core/

$$ModeSim(itm_i, Cluster_s) = \sum_{j=1}^{|Cluster\ s|} simNGram(itm_i, itm_j) \quad (2)$$

c.2. Modes selection

$$\forall ModeCL_s, s = 1 \rightarrow K$$

$$ModeCL_s = \underset{i}{argmax}^{n} (ModeSim(itm_i, Cluster_s)) \quad (3)$$

d. Repeat step (b) to (c) until no TD words has changed cluster membership.

After performing the new proposed version of the k mode algorithm, the obtained results are suggested to be validated by the TD experts in order to enrich TunDiaWN structure, which will be presented in the next section.

4.4 TD groups' validation and TunDiaWN enrichment

In this section, we begin by describing the proposed structure of TunDiaWN. After that, we detail the enrichment task performed by the TD experts. Then, we present a linguistic study performed during the enrichment process.

TunDiaWN structure

As our target language is an Arabic Dialect and therefore likely to share many of the Standard Arabic concepts, we decide to preserve the AWN design. However, the AWN current structure is unable to support the specificities of the Tunisian dialect lexicon. The proposed TunDiaWN structure is then enriched by new features, entities and relations. Moreover, we aim to create a parallel resource which maintains the linkage between Tunisian dialectal, Arabic as well as English synsets and words. That's why AWN and PWN contents are preserved rather than the structures. Thus, the proposed database is designed to be able to support English, Tunisian and Standard Arabic content and correspondence.
In this section, we detail the structure of the proposed TunDiaWN database and we focus on the new features we added to keep up the TD vocabulary particularities, compared to the SA and English ones.

TWN entity types

The database structure incorporates mainly the following entity types: *synset, word, form, synset relations, words relations, annotator:*

***Synset*:** includes English and Arabic synsets. A synset has descriptive information such as Name, POS (Part Of Speech), root (Boolean feature indicating if the target synset is a root or not).

Word: comprises words from different languages. In addition to the unique identifier, every word is described by his value, and a Boolean "valid" attribute which indicates if one word is already validated by experts or not yet.

Form: includes mainly the root of Arabic as well as Tunisian dialect words.

Synsets relations: includes links relating two synsets, like "*has_instance*", "*equivalent*", "*similar*", etc. We preserve here all sunsets' links without adding new ones.

Words relations: two English words can be linked by "*pertainym*" or "*antonym*" relations. There are no added Arabic words relations.

Annotator: is used to indicate who has validated each word. The attribute "*region*" helps to classify words by region and identify where words come from. We assume here that the annotator will do his job according to the background of his native region.

TunDiaWN new features

Since the Tunisian dialect is not a standard language, new features are required to be added to the TunDiaWN resource in order to preserve the TD specificities. We describe below the most important TD characteristics integrated in the proposed resource:

SMS language

In the context of Tunisian dialect, the SMS language is a written form which combines Latin script and some numbers in order to express dialectal words.
The SMS language is widely used especially in social networks and blogs.
Table 5 gives examples of the most used numbers which aim to replace specific Arabic letters.
TD words are illustrated with Latin Script (Latin), Arabic Letters (Ar-L) and using transliteration[6].

[6] Throughout this paper we use the Buckwalter transliteration : http://www.qamus.org/transliteration.htm

Numbers	Arabic replaced letters	Dialectal words			Part of speech	Arabic translation	English translation
		Latin Scrip	Arabic letters	Transliteration			
3	العَيْن ع	3ayyet	عَيِّطْ	Eay~iT	verb	صَاحَ	To cry
5	الخَاء خ	5allé	خَلّى	xal~aY	verb	تَرَكَ	To leave
7	الحاء ح	7outa	حُوتَة	Huwtap	noun	سَمَكَة	A fish
9	القَاف ق	9ale9	قَالِقْ	qaAliq	adjec-tive	ضَجِرٌ	bored

Table 5. TD words written using the SMS language

Foreign words

The use of foreign words is a prominent feature in the Tunisian community due to historical reasons. Foreign words are used in almost everyday conversation.

The following table (table 6) illustrates the use of foreign words next to Tunisian dialect ones in the same sentence.

Tunisian dialect (Latin)	**En tout cas**, n7eb n9ollek **merci** 3alli 3maltou m3aya. Net9ablou mba3ed, **à toute**.
Tunisian dialect (Ar-L)	**أَنْتوُكاَ**، نْحِبّ نْقُلّكْ **مِيرْسِي** عَلّي عْمَلْتُو مْعَايَا. نِتْقَابْلُو مْبَاعِدْ، **اَتُوتْ**
French Translation	**En tout cas**, je veux te dire **merci** pour tout ce que t'as fais pour moi. on se voit après, **à toute**.
English Translation	Anyway, I want to say thank you for everything you've done for me. See you later.

Table 6. Examples of French words widely used in TD communications

A TD corpus study found that pure French origin words are ubiquitous and represent 11.81% of the dialogue corpus (Graja et al, 2010).

Tunisian dialect can also borrow and adapt words from other languages in order to make them sound and behave like TD words.

As an illustration, the TD word " **تْنَرْفِيزْ**/ tonarofiyzo" is derived from the French word "nervosité" and is synonym to the English word "anger".

As can be seen, the foreign words are part of the Tunisian dialect vocabulary. Such words must not be neglected. They must be added to any dictionary of Tunisian dialect lexicon (Graja et al. 2010).

The foreign words used with their original forms are added to the TunDiaWN database.

Concerning the TD words having foreign origins, they are firstly distinguished from other TD words. The second step consists in finding the origin words in other languages, saving them and linking them to the concerned TD words. Consequently, the borrowed TD words are easily identified. Their basic language and words are straightforwardly found and browsed.

Morphology

Since the Tunisian dialect has no standard orthography, one word can be written in many forms using Arabic letters or Latin script. For example, the word "will" can be expressed in different ways: "bech"/ "**بَاشْ**", "bich" /"**بِشْ**", "mich"/"**مِشْ**".

To deal with this situation, our database structure is enriched by a new entity named "morphology" which allows storing all versions of a given TD word.

Sub-dialect group

There are many varieties of Tunisian dialect taking into account the lexical variation depending on Tunisian regions. We can distinguish mainly three sub-dialects in the dialect of each region: *the townspeople*, *peasants/farmers*, *Be-Douin*. This is mainly due to the difference in cultures which adds several different words from different backgrounds having the same meaning. (Graja et al, 2010). The feature "sub-dialect" as well as the "Region" of the annotator are used to give further information about the origin of the target word.

The TD words: "$aAf/**شَافْ**", "roEaY/**رْعَى**", "$obaH/**شْبَحْ**", "gozar/**غْزَرْ** ", are used in different Tunisian regions and are synonyms< to the English word "to look".

TunDiaWN enrichment task

One of our strategic goals is to provide a parallel resource which deal with the lack of parallel TD-SA dictionaries and corpus. Therefore, we proceed by gathering Tunisian dialect and Standard Arabic in one unique structure and maintain the link with the Standard English too.

The starting point of the TunDiaWN enrichment step is the groups of TD words, resulted of per-

forming our clustering based method. The TD roots presumed to be the center of groups are obtained by translating the SA roots available in AWN.
For each TD root, the SA words related to the equivalent SA root are extracted. Two lists of words derived from equivalent roots are available: one is related to a SA root, and the other is from a TD one. The concerned SA synsets are also available.
After that, the TD experts analyze and confront the lists in order to find new synsets enrichment opportunities. The TD words qualified to be retained are those maximizing the synset harmony.
The TD experts must also fill in the necessary attributes related to the added words and manually make the necessary changes and enrichments.
In fact, the added words have to be described according to the new features added to the TunDiaWN database, so as to bring different knowledge of different vocabularies and give all useful details related to the target word.

Linguistic study of the enriched TunDiaWN

The linguistic study of the enrichment possibilities validated by the TD experts shows many important lexical trends in the TD lexicon comparing to the SA vocabulary.
A great part of Arabic synsets is enriched by words that conserve the same SA roots and derivation patterns but appear with small changes in vowels (cf. table 7).

Arabic	Tunisian dialect			Translation
	Ar-L	SMS langage	Transliteration	
قَرَّرَ	قَرِّرْ	9arrer	qar~ir	to decide
زَلَقَ	زْلَقْ	zlo9	zoluq	to slide

Table 7. Example of TD words having SA roots and derivation patterns

We distinguish also words derived from SA roots via the application of specific derivation patterns of TD (cf. table 8). Those words are omnipresent in the TD lexicon.

Moreover, some TD words has identical morphologies comparing to other SA words, but the meaning is far to be similar (cf. table 9).

SA	SA→English translation	TD	TD→English translation
تَعَرَّضَ	To be exposed to	تْعَرِّضْ	to disagree

Table 9. Examples of TD words having similar SA morphologies and different meanings

There is another category of TD words which are very similar to SA words, but use a different preposition.

For example, the SA word "**تَسَبَّبَ بِ**/ttasab~aba bi", which means "to cause", has an equivalent TD word "**تْسَبِّبْ فِي**/tsab~ib fiy" with just different vowels and new preposition.

In some cases, the SA words are linked to TD expressions which have the same meaning, since there are no TD simple equivalent words, as illustrates the following table:

Arabic	Tunisian dialect		Translation
أَزَّمَ, صَعَّبَ	طَلَعْ الما لِلصَّعْدَة	Tal~aEo AlmA lilS~aEodap	To aggravate

Table 10. TD expressions equivalent to SA words

We deduce from this study and the given examples that the Tunisian dialect is marked by a lexical variety which escapes from the standard rules of the Standard Arabic.

5 Conclusion and future works

We have described an approach for building a Tunisian dialect lexical resource which takes advantages of online TD resources and reuses Wordnets of other languages.
The proposed TunDiaWN can be considered as parallel TD-SA resource since it preserves the AWN content. Thanks to the novel added TD attributes, the TunDiaWN design provides, also, great opportunities to deal with the lack of a standard written form and other specificities of the Tunisian dialect.
The construction process begins with the MultiTD corpus construction from many sources. After preprocessing the collected texts, the TD extracted words are gathered according to their common TD roots.

Arabic		أَفْقَرَ	أَضْعَفَ	اِنْتَعَشَ	اِنْتَفَخَ
Tunisian dialect	Arabic Letters	فَقَرْ	ضَعَّفْ	تْنَعْوِشْ	تِنْفَخْ
	Transliteration	faq~ir	DaE~if	tonaEowi$	tinofax
Root		فقر	ضعف	نعش	نفخ
Translation		To beggar	To impoverish	To refresh	To swell

Table 8. Examples of TD words having SA roots and applying specific TD patterns

Our aim at this level is to support the TD experts in the database enrichment task, by giving suggestions of the possible TD words organizations. Now, the proposed TD resource is under construction and evaluation. We plan to improve the coverage of TunDiaWN and looking for other TD specificities not yet covered. We plan also to incorporate the French language into the TunDiaWN content, taking advantages of the available lexical French resource WOLF (Sagot and Fišer, 2008).

Reference

Benoît Sagot and Darja Fišer. 2008. Construction d'un wordnet libre du français à partir de ressources multilingues. In proceeding of TALN conference, Avignon, France.

Christiane Fellbaum. 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Dan Tufis, Dan Cristea and Sofia Stamou. 2004. BalkaNet: Aims, Methods, Results and Perspectives. *A General Overview. Romanian Journal on Information Science and Technology*, Dan Tufiş (ed.), Special Issue on BalkaNet, Romanian Academy, 7 (1–2), 7–41.

David Parapar, Álvaro Barreiro and David E. Losada. 2005. Query expansion using wordnet with a logical model of information retrieval. IADIS AC: 487-494.

E. W. Forgy. 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics - A Journal of the International Biometric Society*, 21: 768–769.

Ganesh Ramakrishnan, Kedar Bellare, Chirag Shah and Deepa Paranjpe. 2003. Generic Text Summarization Using Wordnet for Novelty and Hard. TREC: 303-304.

Giannis Varelas, Epimenidis Voutsakis, Euripides G. M. Petrakis, Evangelos E. Milios, Paraskevi Raftopoulou. 2005. Semantic similarity methods in wordNet and their application to information retrieval on the web. In proceedings of, the 7th annual ACM international workshop on Web information and data management WIDM'07, Bremen, Germany: 10-16.

Grzegorz Kondrak. 2005. N-gram similarity and distance". Proceedings of the Twelfth International Conference on String Processing and Information Retrieval, SPIRE 2005, Buenos Aires, Argentina: 115-126.

Horacio Rodríguez, David Farwell, Javi Farreres, Manuel Bertran, M. Antonia Martí, William Black , Sabri Elkateb, James Kirk, Piek Vossen, Christiane Fellbaum. 2008. Arabic WordNet: current state and future extensions. In Proceedings of The Fourth Global WordNet Conference, Szeged, Hungary.

J. MacQueen. 1967. Some Methods for classification and Analysis of Multivariate Observations. In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press: 281–297.

Ines Zribi, Mariem Ellouze Khemekhem and Lamia Hadrich Belguith. 2013. Morphological Analysis of Tunisian dialect. In proceeding of the International Joint Conference on Natural Language Processing, Nagoya, Japan: 992–996.

Marwa Graja, Maher Jaoua and Lamia Hadrich Belguith. 2010. Lexical Study of A Spoken Dialogue Corpus in Tunisian dialect. In proceeding of the International Arab Conference on Information Technology ACIT'2010, Benghazi-Libya.

Mona Diab and Nizar Habash. 2007. Arabic Dialect Processing Tutorial. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Tutorial Abstracts: 5-6.

Paolo Rosso, Edgardo Ferretti, Daniel Jiménez and Vicente Vidal. 2004. Text Categorization and Information Retrieval using WordNet senses. In Proceeding of the 2nd Global WordNet International conference, Brno, Czech Republic: 299-304.

Piek Vossen. 1998. Introduction to EuroWorNet. *Computers and the Humanities*, 32(2-3), 73-89.

Rihab Bouchlaghem, Aymen Elkhlifi and Rim Faiz. 2010. Automatic extraction and classification approach of opinions in texts. In Proceeding of the 10th International Conference on Intelligent Systems Design and Applications, ISDA 2010, Cairo, Egypt. IEEE 2010: 918-922.

Rahma Boujelbane, Mariem Ellouze Khemekhem and Lamia Hadrich Belguith. 2013. Mapping Rules for Building a Tunisian dialect Lexicon and Generating Corpora. In Proceedings of the International Joint Conference on Natural Language Processing. Nagoya, Japan: 419–428.

Sabri Elkateb , William Black , Horacio Rodríguez , Musa Alkhalifa , Piek Vossen , Adam Pease and Christiane Fellbaum. 2006. Building a WordNet for Arabic. In Proceedings of The fifth international conference on Language Resources and Evaluation; Genoa-Italy: 29-34.

Shehroz S. Khan and Amir Ahmad. 2013. Cluster center initialization algorithm for K-modes clustering. *International journal of Expert Systems with Applications*, 40(18): 7444-7456.

Soo-Min Kim and Eduard Hovy.(2004). Determining the sentiment of opinions. In Proceedings of the

20th international conference on Computational Linguistics COLING '04: 1267–1373.

Violetta Cavalli-Sforza, Hind Saddiki, Karim Bouzoubaa, Lahsen Abouenour, Mohamed Maamouri and Emily Goshey. 2013. Bootstrapping a WordNet for an Arabic dialect from other WordNets and dictionary resources. In Proceedings of the 10th IEEE International Conference on Computer Systems and Applications, Fes/Ifrane, Morocco.

William Black, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, Christiane Fellbaum. 2006. Introducing the Arabic WordNet project. In Proceedings of the Third International WordNet Conference, Fellbaum and Vossen (eds).

Zengyou He, Xaiofei Xu and Shengchun Deng. 2011. Attribute value weighting in k-modes clustering. *International journal of Expert Systems with Applications*, 38(12): 15365-15369.

.Zhexue Huang. 1997. A fast clustering algorithm to cluster very large categorical data sets in data mining. In Proceeding of SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery: 1-8