# Spoken Tunisian Arabic Corpus "STAC": Transcription and Annotation

Inès Zribi[1], Mariem Ellouze[1], Lamia Hadrich Belguith[1], and Philippe Blache[2]

[1] ANLP Research Group, MIRACL Lab., University of Sfax, Tunisia,
ineszribi@gmail.com, mariem.ellouze@planet.tn, l.belguith@fsegs.rnu.tn
[2] Aix-Marseille Université & CNRS LPL, 13100, Aix-en-Provence, France.
philippe.blache@lpl-aix.fr

**Abstract.** Corpora are considered as an important resource for natural language processing (NLP). Currently, the Dialectal Arabic corpora are somewhat limited, particularly in the case of the Tunisian Arabic. In recent years, since the events of the revolution, the increasing presence of spoken Tunisian Arabic in interviews, news and debate programs, the increasing use of language technologies for many spoken languages (e.g., Siri) [6], and the need for works on speech technologies requires a huge amount of well-designed Tunisian spoken corpora.
This paper presents the "STAC" corpus (Spoken Tunisian Arabic Corpus) of spontaneous Tunisian Arabic speech. We present our method used for the collection and the transcription of this corpus. Then, we detail the different stages done to enrich the corpus with necessary linguistic and speech annotations that makes it more useful for many NLP applications.

**Keywords:** Tunisian Arabic, spoken language, corpus transcription, annotation

## 1 Introduction

The colloquial Arabic or Dialectal Arabic (DA) is the natural spoken variety used in daily communication of the Arabic World and is not generally written. Indeed, there is no commonly accepted standard for colloquial Arabic writing system [9]. Today, processing Arabic and spoken dialects is a new area of research that is faced with many challenges. On the one hand, the oral tradition of the DA and the absence of orthographic standards give rise to the difficulty of its automatic processing. These characteristics engender the lack of written resources for the colloquial Arabic. On the other hand, a number of dialects with linguistic differences on phonological, morphological, syntactic, and lexical levels increase the challenges of building tools and resources for all the Arabic dialects.

The training and testing of statistical or symbolic systems in Natural Language Processing (NLP) require the availability of annotated corpora. Our aim consists on developing resources and tools for the Spoken Tunisian Arabic (STA); one of the North African dialects. In the recent years, many researchers were interested in building dialectal Arabic corpora. Generally, the majority of these

contributions aim to develop textual DA corpora or to set spoken transcriptions without linguistic enrichments. That is why we have decided, as a first step in our work, to build a dialectal corpus of transcribed speech that treats multiple themes. Indeed, since the events of the revolution, the volume of data in Tunisian Arabic (TA) (oral and written) has increased. The TA has become progressively used and represented in interviews, news and debate programs instead of Modern Standard Arabic (MSA) [6]. To constitute the corpus, we are basically based on the available audio in the web in order to transcribe and annotate them. In this paper, we present our collection and transcription method for STA. We detail the different stages done to enrich the corpus with necessary linguistic and speech annotations.

Section 2 presents an overview of Tunisian Arabic. Section 3 presents the main related works. Section 4 is dedicated for presenting the transcription challenges for a language with oral tradition. We describe, in section 5, the collection and the transcription of the TA. Finally, we present, in section 6, the different types of annotation for enriching our corpus.

## 2 Tunisian Arabic

Tunisian Arabic (TA) is a dialect of the North African (i.e., the Maghreb) dialects spoken in Tunisia [29]. It is considered as a low variety given that it is neither codified nor standardized even though it is the mother tongue and the variety spoken by all the population in daily usage [24]. Approximately eleven millions people speak at least one of the many regional varieties of TA [29].

The linguistic situation in Tunisia is "poly-glossic" where multiple languages (French, Berber, Italian, etc.) and language varieties coexist (MSA and TA) [16]. Indeed, there are many differences as well as similarities points between TA and MSA in different levels [3]. In addition, TA is distinguished by the presence of words from several other languages. The presence of these languages mainly occurred due to historical facts. We find in Tunisia, a numerous examples of several languages; We find a significant number of expressions and words from European languages such as Spanish, French and Italian, Turkish, and even Maltese. In addition, TA contains several words from the vocabulary of Berber language [29]. Likewise, Tunisian people code switch easily and frequently between MSA, TA and the French language in a conversation [30]. This phenomenon allows the introduction of new words (nouns and verbs) derived from foreign languages.

## 3 Related Works

Today, processing Arabic and spoken dialects is a new area. There are many contributions aiming to develop textual DA corpora ([1], [15], [20], [25], [28], etc.). But, the resources created for Arabic dialects are still in its infancy. The lack of DA corpus is due to the absence of written standards and the lack of

[3] For more details see [29] and [30].

written material for Arabic dialects. We present in this section some works done for building dialectal Arabic corpora.

At present, the major standard Arabic dialects corpora are available through the Linguistic Data Consortium (LDC) by the DARPA EARS program [17] for developing robust speech recognition technology. The LDC provides conversational and broadcast speech with their transcripts. The Levantine Arabic is the object of the LDCs Fisher Levantine Arabic project in which more than 9,400 speakers of the Northern, Southern and Bedwi dialects of Levantine Arabic were participated in collecting 2,000 telephone calls [17]. Furthermore, the data set contains approximately 250 hours of telephone conversations. Each call is up to ten minutes in duration and subjects speak to each other about assigned topics. Moreover, another Arabic colloquial corpus called CALLHOME Egyptian Arabic Speech was dedicated to the Egyptian Arabic ([12], [8]). Indeed, the data set consists of 120 telephone conversations between native speakers of Egyptian dialect. All calls lasted up to 30 minutes. In fact, Egyptian Arabic corpus contains both dialectal and MSA words forms.

As well, the Saudi Arabic dialect was represented by the Saudi Accented Arabic Voice Bank which is very rich in terms of its speech sound content and speaker diversity within the Saudi Arabia [2]. The duration of the total recorded speech is 96 hours distributed among 60,947 audio files. Indeed, the corpus was externally validated and used by IBM Egypt Branch to train their speech recognition engine. English-Iraqi corpus is another Arabic corpora mentioned in [23] and which consists of 40 hours of transcribed speech audio from DARPA's Transtac program. Furthermore, Almeman et al. [3] have building a multi dialect Arabic speech parallel corpus. This corpus is designed to include three Arabic dialects which are the Egyptian dialect, the Gulf dialect, the Levantine dialects and the MSA. The created corpus is limited to a specific domain which is the travel and the tourism. Their corpus is composed of parallel prompts written for the four main varieties of the Arabic language which involved 32 speech hours. It is composed of written MSA prompts translated to dialects and then recorded [3]. To our knowledge, there is no conventional standard used for transcribing dialects.

In a like manner, Masmoudi et al. [18] and Graja et al. [10] have created speech corpora for STA. To our knowledge, they are the only researchers who created a spoken transcribed corpus for TA. The works done consists on building a corpus for limited domain which gathers a set of dialogues between the staff of the National Company of the Tunisian Railways and the customers who seek information about train schedules, fares, reservation, etc. The sizes of the two corpora are respectively 20 hours and 8 hours of transcribed speech.

## 4 Transcription Challenges for a Language with Oral Tradition

The transcription task is the operation which consists of replacing each phoneme and sound in the signal audio to a grapheme language of a writing system [19]. Transcription itself is not a trivial task. It requires a series of operations: choice

of mode transcription, transcription conventions, segmentation, translation, etc. [19]. Transcribing spoken languages that don't have orthographies rules face to many difficulties. The transcriber is faced to two types of problems. The first problems are related to the nature of the oral speech. This kind of problems is shared between all spoken languages. The second type of the problems is associated to the nature of the treated language which is mainly spoken without written tradition. Indeed, any transcriber is faced to many difficulties caused by the speech perception [19]. The bad quality of the signal or the record, the noisy environment, the presence of too many speakers, the overlaps, etc. are the principal causes of the bad listening of the speech and consequently the presence of the errors in the transcripts. Similarly, many transcribers do not always listen to the same sequence of words as others and altogether they may not hear a word or a sequence of words. The wrong listening prevents the transcriber to write faithfully the realized signal. Despite, the transcriber tries involuntarily to correct partial or incorrect words. He always essays to make sense with the perceived elements. These problem increases when the speech is spontaneous dialogue where speakers intersect speech, hesitate, stutter, etc. So, it is difficult to discern and to identify what whom and when is actually spoken.

On the other hand, TA like other Arabic dialects has no standard orthography. Despite, the TA is a variant of the Arabic language. The relationship between these varieties of Arabic does not prevent the differences between them. It is characterized by a rich lexicon which is a collection of words from Latin, Turkish, Berber, and Maltese origin. The massive borrowing and integration to TA has caused the presence of new phonemes that does not belong to the Arabic language. These particularities of the TA make the transcriber unable to write some words even some expressions. Owing to the absence of orthographic rules, the transcription of TA could be with different characters: Arabic characters, Latin characters, alphanumeric (SMS language), etc. This issue presents another challenge for transcribing TA.

In conclusion, most of the difficulties of oral arise in the preliminary phase for the transcription. There is a competition between different alphabets and therefore several orthographic traditions, and transcription tools, methods, and conventions to adopt [19].

## 5 Spoken Tunisian Arabic Corpus "STAC"

The transcribing process consists of two basic steps. The first one is providing voice data in order to be transcribed later. The second step consists in its transcription following transcription guidelines. More details about those two steps are described in the following sections.

### 5.1 Data Collection and Description

The first step in our method for corpus creation is the collection of spoken data. The choice of the speech data content and type is a very important step and

could be the key of further use of our corpus. We choose to provide both manually recorded speech audio (part 2) and audio files downloaded from the web (part 1) in order to improve the reuse of our corpus in new research tendency.

Indeed, there are a lot of free resources available on the web. To facilitate the task of creating a corpus for the STA, we are based on the approach "download" and "save" proposed by [27] for searching and downloading audio files who's the speakers express with Tunisian dialects. Following this approach, we recorded 3 hours and 28 minutes of TA speech from different TV channels and radio stations (*Mosaique radio*, *Tunisian national TV*, *Ettounsiya TV* and *Sfax radio*). It presents a first part of our corpus. These streaming are generally radio and television talk shows, debates, and interactive programs where the general public is invited to participate in discussion by telephone. Having a good amount of spoken recordings is fundamental in the design of the corpus. Also, a high sound quality is required and will be useful for other future processing for example in voice recognition system. To keep the good quality of our corpus, we saved only the files in which the speakers can intervene only on one subject simultaneously. Similarly, we choose records with a good sound quality. Sometimes, the quality sound of the recording can vary considerably over time. So, we filter noisy sequences (music or other non-transcribed noise) that last more than one second. The size of the audio files can vary from several tens of seconds to several minutes. We take care that all records contain more spontaneous speech and the percentage of the dialectal content is very higher than MSA or French content.

The second part of our corpus is about 30 minutes taken from the corpus TuDiCoI (Tunisian Dialect Corpus interlocutor) [10] which is a corpus of spoken dialogue in TA and it is obtained from a railway information service. It gathers a set of conversations recorded in the railway station between the staff and customers who request information about the departure time, price, booking, etc. [10]. We have redone the transcription of this part following our convention guidelines.

Including different themes and speakers of TA make our corpus more generic as possible. It contains spontaneous speech, less spontaneous speech and sometimes prepared speech. In addition, the relatively big number of speakers (about 70 speakers) in our corpus speaking each one with its own style make our corpus a representative sample of TA. We provide both individual and multiple speakers in our collection to identify different aspects of conversational speech. Also, the radio and television records have a varied content. There are a wide variety of speakers and themes (social, health, religious, political, and others). Providing speech data with a variety of themes will increase the size of the vocabulary in our corpus and will be very useful for further application for example theme classification [5]. Indeed, we defined the following themes list in our data selection: religious, political, health, social, and others. The corpus contains dialect data from different Tunisian regions. The dialect of Tunis (the capital of Tunisia) is the most dominant while it is the dialect used in the Tunisian media. We consider it as the standard dialect of Tunisia because it is understood by all Tunisian people. It presents about 90% of the totality of our corpus. The table 1 presents some statistics about our corpus.

**Table 1.** Statistic about STAC corpus.

| **Themes** | **Duration** |
|---|---|
| Social | 01:01:35 |
| Health | 01:30:46 |
| Religious | 00:12:38 |
| Political | 00:50:50 |
| Other | 01:14:42 |
| **Total** | **04:50:31** |
| TA percentage | 97.20% |
| MSA percentage | 0.37% |
| French language percentage | 2.43% |

There are a few works done for creating corpus for STA ([18] and [10]). To our knowledge, our corpus is the first resource for TA which contains different types of annotation and enrichments. It can be a good resource spoken for TA and can be used for different purposes. Despite, the size is relatively small. The total size of our corpus is about 5 hours. It is growing, since there are still new recordings that are planned to being done. In conclusion, STAC could be a good textual resource for STA. It could be useful for creating tools and other resources for TA.

### 5.2 Corpus Transcription

The transcription of spoken speech is a symbolic representation of the spoken language. In the literature, the researchers have defined fours types of transcriptions [19]: phonetic, phonologic, morpho-phonologic, and orthographic. First, the phonetic transcription consists on describing as closely as possible the differences between sounds using the phonetic alphabets: IPA (International Phonetic Alphabet) or SAMPA (Speech Assessment Methods Phonetic Alphabet). It presents the pronunciation of the speech. This type of transcription is expensive in term of time because the transcribers are not familiarized with phonetic alphabets. Second, the phonological transcription consists on describing only the distinctive phonological differences. Then, the morpho-phonologic transcription is a combination between phonological and syntactical notation [19]. It is an analysis and decomposition of all speech constituents. This type of transcription is very expensive in time of writing. Finally, orthographic transcription is a transcription method that employs the standard spelling of the target language. It is easier to the transcriber to use the usual alphabets for a language. After studying these types of transcription, we choose the orthographic transcription. Our choice is justified by different reasons. First, according to our objectives which are the tools development and adaptation in the favour of automatic treating STA, the best choice for our objective is to use the orthographic transcription with the Arabic alphabets. Likewise, the orthographic transcription is easier to the transcriber than to use the phonetic alphabets. Second, we create a corpus, which is easy to read by everybody.

To transcribe our corpus, we choose to use Praat tool[4]. Indeed, the choice was related to our research group needs. This tool allows the analysis of speech in phonetics and also supports speech synthesis, including articulator synthesis[5]. Praat can provide an aligned transcription between speech and text and facilitate the labeling and segmentation of the speech for the linguistic issues due to the tires provided by its interface. The acoustic signal of audio content may correspond to speech, music, noise or/and mixtures of them. Furthermore, in the same record, one or more speakers coexist and discuss many topics in the same time. Also, the recording sound quality may vary significantly over time. As a consequence, it is important to define a set of conventions that specify the orthographic transcription of the different elements which exist in the acoustic signal.

We utilised two works done for TA for transcribing our corpus. The first work is for [30] who presented an Orthographic Transcription of TA (OTTA). The second work is for [29] who proposed a conventional orthography for TA (CODA Tun.) which is an extension of the CODA map [11] to the TA.

**Arabic Orthographic Transcription.** To standardize the orthographic transcription of the TA, we applied the orthographic transcription conventions of TA defined in OTTA and in CODA Tun. The difference between the two conventions is in the phonetic level. Indeed, CODA Tun. is defined in writing TA words without specifying the phonetic differences between MSA and TA in some cases. Contrariwise, OTTA mixes widely between the phonetic and the orthographic transcription. Hence, we created two versions of our corpus using these conventions. Subsequently, the obtained corpus based on two transcription conventions will be useful for the creation of processing tools for TA such as stemmer, morph-syntactic tagger, etc. Also, it is useful for the creation of automatic speech processing systems such as speech synthesis, automatic transcription systems, etc. [30]. Table 2 presents an example of a TA sentence written according the two orthographic conventions.

**Enrichment of the Transcription.** Textual transcription of the speech is not sufficient for developing tools and applications for processing automatic language. Moreover, the standard orthographic transcription doesn't take into consideration the observed phenomena of speech (elisions, disfluency, liaison, noise, etc.). So, we add some orthographic enrichment for the transcribed speech. We followed the enriched orthographic transcription guidelines described in the OTTA directive. The guideline proposed by [30] is an adaptation of the Enriched Orthographic Transcription (TOE in French) [4] which is elaborated by LPL laboratory to transcribe conversational French corpus. The French convention is extended with some precisions and modifications for the transcription of STA [30]. We present here some specifications that we used while transcribing our corpus.

Our transcription method consists of an orthographic transcription that it specifies the typical phenomena of the oral. Indeed, the transcript should be close

[4] `http://www.fon.hum.uva.nl/praat/`
[5] `http://en.wikipedia.org/wiki/Praat`

**Table 2.** An example of sentences in TA.

| | |
|---|---|
| **OTTA** | وباش نحكيو زادة عالزيادات في أسوام القاز ونسألوا عالاءنعكاسات متاعو عالمواطن. |
| | wbAš nHkyw zAdħ ϛAlzyAdAt fy ÂswAm AlGAz wnsÂ lwA ϛAlĂnϛkAsAt mtAϛw ϛAlmwATn.[6] |
| **CODA Tun.** | وباش نحكيوا زادة عالزيادات في اسوام القاز ونسألوا عالانعكاسات متاعه عالمواطن. |
| | wbAš nHkywA zAdħ ϛlAlzyAdAt fy AswAm AlqAz wns ÂlwA ϛAlAnϛkAsAt mtAϛh ϛAlmwATn. |
| **Translation** | *We will talk also about the increase in gas prices and question its impact on the citizen.* |

to the signal. We try to write the speech of each speaker and the overlaps speech as possible as we can hear. Therefore, we use neither acronyms nor abbreviation in the transcripts. Similarly, we do not correct the atypical accords. STAC corpus is composed of conversational speech. The script of each speaker is presented separately in an individual tier. Silence pauses could be at the beginning, mixed with the transcript, and at the end of a speaker Turn. We isolated pauses (silent and noisy) which have a minimal duration of 200 ms in a speaking turn. We mark them with the hash tag symbol "#". Also, we mark the silences those are lesser than 100 ms with the plus symbol "+".

Code switching is a main characteristic of the STA. The orthographic form of each non TA word should refer to its orthographic rules. We use this annotation [lan:X, text] for non TA words to be easily recognized (e.g. [lan:FR, deux] for French language). Table 3 describes some annotations used in our corpus.

## 6 Corpus Annotation

The enrichment of the corpus is key part of NLP. Many systems now are easily developed due the availability of annotated corpora for written and spoken data. Many linguistic annotations could be added to a spoken corpus. Indeed, transcriptions of speech do not contain punctuation marks. Texts contain lexical particularities specific to speech; spoken texts are full of disfluencies. Nevertheless, most of NLP tools should consider these specificities in order to perform the proposed task. For this purpose, we enriched our corpus with different types of annotation (morphosyntactic annotation and disfluencies).

### 6.1 Morpho-syntactic Annotation

The morpho-syntactic annotation consists on marking up a word in a corpus as corresponding to a particular part of speech, based on both its definition, as well as its context. The manual annotation of the corpus is very difficult and

[6] Transliteration is coded with Buckwalter transliteration. For more details about it, see [13].

**Table 3.** Some annotations used in STAC corpus.

| **Transcribed event** | **Notation** | **Representation** | **Examples** |
|---|---|---|---|
| Proper name | Between two hooks | [ortho[7], type] | [صفاقس، مك]<br>[*Sfax, place*] |
| Elisions | The characters related to the omitted phonemes are written between parentheses | otho(c) | ب(ا)ش<br>*not* |
| Non-linguistic noises and Inaudible sequence | Star | * | * مشى<br>* *he walks* |
| Laughers | Ampersand | & | & |
| Tunisian dialect liaison [8] | Between equal signs | otho = letter = ortho | أربعطاش=ن=ألف<br>*fourteen* |
| Reported speech | Between paragraph symbols | \sequence\ | قلت لك\اسكت\<br>*I told to you \shut up* |
| Specific pronunciation | Between square brackets | [ortho, buck[9]] | [جزار, zazza:r] *butcher* |
| Speech while laughing | Between double ampersands | &&ortho&& | باهي&&*ok ok*&& |
| Title | Between quotation marks | "ortho" | " عندي ما نق(و)ل لك "<br>*"I have something to say"* |
| Truncated word | Final dash | ortho- | عس- |

expensive in time. Eventually, the development of NLP tools for TA is still in its infancy. We present in this section our method for morpho-syntactic annotation of TA.

The main idea of our method is to disambiguate the output of the morphological analyzer developed for TA, $Al-Khalil$ [31]. Before presenting our method, we should define the tags set that we used for annotating our corpus. Usually, Al-Khalil TA version returns a list of analysis for a given word with different information (gender, prefix, suffix, number, person, voice, POS, etc). For our task, we keep all these morphological characteristics. We add four grammatical categories (i.e. part of speech) which are related to STA. We define NE, FW,

[7] "Ortho" is the orthographic transcription.
[8] The following rules are applicable when we used OTTA transcription guidelines.
[9] "buck" is the transliteration of Buckwalter [13].

FP, and ONOM for respectively named entity, foreign word, filled pause, and onomatopoeia.

The annotation process starts by extracting speaker's text. STAC corpus incorporates the transcription of many conversations between at least two speakers. Hence, speech text for each speaker is divided into many speech turns. We gather the speech turns for each speaker in a unique text. We, then, segment it manually in utterances. We consider an utterance a semantically meaningful unit. The automatic identification of utterances boundaries is considered as a future work. Speech text includes many annotations; some of them are very useful in the morpho-syntactic annotation process. Some of others such as noise and music are removed. Furthermore, all foreign words, filled pauses, onomatopoeia, and also named entities existing in the corpus are not analysed by $Al-Khalil$ [31]. We tagged them respectively as FW, FP, ONOM and NE.

To annotate our corpus, we utilized an iterative procedure to semi-automatic tagging the unannotated data. The iterative procedure starts by dividing our corpus to 10 folders according to the number of sentences. We begin with a morphological analysis of the first folder of the corpus. Eventually, the analyzer gives to each word a set of analysis. We choose the correct analysis according to the position of the word in the utterance. When the analyzer fails in giving an analysis for a word, we determine for this word a set of morphological features. To decrease the number of unrecognized words, we remove all diacritics from the words when analyzing and we enhance also the Al-Khalil lexicon. We train a first version of our POS tagger with the first part of the corpus completely annotated by hand. We applied a multi-class classifier using a rules-based classifier (Ripper) as the main classifier. The generated rules are used for choosing the correct analysis for each word. We use the result tagger for annotating the second folder of the corpus. We manually corrected the output of the tagger, and added the corrected part to the training corpus. Then, we iterate this process over the different parts of the corpus. At the end of this process, we obtain a larger manually corrected corpus (i.e. for each word, all the morphological analysis are kept and only the correct analysis in a given context is marked). The corpus consists of 42,388 words that are transcribed. It is composed of 2,252 verbs, 1,457 nouns, and 458 adjectives. The proposed method achieved an F-measure score of 87%.

### 6.2 Disfluencies Annotation

Spoken corpus annotation should not be limited to the usual annotations of the written language. Indeed, a speech transcription is a new form of texts with specificities which constitute a practical issue for automatic analysis of spoken texts [7]. So, spoken corpus annotation must take into consideration these specificities, more precisely, the "disfluencies". Indeed, disfluencies are defined as a phenomenon occurring frequently throughout spontaneous speech, and consist of the interruption of the normal course of speech [14]. In fact, there are different types of disfluencies [22]: filled pauses, repetition of words or word sequences, immediate self-corrections, word fragments. Generally, disfluencies

could be combined simultaneously with the association of at least two of phenomena mentioned above. The analysis of disfluencies realized by Shriberg [26] showed that the disfluent sequence can be divided into three regions: the reparandum (word truncation or phrase truncation), break point (filled pauses, silent pauses, etc.) and the repair. Based on Shriberg analysis, Pallaud et al. [21] have defined an annotation schema which reflects the proposed structure of disfluencies. The guideline of annotation is developed in order to annotate disfluencies using Praat. Given the specificities of the TA, especially the code switching, we noticed the presence of some cases of repetition code switching. So, we added some annotations that describe this type of repetition which is specific for the TA. Figure 1 presents an annotated example extracted from our corpus.

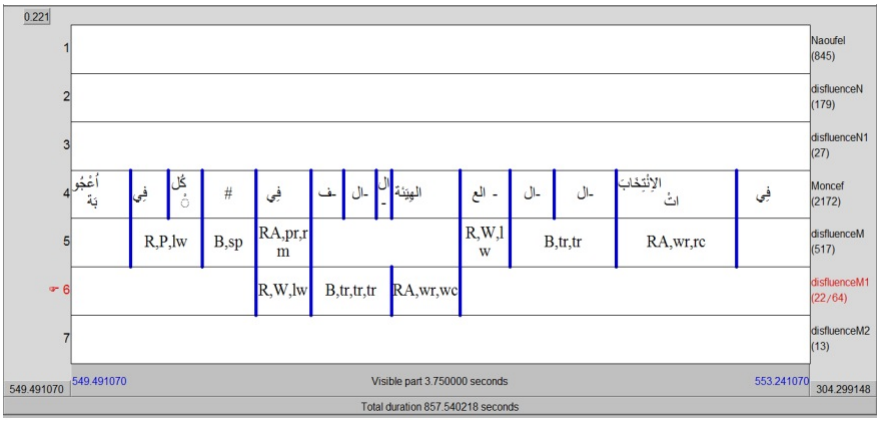


**Fig. 1.** An example of disfluencies annotation

## 7 Conclusion and Future Works.

The creation of an annotated corpus presents a challenge for NLP application. Indeed, the Tunisian Arabic is a language with oral tradition without orthographic rules. This issue makes the creation and the annotation of a corpus for this language more difficult. In this context, we presented our effort for the creation of an annotated corpus for spoken Tunisian Arabic. We presented, firstly, the speech data collection and its transcription according to our orthographic transcription guidelines. Then, we describe its enrichment with morpho-syntactic and disfluencies annotations.

Our corpus consists of 5 transcribed hours and we plan to extend it (i.e. increasing the size of the corpus and adding other themes and domains) basing on the proposed methods of transcription and annotation in order to make the corpus as representative as possible of the spoken Tunisian Arabic. In addition, we plan to use this corpus for developing tools for processing Tunisian Arabic.

Developing a POS Tagger for spoken Tunisian Arabic is the main focus at the moment. We intend also to develop a tool that allows the automatic detection of disfluencies in order to consider these phenomena in automatic parsing spoken Tunisian Arabic.

## References


1. Al-Sabbagh, R., Girju, R.: YADAC: Yet another Dialectal Arabic Corpus. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, May 23-25. pp. 2882–2889 (2012)
2. Alghamdi, M., Alhargan, F., Alkanhal, M., Alkhairy, A., Eldesouki, M., Alenazi, A.: Saudi Accented Arabic Voice Bank. Journal of King Saud University - Computer and Information Sciences, Riyadh, 20, pp. 45–64 (2008)
3. Almeman, K., Lee, M., Almiman, A.A.: Multi dialect Arabic speech parallel corpora. In: First International Conference on Communications, Signal Processing, and their Applications (ICCSPA). pp. 1–6 (2013)
4. Bigi, B., Péri, P., Bertrand, R.: Orthographic Transcription: which enrichment is required for phonetization? In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, May 23-25. pp. 1756–1763 (2012)
5. Bischoff, K., Firan, C.S., Paiu, R., Nejdl, W., Laurier, C., Sordo, M.: Music Mood and Theme Classification - a Hybrid Approach. In: International Society for Music Information Retrieval (ISMIR 2009). pp. 657–662 (2009)
6. Boujelbane, R., Khemakhem, M.E., Belguith Hadrich, L.: Mapping Rules for Building a Tunisian Dialect Lexicon and Generating Corpora. In: Sixth International Joint Conference on Natural Language Processing (IJCNLP 2013), Nagoya, Japan, October 14-18. pp. 419–428 (2013)
7. Dister, A., Constant, M., Prunelle, G.: Normalizing speech transcriptions for Natural Language Processing. In: 3rd International Conference on Spoken Communication (GSCP'09), Naples, Italy, Feb. pp. 507–520 (2009)
8. Duh, K., Kirchhoff, K.: Lexicon Acquisition for Dialectal Arabic Using Transductive Learning. In: Proceedings of EMNLP'06, Sydney, Australia, July 22-23. pp. 399–407 (2006)
9. Elmahdy, M., Gruhn, R., Minker, W., Abdennadher, S.: Modern Standard Arabic Based Multilingual Approach for Dialectal Arabic Speech Recognition. In: The Eighth International Symposium on Natural Language Processing (SNLP), Bangkok, Thailand (2009)
10. Graja, M., Jaoua, M., Belguith Hadrich, L.: Discriminative Framework for Spoken Tunisian Dialect Understanding. In: Proceedings of Statistical Language and Speech Processing - First International Conference, (SLSP 2013), Tarragona, Spain, July 29-31. LNCS, vol. 7978, pp. 102–110 (2013)
11. Habash, N., Diab, M., Rambow, O.: Conventional Orthography for Dialectal Arabic. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, May 23-25. pp. 711–718 (2012)
12. Habash, N., Eskander, R., Hawwari, A.: A Morphological Analyzer for Egyptian Arabic. In: Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology, Montréal, Canada. pp. 1–9 (2012)
13. Habash, N., Soudi, A., Buckwalter, T.: On Arabic Transliteration. In: Arabic Computational Morphology: Knowledge-based and Empirical Methods (2007)

14. Heeman, P., Allen, J.: Detecting and correcting speech repairs. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics. pp. 295–302 (1994)
15. Jarrar, M., Habash, N., Akra, D., Zalmout, N.: Building a Corpus for Palestinian Arabic: a Preliminary Study. In: Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing, Doha, Qatar, October 25. pp. 18–27 (2014)
16. Lawson, S., Sachdev, I.: Codeswitching in Tunisia: Attitudinal and behavioural dimensions. Journal of Pragmatics 32(9), pp. 1343–1361 (2000)
17. Maamouri, M., Buckwalter, T., Cieri, C.: Dialectal Arabic Telephone Speech Corpus : Principles, Tool design, and Transcription Conventions. In: NEMLAR International Conference on Arabic Language Resources and Tools, Cairo, Egypt (2004)
18. Masmoudi, A., Estève, Y., Khemakhem, M.E., Bougares, F., Belguith Hadrich, L.: Phonetic tool for the Tunisian Arabic. In: SLTU'2014, Saint-Petersburg, Russia (2014)
19. Moukrim, S.: Morphosyntaxe et sémantique du "présent" Une étude contrastive à partir de corpus oraux Arabe marocain, berbère tamazight et français (ESLO/LCO). Thesis, Université d'Orléans, December (2010)
20. Mubarak, H., Darwish, K.: Using Twitter to Collect a Multi-Dialectal Corpus of Arabic. In: Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Doha, Qatar, October. pp. 1–7 (2014)
21. Pallaud, B., Blache, P., Bertrand, R.: Codage des annotations de disfluences dans les corpus du CID. pp. 1–5
22. PIU, M., Bove, R.: Annotation des disfluences dans les corpus oraux. In: RÉCITAL 2007. pp. 5–8. Toulouse, France (2007)
23. Precoda, K., Zheng, J., Vergyri, D., Franco, H., Richey, C., Kathol, A., Kajarekar, S.S.: Iraqcomm: a next generation translation system. In: INTERSPEECH 2007, Antwerp, Belgium, August 27-31. pp. 2841–2844 (2007)
24. Saidi, D.: Typology of Motion Event in Tunisian Arabic. In: LingO. pp. 196–203 (2007)
25. Salama, A., Bouamor, H., Mohit, B., Oflazer, K.: YouDACC: the Youtube Dialectal Arabic Comment Corpus. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland (2014)
26. Shriberg, E.E.: Preliminaries to a Theory of Speech Disfluencies. Tech. rep. (1994)
27. Waibel, A., Schultz, T., Vogel, S., Fugen, C., Honal, M., Kolss, M., Reichert, J., Stuker, S.: Towards language portability in statistical speech translation. In: Proceedings of ICASSP'04, May. vol. 3, pp. iii–765–8 vol.3 (2004)
28. Younes, J., Souissi, E.: A quantitative view of Tunisian dialect electronic writing. In: 5th International Conference on Arabic Language Processing, Oujda, Morocco (2014)
29. Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Belguith Hadrich, L., Habash, N.: A Conventional Orthography for Tunisian Arabic. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'2014), Reykjavik, Iceland, May 26-31. pp. 2355–2361. ELRA (2014)
30. Zribi, I., Graja, M., Khmekhem, M.E., Jaoua, M., Belguith Hadrich, L.: Orthographic Transcription for Spoken Tunisian Arabic. In: 14th International Conference CICLing 2013, Proceedings, Part I, Samos, Greece, March 24-30. LNCS, vol. 7816, pp. 153–163. Springer (2013)
31. Zribi, I., Khemakhem, M.E., Belguith Hadrich, L.: Morphological Analysis of Tunisian Dialect. In: Sixth International Joint Conference on Natural Language Processing (IJCNLP 2013), Nagoya, Japan, October 14-18. pp. 992–996 (2013)