SN Computer Science (2024) 5:440
https://doi.org/10.1007/s42979-024-02775-8

ORIGINAL RESEARCH

Detecting Speech Disorders Using A Machine-Learning Guided Method in Spontaneous Tunisian Dialect Speech

Emna Boughariou[1] · **Younès Bahou**[2] · **Lamia Hadrich Belguith**[1]

Received: 26 October 2022 / Accepted: 5 March 2024

Abstract
This work investigates the disfluencies processing task within the natural spoken language comprehension field. We present a transcription-based method with purely linguistic features for detecting disfluencies in spoken Tunisian dialect transcriptions. Disfluencies processing is the task of detecting spontaneous disorders in spoken language transcripts, distinguishing between fluent and disfluent words. The originality of this method is that several disfluency types are processed automatically and for wide domains in the spontaneous spoken Tunisian dialect. Likewise, it incorporates various linguistic features such as morpho-syntactic labels and word synonyms. Syllabic elongations, speech words, word-fragments, and simple repetitions are carried out according to the rule-based approach, while complex repetitions, insertions, substitutions, and deletions are detected using a transition-based model through the machine learning approach. We compare the transition-based model to the sequence-tagging-based model presented in the previous work. Experiments show that both models are relevant to the disfluencies detection task in the spoken Tunisian dialect, the F-Measure rates are respectively 79.81% and 78.97%.

Keywords Disfluencies · Tunisian Dialect · Transcription-based method · Transition-based model · Linguistic features

Introduction

The Tunisian Dialect (TD) is the official mother tongue spoken by all Tunisians, regardless of their origin and social affiliation. It is a subgroup of Arabic dialects usually identified with Maghreb Arabic. The TD has sparked increased interest in the field of the NLP community, like Arabizi transliteration [21], speech recognition [22], dialect identification [4], word-to-word translation [13], sentiment analysis [24], among others. The TD is not standardized or taught, and has no official status. Nevertheless, most Arabic native speakers can not produce a sustained spontaneous speech in Modern Standard Arabic (MSA); in unwritten situations where spoken MSA would normally be expected. Hence, most Tunisian speakers typically resort to frequent code-switching between their dialect and MSA in their daily lives such as in talk shows on radio and TV channels [19]. Regarding this disturbance in spoken TD speech, syntactical and grammatical speech errors, mostly called disfluencies, can occur frequently in all forms of spontaneous speech, whether casual discussions or formal arguments [32].

Disfluencies are characteristic of spontaneous speech that make it different from written text [10]. They are additional speech noises corrected by the speaker and may affect the grammatical flow of utterances. Disfluencies detection presents a significant challenge for tasks dealing with spontaneous speech processing, such as parsing, machine translation, dialogue systems, and other NLP understanding tasks [6]. It helps to recognize disfluent sequences in spoken language transcripts or automatic speech recognition results.

Detecting disfluencies in spoken TD is rarely studied by researchers. Ref. [25] adopted a symbolic rule-based approach to delimit and correct disfluencies automatically using a set of rules and patterns, in a very restricted domain and with a limited TD vocabulary. Ref. [34] proposed transcription conventions to annotate disfluencies in TD corpora that are used later to annotate manually only incomplete words, repetitions and onomatopoeia words in the TD corpus STAC [33].

This work is part of disfluencies processing in the TD. As far as we know, there is no work within the spoken TD processing field which consists of detecting and removing

✉ Emna Boughariou
emnaboughariou@gmail.com

[1] Sfax University, Sfax, Tunisia

[2] HA'IL University, Hail, Kingdom of Saudi Arabia

Published online: 17 April 2024

several types of disfluencies automatically and from open domain transcriptions. In this perspective, our study aims to propose an original method for processing eight types of disfluencies, obviously, syllabic elongations, word-fragments, speech words, simple and complex repetitions insertions, substitutions and deletions. In previous work, we presented a method for detecting and removing disfluencies in TD transcripts, which provides a rule-based approach for simple disfluencies processing and a sequence-based model within the Machine Learning (ML) approach for complex disfluencies processing. The major contributions of this paper, mainly concern the detection of complex disfluencies task. We present a transition-based model for complex disfluencies detection and we compare it with the sequence-based model.

This paper is structured as follows: We present a background of disfluencies processing in spontaneous spoken TD in Section "Background of Disfluencies Processing in TD". We give an overview of the building process of our TD corpus and an analysis of its most significant characteristics in Section "Data. We present our contribution to detecting disfluent regions of utterances in Section "The Proposed Method for Disfluencies Processing in TD". We review the previous work for complex disfluencies processing in TD and we show the comparison results and the critical analysis of both contributions in Section "Experiments and Discussion", before drawing our conclusion and major future work in Section "Conclusion and Perspectives".

Background of Disfluencies Processing in TD

We expose in this section, the different types of disfluencies studied in our work as well as the challenges of handling these phenomena in the context of spontaneous TD speech processing.

Disfluencies Taxonomy

The TD usually includes eight types of disfluencies as follows [5]:

Syllabic elongations are abnormal vowel lengthening of a syllable lasting more than 1 s. In TD speech, the elongations appear usually with either the first (e.g., Utt1) or the last (e.g., Utt2) syllable of the word. The following examples illustrate the two cases:

Utt1: صووووتك موش واضح [Swwwtk mw$ wADH] (Your voiiiice is not clear).

Utt2: فمااااا ترات برك [fmAAAA trAn brk] (Thereee's only one train).

Speech words are characterized by the continuation of the acoustic signal generation during the pause period. They include filled pauses also called hesitations (e.g., اه [āh] (Ah)) and discursive markers (e.g., يعني [y'ny] (meaning)). They are the most frequent disfluencies used in spontaneous oral productions.

Word-fragments are "syllables, speech sounds or single consonants, which are similar to the beginning of the next fully articulated word... [and] they may neither be equal to the whole next word" [11]. They are truncated words started and interrupted by the same speaker (e.g., Utt3). They may be dropped, taken, or replaced.

Utt3: وقتاش يبدا ال المؤتمر [wqtA$ ybdA Al AlmWtmr] (When the the-conference starts).

Simple repetitions are words that occur several times consecutively (e.g., Utt4).

Utt4: ثلاثة أمم ثلاثة وزراء [vlAvp Omm vlAvp wzrA'] (Three emm three ministers).

Complex repetitions can be either one word that is repeated not consecutively (except speech words) in the utterance (e.g., Utt5) or a group of words identically repeated (e.g., Utt6).

Utt5: علاش هذا التبذير علاش [ElA$ h*A Altb*yr ElA$] (Why this extravagance why)

Utt6: بعد ثلاثة سوايع امم ثلاثة سوايع يبدا [bEd vlAvh swAyE Amm vlAvh swAyE ybdA] (After three hours emm three hours starts)

Insertions are the case of correcting a part of the speech by adding new words (e.g., Utt7).

Utt7: عندك مثال كان عندك مثال [Endk mvAl kAn Endk mvAl] (you have an example if you have an example)

Substitutions are the case of correcting a part of the speech by replacing some words with new ones (e.g., Utt8).

Utt8: موضوع يهم المسنين أهه سامحني الشباب [mwDwE yhm Almsnyn Ohh sAmHny Al$bAb] (A topic that interests the elderly euuh forgive me the youth people)

Deletions are the case of correcting a part of the speech by removing words (e.g., Utt9).

Utt9: ترات نورمال إلي يخرج تو ترات إلي يخرج تو [trAn nwrmAl Ily yxrj tw trAn Ily yxrj tw] (Normal train that leaves now train that leaves now)

Disfluencies typology depends usually on the concerned language. For the TD language, 38% of elongation cases affect the first syllable of the word [6]. For research dealing with French, phonological characteristics like schwa and assimilation can be considered as a disfluencies type [7].

The heterogeneity in how the different disfluencies can be detected, brought us to use the following taxonomy: Simple Disfluencies and Complex Disfluencies. Simple disfluencies affect only one minimal token, they include syllabic elongations, speech words, word-fragments, and simple repetitions [5]. Complex disfluencies affect several tokens and may break the morphological flow of the utterance, they include complex repetitions, insertions, substitutions, and deletions. According to [29], complex disfluencies are typically

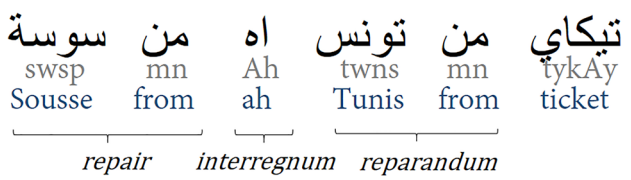

Fig. 1 Sentence with disfluencies annotated according to Shriberg (1994)

assumed to have a tripartite reparandum-interregnum-repair structure as shown in Fig. 1.[1]

The reparandum is the disfluent portion of the utterance that is corrected or abandoned. The interregnum (also called editing term) is the optional portion of the utterance; it could include speech words. The repair is the portion of the utterance that corrects the reparandum.

Challenges for Disfluencies Detection in TD

Detecting disfluencies is a challenging task for spoken dialects including the TD, although this is the most natural, easy, and spontaneous means of communication. The dialects are not standardized, they are not taught, and they have no official status.

Pronunciation variability. The same utterance can be expressed in different ways between speakers to regions, social classes, ages, and areas. The TD is divided into several dialectal areas according to the Tunisian regions. The vocabulary varies through areas, involving phonological, morphological, lexical, and syntactic variations. The personal pronoun 'I' in English is pronounced "آنا [lnA]" in Tunis and Sfax, "آني [lny]" in Sahel, "ناي [nAy]" in El Kef, etc.

Dialogue context. Spontaneous dialogue may contain sub-dialogues of reformulation, clarification, or rectification. Likewise, the speaker can expand his idea over several turns. Several disfluencies cases can be caused by the absence of the dialogue context. In the dialogue below, "الزوز" and "الوحدة" are considered literally as substitution disfluencies, while it is not the case when considering the context of the first speaker turn.

- Speaker 1: سوم الوحدة والا الزوز؟ [swm AlwHdp wAlA Alzwz?] (the price of the once or of both)

- Speaker 2: الزوز, الوحدة بعشرين دينار [Alzwz, AlwHdp bE$ryn dynAr] (both, the once is for twenty dinars)

Out-of-vocabulary words. OOV words are speech recognition errors that correspond to insertions, deletions, and confusions of words generated by automatic recognition systems [3]. OOV words include unknown words that are non-existent words in the recognition language model or in the lexicon, they could be truncated words (e.g., Utt10) and miss-recognized words which are produced in the output of speech recognition, while other words were pronounced (e.g., Utt11).

Utt10: من تون من تونس [mn twn mn twns] (From Tun from Tunis)

Utt11: نحب نمشي لمارس [nHb nm$y lmArs] (I like to go to March)

Long-range dependency. Repairs do not necessarily follow reparandums in some cases. They may be placed after several fluent words not belonging to the disfluency structure. In Utt12, the reparandum "المعلمين" is corrected after three fluent words using the repair "التلامذة" where "اه لا" is the interregnum.

Utt12: المعلمين الي عملو مسيرة اه لا التلامذة [AlmElmyn Aly Emlw msyrp Ah lA AltlAm*p] (the teachers who made a protest ah no the students).

Segmentation of utterances. Sentence boundary detection and disfluencies processing are both being studied increasingly in spontaneous speech studies. The probability of disfluencies was found to be exponentially proportional to the length of the utterance [2]. Utt13 contains a complex repetition of the word تيكاي [tykAy] (ticket) that appears in both positions 3 and 6. While considering its grammatical structure, we should notice that Utt13 must be split into two utterances as follows:

باهي اعطيني تيكاي [bAhy AETyny tykAy] (ok give me a ticket) and بقداه هي التيكاي [bqdAh hy AltykAy] (how much is the ticket).

Utt13: باهي اعطيني تيكاي بقداه هي التيكاي [bAhy AETyny tykAy bqdAh hy AltykAy] (ok give me a ticket how much is the ticket).

Irregular word order. The order of the words in oral utterances is not always respected and this is without affecting the semantics conveyed. In a given Tunisian verbal utterance, the canonical word order can generally follow three syntactic structures, namely, Subject-Verb-Object (SVO), Verb-Subject-Object (VSO), and Object-Verb-Subject (OVS). Since TD is an irregular language, the syntactical order of words may change in the same disfluent utterance. In (Utt14), "وقتاش يخرج تران [wqtA$ yxrj trAn], VSO" is a repetition of "التران وقتاش يخرج [AltrAn wqtA$ yxrj], SVO" translated both to "when does the train leave".

Utt14: وقتاش يخرج تران نحب نعرف التران وقتاش يخرج لتوزر [wqtA$ yxrj trAn nHb nErf AltrAn wqtA$ yxrj ltwzr] (when does the train leave to Tozeur).

Compound words. Poly-lexical expressions such as compound words include several linguistic phenomena for which the syntactic and semantic properties only partially overlap. Disfluent compound words represent 2.7% of 4.6% cases of disfluent units in [9] study. Disfluencies can appear inside compound words (e.g., منزل امم بوزيان [mnzl emm bwzyAn]

[1] The translation from TD to English is right-to-left and word-for-word.

(Manzel Bouzayen), a commune in the Sidi Bouzid governorate in Tunisia) or in its outside (e.g., خا خارق للعادة [xA xArq llEAdp] (extraordinary)).

Voluntary repetitions. Voluntary word repetition is meant to highlight a description of reality and can also contribute to an argument. The repetition can then be either semantic (e.g., anaphora) or grammatical. In French, some words are repeated for syntactic reasons (e.g., *nous nous sommes*, for pronominal verbs). In TD vocabulary, we noticed voluntary repetitions used frequently, such as "تو تو [tw tw] (now)" and "كيف كيف [kyf kyf] (the same)".

Synonymy. The speaker may replace some words with their synonyms, which is unnecessary for the syntactic structure of the utterance. In Utt15, "مشى [m$Y]" may be the synonym of "خرج [xrj]". Besides, for languages that code-switch between several languages such as TD, a word can be repeated by its similar to other foreign languages.

Utt15: المدير مشى خرج [Almdyr m$Y xrj] (the director walked left)

Enumeration. An enumeration consists of successively detailing various elements of which a generic concept or an overall idea is composed. With the absence of coordinating conjunctions, the enumeration can be interpreted as a substitution disfluencies type in which the speaker replaces the words to catch up and correct his utterance, as the following example shows.

Utt16: باش يفسر يفهم [bA$ yfsr yfhm] (in order to explain understood)

Agglutination. Arabic TD is an agglutinative language. New dialectal affixes and suffixes are added when others are removed compared to the MSA morphology. The negation particle "ما [mA]" and the negation letter "ش [$]" attached respectively to the beginning and at the end of the verbs replace the negation particles of the MSA "لم [lm]" and "لن [ln]", for example, "مامشيتش [mAmyt] (I did not go)" whose root of the word is the verb "مشيت [m$yt] (I went)". Thus, the MSA interrogative clitics "ا [A]" and "هل [hl]" are replaced by "شي [$y]" attached to the end of the word, as in "مشيتشي [myty] (Did you go)". Likewise, the proclitics "ه [h]", "ع [E]" and "م [m]" agglutinated to the definite article "ال [Al]" are the result of reduction of close demonstrative pronouns ("هذا [h * A] (this)", "هذه [h * h] (this)", "هؤلاء [hWlA ’] (these, ’feminine / masculine plural’)", "هاتان [hAtAn] (these, ’feminine duel’)", "هذان [h * An] (these, ’masculine duels’)"), the preposition "من [mn] (de)" and the coordinating conjunction "مع [mE] (with)", respectively.

Diacritics lack. The absence of Arabic vowels (i.e., diacritical marks placed above or below the Arabic letters) leads to lexical ambiguity, given the polysemy nature of non-vowel words., which results in problems with the automatic analysis, especially in morpho-syntactic tagging. "درس [drs]" can have different vowellations, such as "دَرَسَ [darasa] (he studied)", "دَرْسٌ [darsun] (a lesson)", "دَرَّسَ [dar sa] (he taught)", etc.

Syntactical dependencies. The TD is characterized by the fact that the possessive pronouns or adjectives can be reduced and agglutinated to the nouns in their final position. In (Utt17), the expression "your phone" is written in two forms "تليفون متاعك [tlyfwn mtAEk]" where the possessive pronoun "متاعنا [mtAEnA]" is detached from the noun and "تليفونك [tlyfwnk]" whose the possessive pronoun is reduced and agglutinated to the noun.

Utt17: هز تليفونك قتلك هز تليفون متاعك [hz tlyfwnk qtlk hz tlyfwn mtAEk] (take your phone I told you to take your phone)

Alternation of languages and dialects. TD is a spoken variety of Arabic and presents a mode of communication built on the alternation between several languages and dialects. Tunisians express themselves spontaneously using usually three languages namely, MSA, TD and French. The TD itself represents a mosaic of languages many of which words and expressions are borrowed from French, Maltese, English, Turkish, and Spanish as a result of trade movements and colonization over the centuries. These words and expressions can be used daily without any phonological or morphological modification.

Reuse of borrowed words. Some foreign words, especially those from the French language, are affected by morphological changes to express an action, an order, or the possession of objects. A borrowed verb is morphologically derived to produce adjectives, nouns, and the conjugation of that verb. Borrowed nouns also undergo morphological derivations including, the verb, the adjective, and the noun. These changes are different from the real derivation of the word real language. For example, the French verb "*gérer*" (to manage) is conjugated to "يجاري [yjAry] (he manages)" instead of "*il gère*" in French. This reuse and derivation feature is applied also to MSA words. For example, adding the affix "جي [jy]" to names like "بنك [bnk] (bank)" indicates the profession "بنكاجي [bnkAjy] (banker)".

Data

Our study is carried out using the Disfluencies Corpus from Tunisian Arabic Transcription ’DisCoTAT’ [5]. It consists of transcribed utterances coming mainly from recordings of railway information services and Tunisian TV channels and radio programs.

DisCoTAT is composed of two parts. The first part consists of 38.627 utterances collected from three TD existing corpora (i.e., STAC [33], TUDICOI [12], and TARIC [20]). STAC (Spoken Tunisian Arabic Corpus) is composed of 3 h and 28 min (7.788 utterances) of TD speech recordings collected from different TV channels and radio

Table 1 Simple disfluencies distribution in the DisCoTAT corpus

Syllabic elongations	Speech words	Word-fragments	Simple repetitions
185	300	1213	1342

Table 2 Complex disfluencies distribution in the DisCoTAT corpus

Complex repetitions	Insertions	Substitutions	Deletions
809	282	495	132

Table 3 Examples of utterances with disfluencies

STAC	بالحق إنسان ولى يلوج يعمل معناتها ديفولمون
	dyfwlmwn mEnAthA yEml ylwj wlY InsAn bAlHq
	In truth, a human had become seek to make I mean release
TUDICOI	من هنا من هنا وقتاش يخرج لتونس
	ltwns yxrj wqtA$ hnA mn hnA mn
	From here, from here, when goes out to Tunisia
TARIC	قداش هي من تونس للجم بروميار كلاس بقداش
	bqdA$ klAs brwmyAr lljm twns mn hy qdA$
	How much is she from Tunis to Eljam first-class how much
DisCoTAT	مالمستحسن يمشي لل اه للطبيب
transcripts	llTbyb Ah ll ym$y mAlmstHsn
	It is recommended he goes to a ah to a doctor

stations. TUDICOI (TUnisian DIalect COrpus Interlocutor) is composed of 1.825 dialogues composed of 12.182 utterances. TARIC (Tunisian Arabic Railway Interaction Corpus) consists of 20 h of transcribed speech. It is composed of 4.662 dialogues with 18.657 utterances. TUDICOI and TARIC utterances consist of railway information services (e.g. train schedule, train destination, tariffs, etc.). However, only 21% of collected utterances contain disfluencies phenomena.

The second part of DisCoTAT is about 2 h of recordings obtained from different TV channels and radio stations (i.e., Mosaique radio, Sfax radio, and Nessma TV). The transcription is done manually according to OTTA and CODA-TUN conventions [34]. Only disfluent utterances are transcribed to increase the number of disfluencies occurrences in the corpus. To date, the number of transcripts is about 406 disfluent utterances. The total number of disfluent utterances in DisCoTAT is about 3780 composed of 4757 disfluency phenomena. 80% of utterances are used for training and 20% of utterances are used for evaluation. Tables 1 and 2 illustrate the distribution of disfluencies types in DisCoTAT.

However,DisCoTAT is composed of a mosaic of words coming from various languages mainly TD (62%), MSA (17%), French (13%), and Others (8%). Table 3 presents an example of DisCoTAT utterance.

DisCoTAT is enriched with two types of annotation: morpho-syntactic annotation using TD-WordNet and handcrafted complex disfluencies annotation using the annotator tool DisAnT [5]. A given utterance goes through two phases of processing. The first phase is automatic, it consists of applying a set of pre-processing tasks such as lexical analysis, POS tagging, and simple disfluencies processing. The second phase is manual, it consists of identifying the disfluencies boundaries in the utterance.

The Proposed Method for Disfluencies Processing in TD

In pursuit of our previous work [6], we present in this section, a transcription-based method guided by linguistic features, to handle disfluencies removal from transcribed utterances of spoken TD.

We have taken up the method proposed in our previous work, in particular pre-processing and simple disfluencies processing steps. However, we propose, in this paper, a transition-based model to carry out the complex disfluencies processing step. Figure 2 shows the steps of the proposed method through a TD utterance example.

Pre-processing Step

The pre-processing step is essential for the detection of different types of disfluencies, it allows the utterance to be adapted for downstream steps.

Tokenization

The purpose of the tokenization consists of segmenting the utterance into tokens. A token is either a word or a group of words (i.e., compound words). For example, "موش [mw$] (is not)" and "نورمال [nwrmAl] (normal)" constitute one token "موش-نورمال (abnormal)" labelled with ("locution").

Morph-Syntactic Analysis

Tokens found are labelled with POS tags using the TD-WordNet lexicon [5]. The morpho-syntactic analysis allows lemmatizing no-labelled words based on the TD-WordNet list of prefixes and suffixes to find their POS tag. For example, the word " قلتلك [qltlk] (I told you)" is an inflected form of the verb "قال [qAl] (tell)", concatenated with the suffix "لك [lk]" which refers to a singular pronoun.

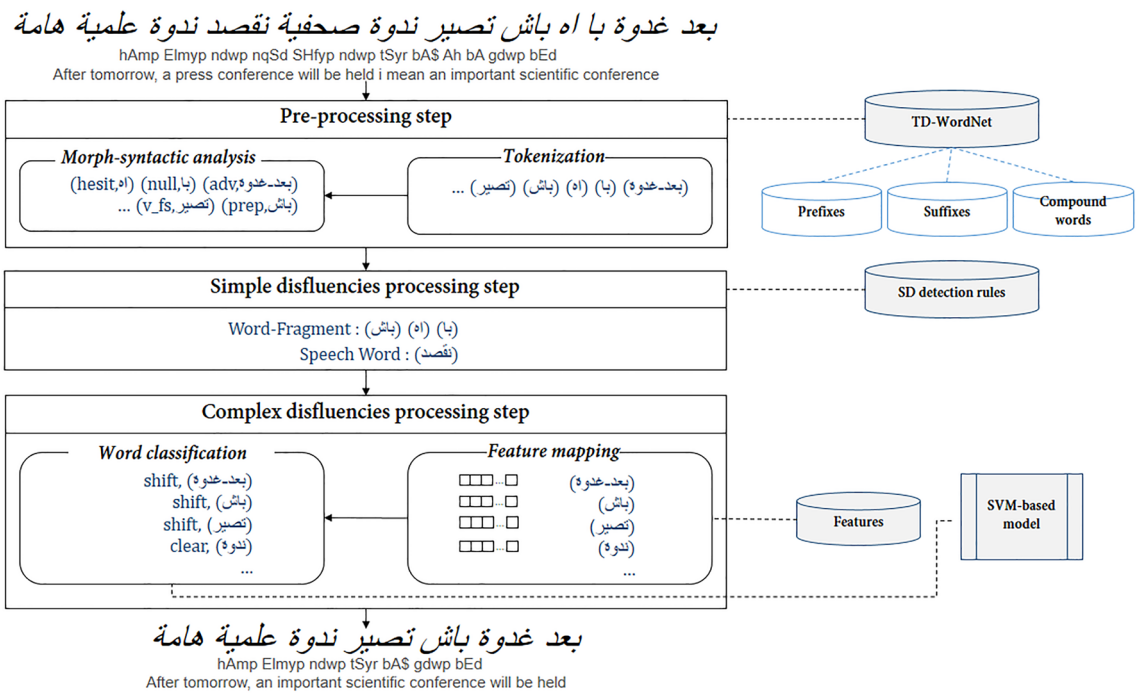

Fig. 2 Steps of the proposed method

Simple Disfluencies Processing Step

Simple disfluencies are processed using a rule-based approach. For detecting simple disfluencies, we designed a set of detection rules based mainly on POS tags of words. In this work, we have integrated the semantic feature 'word synonyms' to improve the detection performance of simple repetition phenomena.

Syllabic Elongations

Syllabic elongations processing consists of detecting and correcting words that are *(i)* not POS-tagged and *(ii)* contain more than two extensions in the first or last syllable.

Speech Words

Detecting speech words is a word-based matching of words tagged ("Marq_Disc", discourse mark) and ("Marq_Hesit", hesitation mark). Although speech words fall into the simple disfluencies category, they are removed until the next step, as they help to detect complex disfluencies.

Word-Fragments

Word-fragments are lexical syllable repetitions. Their processing consists of detecting and removing words that *(i)* may not be POS-tagged and *(ii)* are an integral part of the following word.

Simple Repetitions

Simple repetition processing consists of detecting and removing words that occur several times consecutively. Indeed, repetition can be lexical (i.e., with the same word) or semantic (i.e., with a synonym).

Speech words that appear inside word-fragment cases (e.g., Utt18) or simple repetition cases (e.g., Utt19) are removed consistently with the appropriate disfluency.

Utt18: "ترانمتا اه متاع تونس [trAn mtA Ah mtAE twns] (The train of euh of-Tunis)".

Utt 19: "المعرض يبدا اليوم آآ ليوما [AlmErD ybdA Alywm euuh lywmA] (The show starts today ahh today)".

Complex Disfluencies Processing Step

Previous studies on the disfluencies processing task fall into four main approaches:

The **Noisy Channel Models (NCM) based** approach [2, 17] uses NCM with Tree Adjoining Grammars (TAG) [14] to find the similarity between the disfluent chunk of the utterance and its correction as an indicator of disfluencies.

NMC models are not suitable for detecting various types of disfluencies, notably insertions and deletions in which repetitions do not necessarily accompany them. In

[35], 62% of the words of the Reparandum are identical to the words of the Repair. In our study corpus, in approximately 32% of DC cases, the Repair does not contain repeated words from the Reparandum.

The **transition-based** approach [16, 30] uses transition-based analysis models that detect disfluencies while simultaneously identifying the utterance's syntactic tree. Disfluencies detection is achieved by adding new actions to the parser to detect and remove the disfluent parts of the utterance and their dependencies.

The advantage of transition models lies in the fact that two different tasks (i.e., syntactic analysis and disfluency processing) can be carried out simultaneously. Likewise, they can capture the contiguous syntactic dependencies of disfluencies as well as segment-level information. In contrast, joint models require large annotated treebanks, containing both disfluent and syntactic structure annotations for the training phase. Additionally, they introduce an additional annotated syntactic structure that is very expensive to produce and can cause noise by significantly enlarging the output search space.

The **sequence-tagging based** approach [1, 18] uses word sequence-tagging models. It is based on statistical models that predict the class of a token according to the BIO encoding schema [26]. A model labels words as being inside or outside of the edit region.

Sequence-tagging-based models make it possible to capture distant dependencies between Reparandum and Repair even in long utterances using neural networks. The disadvantage of these models lies in the fact that they require large volumes of annotated data.

The **seq2seq transformer-based** approach [27, 28] is inspired by the machine translation task which considers disfluent text as the source language and the fluent text as the target language. The Transformer is a seq2seq neural type which has the particularity of using only the attention mechanism and no recurrent or conventional network.

The seq2seq models using the attention mechanism make it possible to almost perfectly detect the Repair which is far from the Reparandum. In contrast, seq2seq models must rely on a large amount of data.

This work is part of the disfluencies processing project in the Tunisian dialect. We propose to create and compare several disfluency processing models from the existing state-of-the-art methods. In a previous work, we presented our classification-based model. In this present work, we propose to handle complex disfluencies using a transition-based model, which detects and removes the disfluent chunk of the utterance with a set of transition actions and without syntactic dependencies inspired by [30]. We also incorporate semantic features based mainly on word synonyms to perform complex repetition detection in addition to morpho-syntactic features. The main idea of this work is to detect only the disfluent words (i.e., reparandum and interregnum) with ignoring the rest of the utterance's words (i.e, fluent words, and repair). Identifying reparandums is the most challenging task in disfluencies processing. They may be in subjective form, occur at different places, vary in length, and in some cases, they could be nested [30]. We display both model comparisons in section "Complex Disfluencies Processing Evaluation". In future work, we aim to implement the other disfluency detection methods.

Model presentation

A disfluent utterance is presented by the tuple (A,I,D,O) where:

- Action (A) presents the history of the actions,
- Input (I) presents words not yet processed,
- Disfluent (D) presents words considered to be disfluent,
- Output (O) presents words considered to be fluent.

The model results a sequence of binary tags denoted as $D^n_{wi} = d_{w1}, d_{w2}, .., d_{wn}$, which means that w_i is either fluent or disfluent. The best sequence of tags D^* given W^n_i [30] is:

$$D^* = argmax_D\left(D^n_1 | W^n_1\right)$$

At the instant t_0, I contains W^n_i, and O, D and A are initially empty. For each a w_i, the module predicts the transition action $a_i = P(A^w_i | w^n_i, a_(i-1))$ where $A^w_i = \{clear;shift\}$ and $a_(i-1)$ is a dynamic feature:

- clear: moves w_i from I to D, and
- shift: moves w_i from I to O.

The model stops when I is empty. Since the model aims to predict a transition action for W^n_1, we implemented a Binary Classifier Transitions (BCT) proposed by [31] based on word sequence W^n_1 and feature vectors $V_{w^n_i} = v^m_1.w_1, v^m_1.w_2, .., v^m_1.w_n$ presented in Section "Features of the Binary Classifier". Algorithm 1 summarizes the BCT's main steps.

Table 4 Features of the binary classifier

Word-based features
POS tag of the current word
The current word is subordinating conjunction
The current word is coordinating conjunction
The current word is an interjection
Utterance-based features
Repetition number of the current word in the utterance
POS of the n preceding word with n = (1,2,3)
POS of the n next word with n = (1,2,3)
The current word starts with an uncompleted word in a window of 3 preceding words
The current word is repeated in a window of 3 preceding words
The current word is repeated in a window of 3 next words
The current word is preceded by a subordinating or coordinating conjunction
The current word is followed by a subordinating or coordinating conjunction
Presence of a possession adjective or pronoun related and concatenated to a similar word in the utterance
The current word has a synonym in a window of 3 preceding words
The current word has a synonym in a window of 3 next words
Dynamic feature
Class of the n preceding words with n = (1,2,3)

Algorithm 1 Transition Model Algorithm

Require: Set of $W_1^n = w_1, w_2, .., w_n$
Ensure: Set of $D_1^n = d_{w1}, d_{w2}, .., d_{wn}$
O,D,A are empty, $I \leftarrow W_1^n$
while $I \neq empty$ **do**
  $a_i \leftarrow P(A_i^w | w_i^n, a_{(i-1)})$
  if a_i is $clear$ **then**
    add w_i to D
  else
    add w_i to O
  end if
  remove w_i from I
  add a_i to A
end while

Table 5 Classification algorithm results

	LibSVM	SMO	PART	J48	BayesNet
F-Mesure rate	79.81%	79.27%	76.14%	75.88%	72.63%

Features of the Binary Classifier

Labelling a word entity is based on a set of observations that are introduced to a classifier as a set of feature vectors. The task of detecting disfluencies is mainly related to either prosodic or linguistic features. Prosodic information (such as duration, rhythm, etc.) are omitted in our work since it belongs to the processing of transcripts. However, we rely on only linguistic features presented in Table 4.

We used contextual features with a window of ± 3 words. We experimented with a window ± 1 word, a window ± 2 word, a window ± 3 word and a window ± 4 word. The choice of the word window is justified by the fact that the TD utterances are not too long, the corpus analysis shows that the repair starts after a window that does not exceed three words after the disfluent part, and this does not take into account the interregnum. Finally, we used the dynamic criterion. It considers the class assigned dynamically to the three previous words.

Table 6 Evaluation of simple disfluency types

Disfluencies type	Recall (%)	Precision (%)	F-Measure (%)
Syllabic elongation	99.30	97.98	98.6
Speech words	100	100	100
Word-fragments	100	100	100
Simple repetitions	96.60	93.85	95.21

Model Generation

The binary classifier is built using the ML algorithm SVM, which gives high performance in binary classification tasks [23]. SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. We have experimented with various ML classification algorithms using the open-source library WEKA,[2] and we found that libSVM [8], an implementation of SVM, achieves higher results (79.81%). Table 5 summarizes the experiment results of five classification algorithms.

In the model parameters, the batch size is fixed to 100 instances. We used the kernel function Sigmoid [15]:

$$K(X, Y) = tanh\left(\gamma X^T Y + r\right)$$

The Sigmoid kernel takes two parameters, γ (i.e., the scaling parameter of the input data) and r (i.e., a shifting parameter that controls the threshold of mapping). γ and r are fixed to 0. The loss function used is 0.1. It defines the error between the predicted target and the given target value.

We tested the model using the k-fold cross-validation option with k = 10. Data are spliced into 10 parts (i.e., 10 folds) for which the algorithm runs 10 times.

Experiments and Discussion

In this section, we report the evaluations we performed using the evaluation portion data of DisCoTAT. We have implemented the method using the Java programming language with the NetBeans environment. We used Recall, Precision and F-Measure metrics.

The overall method gives good results. The recall, precision and F-measure metrics are 95.39%, 82.24% and 88.33%, respectively. Also, we evaluated the main steps of the method. The next sections present the evaluation results and analysis of simple (Section "Simple Disfluencies Processing Evaluation") and complex (Section "Complex Disfluencies Processing Evaluation") disfluencies processing steps.

Simple Disfluencies Processing Evaluation

The simple disfluencies processing step achieved promising results. Rates are projected in Table 6.

The performance of how well the simple disfluencies processing module can detect speech words and word-fragments is mainly due to the efficiency of the pre-processing step. The TD-WordNet lexicon covers all instances of hesitation and discourse marks identified for the TD vocabulary. The effect of the lemmatization task that deals with the recognition of no-labelled words using their prefix and suffixes, improved the detection of simple disfluency types. In addition, we tested the contribution of the semantic feature to the process of detecting simple repetitions. We have thus obtained a 8.5% improvement compared to the previous work. In Utt20, both "الترينو [Altrynw]" and "ترانٍ [trAn]" are synonyms and mean (train). A simple repetition disfluencies case is successfully detected.

Utt20: ترانٍ ترينو يخرج الاربعة [trAn trynw yxrj AlArbEp] (Train train leaves at four).

Complex Disfluencies Processing Evaluation

We have evaluated the efficiency of the complex disfluencies processing step by experimenting with two models. In this present paper, we presented our transition-based model, which detects and removes the disfluent chunk of the utterance with a set of transition actions without syntactic dependencies. It achieved an F-Measure score of 79.81%.

In our previous work [6], a sequence-tagging-based model for complex disfluencies processing is proposed. The model classifies the utterance's words into six classes based on the reparandum-interregnum-repair structure and following the BIO encoding [26]. Tokens can be labelled as B_RM (i.e., the beginning of the reparandum), I_RM (i.e., belongs to the reparandum part), B_RP (i.e., the beginning

Table 7 Evaluation of complex disfluency types

Disfluencies type	Sequence-tagging based			Transition-based		
	Recall	Precision	F-Measure	Recall	Precision	F-Measure
Complex repetitions	91.63%	86.7%	89.10%	89.87%	83.77%	86.71%
Insertions	78.98%	73.48%	76.13%	82.59%	78.36%	80.42%
Substitutions	82.67%	78.90%	80.74%	82.82%	79.1%	80.92%
Deletions	71.57%	68.32%	69.91%	75.57%	67.32%	71.21%

[2] https://www.cs.waikato.ac.nz/ml/weka/.

of the repair part), I_RP (i.e., belongs to the repair part), IP (i.e., Interregnum (i.e., belongs to speech words), or O: (i.e., fluent word). The sequence-tagging-based model is handled with the statistical CRF algorithm [1] and uses the same features used for the transition-based model. It achieved an F-Measure score of 78.97% with considering semantic features.

Through rates projected in Table 7, we notice that the performances of the two models are close. The transition-based model gives a slight improvement in the detection performance of insertions, modifications, and deletions. We notice that the results for detecting complex repetitions have slightly degraded using the transition-based model. In fact, in some cases where it is a complex repetition of the discontinuous type (i.e., other entity inserted between the two repeated words), as shown in Utt21, the transition-based model considers this disfluency to be of the deletion type and deletes the word inserted with the reparandum. In this case, we admit that the detection of the repair as well as the reparandum is mandatory. However, this error does not affect the syntactic harmony in the utterance.

Utt21: الخدام أنا الخدام إلي يتضر [AlxdAm OnA AlxdAm Ily ytDr] (The worker, I am the worker who is harmed).

Considering the challenges of disfluencies processing in TD presented in section 2.2, the models have overcome several difficulties:

- The model can capture long-range dependency of disfluencies due to the context of a ± 3 window of neighbouring words, applied to several features,
- The wealth of the TD-WordNet led to overcoming compound words, voluntary repetitions, synonyms and pronunciation variability,
- The lemmatization task in the pre-processing step therefore makes it possible to overcome the difficulty of agglutination linked to the Arabic language and TD specifically. Also, the simple disfluencies processing step facilitates the task of complex disfluencies detection since it consists of eliminating these phenomena which disturb the syntactic and semantic flow of the utterance,
- The proposed models can detect several complex disfluency structures even with the presence of OOV words,
- The efficiency of the selected features led to treating syntactical dependencies that appear within complex disfluencies, notably those dealing with possession adjectives or pronouns, as we explained in section 2.2,
- The transition-based model can detect reformulations, a disfluencies type that we did not address in our contribution, with an F-measure rate of 42.7%. Utt22 demonstrates a sentence reformulation case. The speaker breaks his speech and starts a new utterance. Utt22: مع وقتاش يخرج... ثمة تران توا لسوسة [mE wqtA$ yxrj... vmp trAn twA lswsp] (When leaves... Is there a train now to Sousse.).

Also, to validate the learning features set, we trained the models without considering the grammatical gender and number of POS tags. Consequently, this syntax information contributes to improving the performance of both complex disfluencies detection models by 8.85% (Sequence-tagging based) and 8% (transition-based).

However, the major error analysis cases are mainly due to the following reasons:

- POS tagging without considering Arabic vowels can generate multiple POS tags for a given word, for example, "المقابلة [AlmqAblp]" means both (game, Noun) and (across, Adverb) in TD-WordNet,
- In the case where coordinating conjunction is omitted, the enumeration can be interpreted as a modification disfluency type insofar as the speaker replaces the words to catch up and correct his enunciation,
- Foreign words undergo uncontrollable lexical changes that we cannot pin down their various phonological and morphological derivatives in a lexical base,
- The absence of a discursive context in the utterance or dialogue contributes to various problems, in particular annotation confusion. Some cases of disfluencies, nested for example, have a very complex structure, therefore, they can undergo different annotations depending on the annotator.

Conclusion and Perspectives

In this paper, we investigated the field of disfluencies processing in the spontaneous Tunisian spoken dialect. First, we presented our study corpus DisCoTAT which consists of transcribed utterances coming mainly from recordings of Tunisian TV channels and radio programs. Then, we proposed our method to detect eight types of disfluencies. We constructed a set of detection rules based mainly on POS tags of words for simple disfluencies. We also proposed a transition-based model for the complex disfluencies detection task, which was evaluated and compared to another model based on sequence-tagging. The comparison results proved that both transition-based and sequence-tagging models are efficient, with a slight improvement for the transition-based model.

The originality of our contribution is due to that eight types of disfluencies are detected and corrected automatically, where complex disfluencies are detected using stochastic models derived from ML techniques with various

linguistic features and that transcriptions are issued from a wide domain in the spontaneous spoken TD.

In future work, we intend to implement and test other disfluencies detection models like seq2seq transformers. Also, we aim to add acoustic features (e.g., duration, intensity, pitch, etc.) to the linguistic features.

Author Contributions All authors contributed to the study. The first draft of the manuscript was written by Emna Boughariou and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Not applicable.

Data Availability Statement Data available on request from the authors.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

Research Involving Human and /or Animals: Not applicable.

Informed Consent Not applicable.

References

1. Alharbi S, Hasan M, Simons AJ, Brumfitt S, Green P. Sequence labeling to detect stuttering events in read speech. Comput Speech Lang. 2020;62: 101052.
2. Bach N, Huang F. Noisy bilstm-based models for disfluency detection. In: Proc Interspeech. 2019;2019:4230–4.
3. Bahou Y, Maaloul M, Boughariou E. Towards the supervised machine learning and the conceptual segmentation technique in the spontaneous Arabic speech understanding. In: Procedia Computer Science, ACLING2017. UAE: Dubai; 2017. p. 225–32.
4. Bouamor H, Hassan S, Habash N. The madar shared task on Arabic fine-grained dialect identification. In: Proceedings of the Fourth Arabic Natural Language Processing Workshop, 2019; p. 199–207.
5. Boughariou E, Bahou Y, Hadrich Belguith L. Linguistic resources construction: Towards disfluency processing in spontaneous Tunisian dialect speech. In: International Conference on text, speech, and dialogue, 2019; pages 316–328. Springer.
6. Boughariou E, Bahou Y, Hadrich Belguith L. Classification based method for disfluencies detection in spontaneous spoken Tunisian dialect. In: Proceedings of SAI Intelligent Systems Conference, 2020; p. 182–195. Springer.
7. Bouraoui J-L, Vigouroux N. Traitement automatique de disfluences dans un corpus linguistiquement contraint. In: Actes de TALN, 2009; p. 117.
8. Chang C-C, Lin C-J. Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol (TIST). 2011;2(3):1–27.
9. Constant M, Tellier I. Evaluating the impact of external lexical resources into a crf-based multiword segmenter and part-of-speech tagger. In: 8th International Conference on Language Resources and Evaluation (LREC'12), 2012; p. 646–650.
10. Dong Q, Wang F, Yang Z, Chen W, Xu S, Xu B. Adapting translation models for transcript disfluency detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019;33:6351–8.
11. Germesin S, Becker T, Poller P. Hybrid multi-step disfluency detection. In: International Workshop on Machine Learning for Multimodal Interaction, 2008; p. 185–195. Springer.
12. Graja M, Jaoua M, Hadrich Belguith L. Statistical framework with knowledge base integration for robust speech understanding of the Tunisian dialect. IEEE/ACM Trans Audio Speech Lang Process (TASLP). 2015;23(12):2311–21.
13. Ismail SB, Boukédi S, Haddar K. Hpsg grammar supporting Arabic preference nouns and its tdl specification. In: International Conference on Arabic language processing, 2019; p. 221–234. Springer.
14. Johnson M, Charniak E. A TAG-based noisy-channel model of speech repairs. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), 2004; p. 33–39, Barcelona, Spain.
15. Lin H-T, Lin C-J. A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. Neural Comput. 2003;3(1–32):16.
16. Lou PJ, Anderson, P, Johnson M. Disfluency detection using auto-correlational neural networks. 2018. arXiv preprint arXiv:1808.09092.
17. Lou PJ, Johnson M. Disfluency detection using a noisy channel model and a deep neural language model. 2018. arXiv preprint arXiv:1808.09091.
18. Lu Y, Gales M, Knill K, Manakul P, Wang Y. Disfluency detection for spoken learner English. In Proc. SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education, 2019; p. 74–78.
19. Masmoudi A, Bougares F, Khmekhem ME, Estève Y, Hadrich Belguith L. Automatic speech recognition system for Tunisian dialect. Lang Resour Eval. 2018;52(1):249–67.
20. Masmoudi A, Khmekhem ME, Esteve Y, Hadrich Belguith L, Habash N. A corpus and phonetic dictionary for Tunisian Arabic speech recognition. In LREC, 2014; p. 306–310.
21. Masmoudi A, Khmekhem ME, Khrouf M, Hadrich Belguith L. Transliteration of Arabizi into Arabic script for Tunisian dialect. ACM Trans Asian Low-Resour Lang Inf Process. 2019;19(2):1–21.
22. Masmoudi A, Laatar R, Ellouze M, Hadrich Belguith L. Semantic language model for Tunisian dialect. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), 2019; p. 720–729, Varna, Bulgaria. INCOMA Ltd.
23. Mathur A, Foody G. Multiclass and binary svm classification: implications for training and classification users. Geosci Remote Sens Lett IEEE. 2008;5:241–5.
24. Messaoudi A, Haddad H, HajHmida MB, Fourati C, Hamida AB. Learning word representations for Tunisian sentiment analysis. 2020. arXiv preprint arXiv:2010.06857.
25. Neifar W, Bahou Y, Graja M, Jaoua M. Implementation of a symbolic method for the tunisian dialect understanding. In: Proceedings of 5th International Conference on Arabic Language Processing, Oujda, Maroc; 2014.
26. Ramshaw LA, Marcus MP. Text chunking using transformation-based learning. In: Armstrong S, Church K, Isabelle P, Manzi S, Tzoukermann E, Yarowsky D, editors. Natural language processing using very large corpora. Springer; 1999. p. 157–76.
27. Rocholl JC, Zayats V, Walker DD, Murad NB, Schneider A, Liebling DJ. Disfluency detection with unlabeled data and small bert models. 2011. arXiv preprint arXiv:2104.10769.

28. Rohanian M, Hough J. Best of both worlds: Making high accuracy non-incremental transformer-based disfluency detection incremental. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021; p. 3693–3703.
29. Shriberg EE. Preliminaries to a theory of speech disfluencies. Thèse de doctorat: University of California, Berkeley; 1994.
30. Wang S, Che W, Zhang Y, Zhang M, Liu T. Transition-based disfluency detection using lstms. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017; p. 2785–2794.
31. Wu S, Zhang D, Zhou M, Zhao T. Efficient disfluency detection with transition-based parsing. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on natural language processing (Volume 1: Long Papers), 2015; p. 495–503.
32. Zayats V, Ostendorf M, Hajishirzi H. Disfluency detection using a bidirectional lstm. 2016. arXiv preprint arXiv:1604.03209.
33. Zribi I, Ellouze M, Hadrich Belguith L, Blache P. Spoken Tunisian Arabic corpus "stac'': transcription and annotation. Res Comput Sci. 2015;90:123–35.
34. Zribi I, Graja M, Khmekhem ME, Jaoua M, Hadrich Belguith L. Orthographic transcription for spoken Tunisian Arabic. In: International Conference on intelligent text processing and computational linguistics, 2013; p. 153–163. Springer.
35. Zwarts S, Johnson M. The impact of language models and loss functions on repair disfluency detection. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011; p. 703–711, Portland, Oregon, USA. Association for Computational Linguistics.