Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6279–6286
Marseille, 11–16 May 2020

TArC: Incrementally and Semi-Automatically Collecting a Tunisian Arabish Corpus

Elisa Gugliotta[1,2], **Marco Dinarelli**[1]
1. LIG, Bâtiment IMAG - 700 avenue Centrale - Domaine Universitaire de Saint-Martin-d'Hères, France
2. University of Rome *Sapienza*, Piazzale Aldo Moro 5, 00185 Roma, Italy
elisa.gugliotta@uniroma1.it, marco.dinarelli@univ-grenoble-alpes.fr

Abstract

This article describes the constitution process of the first morpho-syntactically annotated Tunisian *Arabish* Corpus (TArC). Arabish, also known as *Arabizi*, is a spontaneous coding of Arabic dialects in Latin characters and *arithmographs* (numbers used as letters). This *code-system* was developed by Arabic-speaking users of social media in order to facilitate the writing in the Computer-Mediated Communication (CMC) and text messaging informal frameworks. There is variety in the realization of Arabish amongst dialects, and each Arabish code-system is under-resourced, in the same way as most of the Arabic dialects. In the last few years, the focus on Arabic dialects in the NLP field has considerably increased. Taking this into consideration, TArC will be a useful support for different types of analyses, computational and linguistic, as well as for NLP tools training. In this article we will describe preliminary work on the TArC semi-automatic construction process and some of the first analyses we developed on TArC. In addition, in order to provide a complete overview of the challenges faced during the building process, we will present the main Tunisian dialect characteristics and their encoding in Tunisian Arabish.
Keywords: Tunisian Arabish Corpus, Arabic Dialect, Arabizi

1 Introduction

Arabish is the romanization of Arabic Dialects (ADs) used for informal messaging, especially in social networks.[1] This writing system provides an interesting ground for linguistic research, computational as well as sociolinguistic, mainly due to the fact that it is a spontaneous representation of the ADs, and because it is a linguistic phenomenon in constant expansion on the web. Despite such potential, little research has been dedicated to Tunisian Arabish (TA). In this paper we describe the work we carried to develop a flexible and multi-purpose TA resource. This will include a TA corpus, together with some tools that could be useful for analyzing the corpus and for its extension with new data.

First of all, the resource will be useful to give an overview of the TA. At the same time, it will be a reliable representation of the Tunisian dialect (TUN) evolution over the last ten years: the collected texts date from 2009 to present. This selection was done with the purpose to observe to what extent the TA orthographic system has evolved toward a writing convention. Therefore, the TArC will be suitable for phonological, morphological, syntactic and semantic studies, both in the linguistic and the Natural Language Processing (NLP) domains. For these reasons, we decided to build a corpus which could highlight the structural characteristics of TA through different annotation levels, including Part of Speech (POS) tags and lemmatization. In particular, to facilitate the match with the already existing tools and studies for the Arabic language processing, we provide a transcription in Arabic characters at token level, following the Conventional Orthography for Dialectal Arabic guidelines *CODA** (CODA *star*) (Habash et al., 2018) and taking into account the specific guidelines for TUN (CODA TUN) (Zribi et al., 2014). Furthermore, even if the translation is not the main goal of this research, we have decided to provide an Italian translation of the TArC's texts.[2]

Even though in the last few years ADs have received an increasing attention by the NLP community, many aspects have not been studied yet and one of these is the Arabish code-system. The first reason for this lack of research is the relatively recent widespread of its use: before the advent of the social media, Arabish usage was basically confined to text messaging. However, the landscape has changed considerably, and particularly thanks to the massive registration of users on Facebook since 2008. At that time, in Tunisia there were still no Arabic keyboards, neither for Personal Computers, nor for phones, so Arabic-speaking users designed TA for writing in social media (Table 1). A second issue that has held back the study of Arabish is its lack of a standard orthography, and the informal context of use. It is important to note that also the ADs lack a standard code-system, mainly because of their oral nature. In recent years the scientific community has been active in producing various sets of guidelines for dialectal

[1] Also known as Arabizi (from the combination between the Arabic words 'Arab', ['ʕarab] and 'English' [iːnˈɡliːziː], or the English one: 'easy', ['iːzi]), Franco-Arabic, Arabic Chat Alphabet, ACII-ized Arabic, and many others. 'Arabish' is probably the result of the union between ['ʕarab] and 'English'.

[2] We considered the Italian translation as an integrated part of the annotation phase that would have cost us less effort in addition to us if carried out in our mother-tongue. The possibility of a TArC English translation is left open for a later time.

Arabic writing in Arabic characters: CODA (Conventional Orthography for Dialectal Arabic) (Habash et al., 2012).
The remainder of the paper is organized as follows: section 2 is an overview of NLP studies on TUN and TA; section 3 describes TUN and TA; section 4 presents the TArC corpus building process; section 5 explains preliminary experiments with a semi-automatic transcription and annotation procedure, adopted for a faster and simpler construction of the TArC corpus; conclusions are drawn in section 6.

2 Related Work

In this section, we provide an overview of work done on automatic processing of TUN and TA. As briefly outlined above, many studies on TUN and TA aim at solving the lack of standard orthography. The first Conventional Orthography for Dialectal Arabic (CODA) was for Egyptian Arabic (Habash et al., 2012) and it was used by Bies et al. (2014) for Egyptian Arabish transliteration into Arabic script. The CODA version for TUN (CODA TUN) was developed by Zribi et al. (2014), and was used in many studies, like Boujelbane (2015). Such work presents a research on automatic word recognition in TUN. Narrowing down to the specific field of TA, CODA TUN was used in Masmoudi et al. (2015) to realize a TA-Arabic script conversion tool, implemented with a rule-based approach. The most extensive CODA is CODA*, a unified set of guidelines for 28 Arab city dialects (Habash et al., 2018). For the present research, CODA* is considered the most convenient guideline to follow due to its extensive applicability, which will support comparative studies of corpora in different ADs. As we already mentioned, there are few NLP tools available for Arabish processing in comparison to the amount of NLP tools realized for Arabic. Considering the lack of spelling conventions for Arabish, previous effort has focused on automatic transliteration from Arabish to Arabic script, e.g. Chalabi and Gerges (2012), Darwish (2013), and Al-Badrashiny et al. (2014). These three work are based on a character-to-character mapping model that aims at generating a range of alternative words that must then be selected through a linguistic model. A different method is presented in Younes et al. (2018b), in which the authors present a sequence-to-sequence-based approach for TA-Arabic characters transliteration in both directions (Sutskever et al., 2014; Younes et al., 2018b).
Regardless of the great number of work done on TUN automatic processing, there are not a lot of TUN corpora available for free (Younes et al., 2018a). To the best of our knowledge there are only five TUN corpora freely downloadable: one of these is the PADIC (Mourad, Abbas, 2017), composed of 6,400 sentences in six Arabic dialects, translated in *Modern Standard Arabic* (MSA), and annotated at sentence level.[3] Two other corpora are the Tunisian Dialect Corpus Interlocutor (TuDiCoI) (Graja, 2010) and the Spoken Tunisian Arabic Corpus (STAC) (Zribi, 2015), which are both morpho-syntactically annotated. The first one is a spoken task-oriented dialogue corpus, which gathers a set of conversations between staff and clients recorded in a railway station. TuDiCoI consists of 21,682 words in client turns (Graja et al., 2013).[4] The STAC is composed of 42,388 words collected from audio files downloaded from the web (as TV channels and radio stations files) (Zribi et al., 2015). A different corpus is the TARIC (Masmoudi, 2014), which contains 20 hours of TUN speech, transcribed in Arabic characters (Masmoudi et al., 2014).[5] The last one is the TSAC (Medhaffar et al., 2017), containing 17k comments from Facebook, manually annotated to positive and negative polarities (Medhaffar et al., 2017). This corpus is the only one that contains TA texts as well as texts in Arabic characters. As far as we know there are no available corpora of TA transcribed in Arabic characters which are also morpho-syntactically annotated. In order to provide an answer to the lack of resources for TA, we decided to create TArC, a corpus entirely dedicated to the TA writing system, transcribed in CODA TUN and provided with a lemmatization level and POS tag annotation.

3 Characteristics of Tunisian Arabic and Tunisian Arabish

The Tunisian dialect (TUN) is the spoken language of Tunisian everyday life, commonly referred to as الدَّارِجَة, *ad-dārija*, العَامِّيَّة, *al-ʿāmmiyya*, or التُّونْسِي, *ət-tūnsī*. According to the traditional diatopic classification, TUN belongs to the area of Maghrebi Arabic, of which the other main varieties are Libyan, Algerian, Moroccan and the Ḥassānīya variety of Mauritania[6] (Durand, 2009). Arabish is the transposition of ADs, which are mainly spoken systems, into written form, thus turning into a quasi-oral system (this topic will be discussed in section 3.2). In addition, Arabish is not realized through Arabic script and consequently it is not subject to the Standard Arabic orthographic rules. As a result, it is possible to consider TA as a faithful written representation of the spoken TUN (Akbar, 2019).

3.1 Tunisian Arabic

The following list provides an excerpt of the principal features of TUN, which, through the TArC, would be researched in depth among many others.[7]
At the phonetic level, some of the main characteristics of TUN, and Maghrebi Arabic in general, are the

[3]The Arabic dialects of the PADIC are: TUN (Sfax), two dialects of Algeria, Syrian, Palestinian and Moroccan (Meftouh et al., 2018).

[4]The annotation was carried out only for 7,814 word.

[5]The 20 hours recorded are equivalent to 71,684 words.

[6]The main geographical macro-areas, also called geolects, are the area of the Levant or Syro-Palestinian, Egypt and Sudan, Mesopotamia, Maghreb (North Africa) and the Arabian Peninsula.

[7]For a detailed description please refer to (Durand, 2009), (Marcais, 1977).

following:

* Strong influence of the Berber substratum, to which it is possible to attribute the conservative phonology of TUN consonants.
* Presence of new emphatic phonemes, above all [ɾˤ], [lˤ], [bˤ].
* Realization of the voiced post-alveolar affricate [ʤ] as fricative [ʒ].
* Overlapping of the pharyngealized voiced alveolar stop [dˤ], <ض>, with the fricative [ðˤ], <ظ>.
* Preservation of a full glottal stop [ʔ] mainly in cases of loans from Classical Arabic (CA) or exclamations and interjections of frequent use.
* Loss of short vowels in open syllables.
* Monophthongization.[8] In TUN <بَيت>, ['baijt], 'house', becomes ['biːt] meaning 'room'.
* Palatalization of ā: Imāla, <إمالة>, literally 'inclination'. (In TUN the phenomenon is of medium intensity.) Thereby the word <باب>, ['baːb], 'door', becomes ['bɛːb].
* Metathesis[9] that in TUN results in: '(he) has understood': <فْهِم>, ['fhəm], '(she) has understood': <فِهْمِت>, ['fəhmət] or 'leg': <رْجِل>, ['rʒəl], 'my leg': <رِجْلِي>, ['rəʒli].

Regarding the morpho-syntactic level, the TUN presents:

* Addition of the prefix /-n/ to first person verbal morphology in *muḍāri'* (imperfective).
* Realization of passive-reflexive verbs through the morpheme /-t/[10] prefixed to the verb as in the example: <سوريّة مالحَفْصيّة تْتِلْبِس>, [suːˈriːjːa məl-ħafˈsˤijːa t-ˈtəlbəs], 'the shirts of Ḥafṣiya[11] are not bad', (lit: 'they dress').
* Loss of gender distinction at the 2nd and 3rd persons, at verbal and pronominal level.
* Disappearance of the dual form from verbal and pronominal inflexion. There is a residual of pseudo-dual in some words fixed in time in their dual form.
* Loss of relative pronouns flexion and replacement with the invariable form <اِلّي>, [əlːi].
* Use of presentatives /ɾā-/ and /hā-/ with the meaning of 'here', 'look', as in the example in TUN: <راني مَخْنوق>, ['ɾaːniː məxˈnuːq], 'here I am asphyxiated (by problems)', or in <هاك دَبَّرْتْها>, ['haː-k dəˈbːərt-haː], 'here you are, finding it (the solution)' hence: 'you were lucky'.
* Presence of circumfix negation marks, such as <<ما>, [ma] + verb + <ش> , [ʃ]>. The last element of this structure must be omitted if there is another negation, such as the Tunisian adverb <عُمْر>, ['ʕomr], 'never', as in the structure: <['ʕomr] + personal pronoun suffix + [mə] + perfect verb>. This construction is used to express the concept of 'never having done' the action in question, as in the example: <عُمري ما كُنْت نِتْصَوُّر...>, ['ʕomr-i ma 'kənt nətsˤawːər], 'I never imagined that...'.
Instead, to deny an action pointing out that it will never repeat itself again, a structure widely used is <[ma] + ['ʕaːd] + [ʃ] + imperfective verb>, where the element within the circumfix marks is a grammaticalized element of verbal origin from CA: <عاد>, ['ʕaːd], meaning 'to go back, to reoccur', which gives the structure a sense of denied repetitiveness, as in the sentence: <هو ما عادِش يرْجَع>, ['huːwa ma 'ʕaːd-ʃ 'jərʒaʕ], 'he will not come back'.
Finally, to deny the nominal phrase, in TUN both the <موش>, ['muːʃ], and the circumfix marks are frequently used. For the negative form of the verb 'to be' in the present, circumfix marks can be combined with the personal suffix pronoun, placed between the marks, as in <مانِيش>, [maˈniːʃ], 'I am not'.
Within the negation marks we can also find other types of nominal structures, such as: <[fiː] + ['bɛːl]('mind') + personal pronoun suffix>, which has a value equivalent to the verb 'be aware of', as in the example: <ما في باليش>, [ma fiː bɛːl-ˈiː-ʃ], 'I did not know'.

3.2 Tunisian Arabish

As previously mentioned, we consider Arabish a quasi-oral system. With *quasi-orality* it is intended the form of communication typical of Computer-Mediated Communication (CMC), characterized by informal tones, dependence on context, lack of attention to spelling and especially the ability to create a sense of collectivity (Hert, 1999)[12].
TA and TUN have not a standard orthography, with the exception of the CODA TUN. Nevertheless, TA is a spontaneous code-system used since more than ten years, and is being conventionalized by its daily usage. From the table 1, where the coding scheme of TA is illustrated, it is possible to observe that there is no one-to-one correspondence between TA and TUN characters and that often Arabish presents overlaps in the encoding possibilities. The main issue is represented by the not proper representation by TA of the emphatic phones: [ðˤ], [tˤ] and [sˤ].
On the other hand, being TA not codified through the Arabic alphabet, it can well represent the phonetic realization of TUN, as shown by the following examples:

[8] Reduction of the diphthongs [aw] and [aj] to [uː] and [iː] in pre-Hilalian dialects, and to [oː] and [eː] in the Hilalian ones.

[9] Transposition of the first vowel of the word. It occurs when non-conjugated verbs or names without suffix begin with the sequence CCvC, where C stands for ungeminated consonant, and 'v' for short vowel. When a suffix is added to this type of name, or a verb of this type is conjugated, the first vowel changes position giving rise to the CvCC sequence.

[10] The morpheme /-t/ can be traced back to the same morpheme present in the V and VI verbal patterns of CA (Mion, 2004).

[11] Ḥafṣiya is a neighborhood in the Medīna of Tunis, known for its great daily frīp (second-hand market).

[12] Even though the CMC is generally a type of asynchronous communication.

IPA	*TUN*	*TA*	*IPA*	*TUN*	*TA*
[aː]	ة	a, e, h	[a][aː]	ى, ا	a, e, é, è
[ʔ]	ء	2	[ðˤ]	ض	dh, th, d
[b]	ب	b, p	[tˤ]	ط	6, t
[t]	ت	t	[ðˤ]	ظ	th, dh
[θ]	ث	th	[ʕ]	ع	3, a
[ʒ]	ج	j	[ɣ]	غ	4, gh
[ħ]	ح	7, h	[f]	ف	f
[x]	خ	5, kh	[q]	ق	9, q
[d]	د	d	[k]	ك	k
[ð]	ذ	dh	[l]	ل	l
[r]	ر	r	[m]	م	m
[z]	ز	z	[n]	ن	n
[s]	س	s	[h]	ه	8, h
[ʃ]	ش	ch, (sh)	[w][uː]	و	ou, w
[sˤ]	ص	s	[j][iː]	ي	i, y

Table 1: Arabish code-system for TUN

***** The Arabic alphabet is generally used for formal conversations in Modern Standard Arabic (MSA), the Arabic of formal situations, or in that of Classical Arabic (CA), the Arabic of the Holy Qur'ān, also known as 'The Beautiful Language'. Like MSA and CA, also Arabic Dialects (ADs) can be written in the Arabic alphabet, but in this case it is possible to observe a kind of hypercorrection operated by the speakers in order to respect the writing rules of MSA. For example, in TUN texts written in Arabic script, it is possible to find a 'silent vowel' (namely an epenthetic 'alif <ا>) written at the beginning of those words starting with the sequence '#CCv', which is not allowed in MSA.

***** Writing TUN in Arabic script, the Code-Mixing or Switching in foreign language will be unnaturally reduced.

***** As described in table 1, the Arabic alphabet is provided with three short vowels, which correspond to the three long ones: [aː], [uː], [iː], but TUN presents a wider range of vowels. Indeed, regarding the early presented characteristics of TUN, the TA range of vowels offers better possibility to represent most of the TUN characteristics outlined in the previous subsection, in particular:

- Palatalization.
- Vowel metathesis.
- Monophthongization.[13]

4 Tunisian Arabish Corpus

In order to analyze the TA system, we have built a TA Corpus based on social media data, considering this as the best choice to observe the quasi-oral nature of the TA system.

[13]Regarding the last two phenomena, they can be visible in Arabic script only in case of texts provided with short vowels, which are quite rare.

4.1 Text collection

The corpus collection procedure is composed of the following steps:

1. Thematic categories detection.
2. Match of categories with sets of semantically related TA keywords.
3. Texts and metadata extraction.

Step 1. In order to build a Corpus that was as representative as possible of the linguistic system, it was considered useful to identify wide thematic categories that could represent the most common topics of daily conversations on CMC.

In this regard, two instruments with a similar thematic organization have been employed:

- **'A Frequency Dictionary of Arabic'** (Buckwalter and Parkinson, 2014) In particular its 'Thematic Vocabulary List' (TVL).
- **'Loanword Typology Meaning List'** A list of 1460 meanings[14] (LTML) (Haspelmath and Tadmor, 2009).

The TVL consists of 30 groups of frequent words, each one represented by a thematic word. The second consists of 23 groups of basic meanings sorted by representative word heading. Considering that the boundaries between some categories are very blurred, some categories have been merged, such as 'Body' and 'Health', (see table 2). Some others have been eliminated, being not relevant for the purposes of our research, e.g. 'Colors', 'Opposites', 'Male names'. In the end, we obtained 15 macro-categories listed in table 2.

Step 2. Aiming at easily detect texts and the respective *seed URLs*, without introducing relevant query biases, we decided to avoid using the category names as query keywords (Schäfer and Bildhauer, 2013). Therefore, we associated to each category a set of TA keywords belonging to the basic Tunisian vocabulary. We found that a semantic category with three meanings was enough to obtain a sufficient number of keywords and URLs for each category. For example, to the category 'Family' the meanings: 'son', 'wedding', 'divorce' have been associated in all their TA variants, obtaining a set of 11 keywords (table 2).

Step 3. We collected about 25,000 words and the related metadata as first part of our corpus, which are being semi-automatically transcribed into Arabic characters (see next sections). We planned to increase the size of the corpus at a later time. Regarding the metadata, we have extracted the information published by users, focusing on the three types of information generally used in ethnographic studies:

1. Gender: Male (M) and Female (F).

[14]The 'Loanword Typology Meaning List' is a result of a joint project by Uri Tadmor and Martin Haspelmath: the 'Loanword Typology Project' (LWT), launched in 2004 and ended in 2008.

Macro-Categories	Words Associated
1. Family *son, wedding, divorce*	weld, wild, 3ars, 3ers, tla9, 6la9, tlaq, 6laq, tle9, tleq, 6leq
2. Clothing *dress, shoes, t-shirt*	robe, lebsa, rouba, sabat, spedri, spadri, marioul, maryoul, meryoul, merioul
3. Automobiles *gasoil, engine, occasion*	mazout, motor, moteur, motour, forsa
4. Animals *cock, dog, cat*	sardouk, kelb, kalb, 9attous, gattous
5. Body and Health *sick, doctor, health*	maridh, marith, mridh, ettbib, tbib, sa77a, sa7a, sahha, saha

Table 2: Example of the fifteen thematic categories

2. Age range: [10-25], [25-35], [35-50], [50-90].
3. City of origin.

4.2 Corpus Creation

In order to create our corpus, we applied a word-level annotation. This phase was preceded by some data pre-processing steps, in particular tokenization. Each token has been associated with its annotations and metadata (table 3). In order to obtain the correspondence between Arabish and Arabic morpheme transcriptions, tokens were segmented into morphemes. This segmentation was carried out completely manually for a first group of tokens.[15] In its final version, each token is associated with a total of 11 different annotations, corresponding to the number of the annotation levels we chose. An excerpt of the corpus after tokens annotation is depicted in table 3.
For the sake of clarity, in table 3 we show:
* The A column, *Cor*, indicates the token's source code. For example, the code *3fE*, which stands for *3rab fi Europe*, is the forum from which the text was extracted.
* The B column, *Textco*, is the publication date of the text.
* The C column, *Par*, is the row index of the token in the paragraph.
* The D column, *W*, is the index of the token in the sentence. When 'W' corresponds to a range of numbers, it means that the token has been segmented in to its components, specified in the rows below.
* The E column, *Arabiʃ*, corresponds to the token transcription in Arabish.
* The F column, *Tra*, is the transcription into Arabic characters.
* The G column, *Ita*, is the translation to Italian.

[15]Arabic, in general, is a language with a high level of synthesis, that means that it can concentrate within a token more syntactic and grammatical information through the addition of different morphemes.

* The H column, *Lem*, corresponds to the lemma.
* The I column, *POS*, is the Part-Of-Speech tag of the token. The tags that have been used for the POS tagging are conform to the annotation system of Universal Dependencies.
* The last three columns (J, K, L) contain the metadata: *Var*, *Age*, *Gen*.

A	B	C	D	E	F	G
Cor	Textco	Par	W	Arabiʃ	Tra	Ita
3fE	150902	2	1	kifech	كيفاش	come
3fE	150902	2	2	tchou-fou	تشوفوا	vi pare
3fE	150902	2	3-4	l3icha	العيشة	la vita
3fE	150902	2	3	l	الـ	-
3fE	150902	2	4	3icha	عيشة	-
3fE	150902	2	5-6	fil	فالـ	all'
3fE	150902	2	5	f	فـ	-
3fE	150902	2	6	il	الـ	-
3fE	150902	2	7	4orba	غربة	estero
3fE	150902	2	8	?	؟	?

H	I	J	K	L
Lem	POS	Var	Age	Gen
كيفاش	adv	Bnz	25-35	M
شاف	verb	Bnz	25-35	M
عيشة	noun	Bnz	25-35	M
الـ	det	Bnz	25-35	M
عيشة	noun	Bnz	25-35	M
في	prep	Bnz	25-35	M
في	prep	Bnz	25-35	M
الـ	det	Bnz	25-35	M
غربة	noun	Bnz	25-35	M
؟	pct	Bnz	25-35	M

Table 3: An Excerpt of the TArC structure. In the column *Var*, 'Bnz' stands for 'Bizerte' a northern city in Tunisia. Glosses: w1:*how*, w2:*do you(pl) see*, w3-4:*the life*, w5-6:*at the*, w7:*outside*, w8:*?*

Since TA is a spontaneous orthography of TUN, we considered important to adopt the CODA* guidelines as a model to produce a unified lemmatization for each token (column *Lem* in table 3). In order to guarantee accurate transcription and lemmatization, we annotated manually the first 6,000 tokens with all the annotation levels.
Some annotation decisions were taken before this step, with regard to specific TUN features:
* **Foreign words.** We transcribed the Arabish words into Arabic characters, except for Code-Switching terms. In order to not interrupt the sentences continuity we decide to transcribe Code-Mixing terms into Arabic script. However, at the end of the corpus creation process, these words will be analyzed, making the distinction between acclimatized loans and Code-Mixing. The first ones will be transcribed into Arabic

characters also in *Lem*, as shown in table 4. The second ones will be lemmatized in the foreign language, mostly French, as shown in table 5.

*** Typographical errors.** Concerning typos and typical problems related to the informal writing habits in the web, such as repeated characters to simulate prosodic features of the language, we have not maintained all these characteristics in the transcription (column *Tra*). Logically, these were neither included in *Lem*, according to the CODA* conventions, as shown in table 5.

W	Arabiʃ	Tra	Ita	Lem	POS
4	konna	كنّا	siamo stati	كان	verb
5	far7anin	فرحانين	contenti	فرحان	adj
6	,	,	,	,	punct
7	merci	مرسي	grazie	مرسي	intj

Table 4: Loanword example in the corpus. Glosses: w4:*we were*, w5:*happy*, w6:*,* , w7:*thanks*

W	Arabiʃ	Tra	Ita	Lem	POS
1	R7	recette	ricetta	recette	noun
2	patee	pâté	patè	pâté	noun
3	dieri	دياري	fatto in casa	دياري	adj
4	w	و	e	و	cconj
5	bniiiiin	بنين	buonissimo	بنين	adj

Table 5: Prosody example in the corpus. Glosses: w1:*recipe*, w2:*pâté*, w3:*homemade*, w4:*and*, w5:*delicious*

*** Phono-Lexical exceptions.** We used the grapheme <ڤ>, [q], only in loanword transcription and lemmatization. As can be seen in table 6, the Hilalian phoneme [g] of the Turkish loanword 'gawriyya', has been transcribed and lemmatized with the grapheme <ڤ>, [g].

W	Arabiʃ	Tra	Ita	Lem	POS
1	Mtala9	مطلق	divorziato	مطلق	noun
2	min	من	da	من	noun
3	gawriya	ڤاورية	(un')europea	ڤاوري	adj

Table 6: Phono-Lexical exceptions in the corpus. Glosses: w1:*divorced*, w2:*from*, w3:*European(f)*

*** Glottal stop.** As explained in CODA TUN, real initial and final glottal stops have almost disappeared in TUN. They remain in some words that are treated as exceptions, e.g. <أسئلة>, ['ʔasʔla], 'question' (Zribi et al., 2014). Indeed, we transcribe the glottal stops only when it is usually pronounced, and if it does not, we do not write the glottal stops at the beginning of the word or at the end, neither in the transcription, nor in the lemmas.

*** Negation Marks.** CODA TUN proposes to keep the MSA rule of maintaining a space between the first negation mark and the verb, in order to uniform CODA TUN to the first CODA (Habash et al., 2012). However, as Zribi et al. (2014) explains, in TUN this rule does not make really sense, but it should be done to preserve the consistency among the various CODA guidelines. Indeed, in our transcriptions we report what has been produced in Arabish following CODA TUN rules, while in lemmatization we report the verb lemma. At the same time we segment the negative verb in its minor parts: the circumfix negation marks and the conjugated verb. For the first one, we describe the negative morphological structure in the *Tra* and *Lem* columns, as in table 7. For the second one, as well as the other verbs, we provide transcription and lemmatization.

W	Arabiʃ	Tra	Ita	Lem	POS
14-15	manajem-nech	ما نجمناش	non abbiamo potuto	نجّم	verb
14	ma + ch	ما+ش	-	ما+V+ش	part
15	najemne	نجمنا	-	نجّم	verb

Table 7: Circumfix negation marks in the corpus. Glosses: w14-15:*we could not*

5 Incremental and Semi-Automatic Transcription

In order to make the corpus collection easier and faster, we adopted a semi-automatic procedure based on sequential neural models (Dinarelli and Grobol, 2019b; Dinarelli and Grobol, 2019a). Since transcribing Arabish into Arabic is by far the most important information to study the Arabish code-system, the semi-automatic procedure concerns only transcription from Arabish to Arabic script.

In order to proceed, we used the first group of (roughly) 6,000 manually transcribed tokens as training and test data sets in a 10-fold cross validation setting with 9-1 proportions for training and test, respectively. As we explained in the previous section, French tokens were removed from the data. More precisely, whole sentences containing *non-transcribable* French tokens (code-switching) were removed from the data. Since at this level there is no way for predicting when a French word can be transcribed into Arabic and when it has to be left unchanged, French tokens create some noise for an automatic, probabilistic model. After removing sentences with French tokens, the data reduced to roughly 5,000 tokens. We chose this amount of tokens for annotation blocks in our incremental annotation procedure.

We note that by combining sentence, paragraph and token index in the corpus, whole sentences can be reconstructed. However, from 5,000 tokens roughly 300 sentences could be reconstructed, which are far too

few to be used for training a neural model.[16] Instead, since tokens are transcribed at morpheme level, we split Arabish tokens into characters, and Arabic tokens into morphemes, and we treated each token itself as a sequence. Our model learns thus to map Arabish characters into Arabic morphemes.

The 10-fold cross validation with this setting gave a token-level accuracy of roughly 71%. This result is not satisfactory on an absolute scale, however it is more than encouraging taking into account the small size of our data. This result means that less than 3 tokens, on average, out of 10, must be corrected to increase the size of our corpus. With this model we automatically transcribed into Arabic morphemes, roughly, 5,000 additional tokens, corresponding to the second annotation block. This can be manually annotated in at least 7,5 days, but thanks to the automatic annotation accuracy, it was manually corrected into 3 days.[17] The accuracy of the model on the annotation of the second block was roughly 70%, which corresponds to the accuracy on the test set. The manually-corrected additional tokens were added to the training data of our neural model, and a new block was automatically annotated and manually corrected. Both accuracy on the test set and on the annotation block remained at around 70%. This is because the block added to the training data was significantly different from the previous and from the third. Adding the third block to the training data and annotating a fourth block with the new trained model gave in contrast an accuracy of roughly 80%. This incremental, semi-automatic transcription procedure is in progress for the remaining blocks, but it is clear that it will make the corpus annotation increasingly easier and faster as the amount of training data will grow up.

Our goal concerning transcription, is to have the 25,000 tokens mentioned in section 4.1 annotated automatically and manually corrected. These data will constitute our gold annotated data, and they will be used to automatically transcribe further data.

6 Conclusions

In this paper we presented TArC, the first Tunisian Arabish Corpus annotated with morpho-syntactic information. We discussed the decisions taken in order to highlight the phonological and morphological features of TUN through the TA corpus structure. Concerning the building process, we have shown the steps undertaken and our effort intended to make the corpus as representative as possible of TA. We therefore described the texts collection stage, as well as the corpus building and the semi-automatic procedure adopted for transcribing TA into Arabic script, taking into account CODA* and CODA TUN guidelines. At the present stage of research, TArC consists of 25.000 tokens, however our work is in progress and for future research we plan to enforce the semi-automatic transcription, which has already shown encouraging results (accuracy = 70%). We also intend to realize a semi-automatic TA Part-Of-Speech tagger. Thus, we aim to develop tools for TA processing and, in so doing, we strive to complete the annotation levels (transcription, POS tag, lemmatization) semi-automatically in order to increase the size of the corpus, making it available for linguistic analyses on TA and TUN.

[16] Preliminary experiments gave indeed quite poor results, below 50% token-level accuracy on average.

[17] We based our estimations of the annotation time needed on the time we spent correcting tokens, which is actually faster because tokens are already transcribed, they don't need to be transcribed from scratch.

7 Bibliographical References

Akbar, R. (2019). Arabizi among kuwaiti youths: Reshaping the standard arabic orthography. *International Journal of English Linguistics*, 9(1):301–323.

Al-Badrashiny, M., Eskander, R., Habash, N., and Rambow, O. (2014). Automatic transliteration of romanized dialectal arabic. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 30–38.

Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., Strassel, S., Habash, N., Eskander, R., and Rambow, O. (2014). Transliteration of arabizi into arabic orthography: Developing a parallel annotated arabizi-arabic script sms/chat corpus. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 93–103.

Boujelbane, R. J. (2015). *Traitements linguistiques pour la reconnaissance automatique de la parole appliquée à la langue arabe: de l'arabe standard vers l'arabe dialectal*. Ph.D. thesis, Aix-Marseille.

Buckwalter, T. and Parkinson, D. (2014). *A frequency dictionary of Arabic: Core vocabulary for learners*. Routledge.

Chalabi, A. and Gerges, H. (2012). Romanized arabic transliteration. In *Proceedings of the Second Workshop on Advances in Text Input Methods*, pages 89–96.

Darwish, K. (2013). Arabizi detection and conversion to arabic. *CoRR*, abs/1306.6755.

Dinarelli, M. and Grobol, L. (2019a). Hybrid neural models for sequence modelling: The best of three worlds. *CoRR*.

Dinarelli, M. and Grobol, L. (2019b). Seq2biseq: Bidirectional output-wise recurrent neural networks for sequence modelling. *CoRR*, abs/1904.04733.

Durand, O. (2009). *Dialettologia araba*. 'Sapienza' University of Rome, 'Studi Orientali' Faculty.

Graja, M., Jaoua, M., and Hadrich-Belguith, L. (2013). Discriminative framework for spoken tunisian dialect understanding. In *International Conference on Statistical Language and Speech Processing*, pages 102–110. Springer.

Habash, N., Diab, M., and Rambow, O. (2012). Conventional orthography for dialectal Arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 711–718, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Habash, N., Eryani, F., Khalifa, S., Rambow, O., Abdulrahim, D., Erdmann, A., Faraj, R., Zaghouani, W., Bouamor, H., Zalmout, N., Hassan, S., Al-Shargi, F., Alkhereyf, S., Abdulkareem, B., Eskander, R., Salameh, M., and Saddiki, H. (2018). Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Haspelmath, M. and Tadmor, U. (2009). *Loanwords in the world's languages: a comparative handbook*. Walter de Gruyter.

Hert, P. (1999). Quasi-oralité de l'écriture électronique et sentiment de communauté dans les débats scientifiques en ligne. *Réseaux*, 17(97).

Marcais, P. (1977). Esquisse grammaticale de l'arabe maghrébin. *Langues d'Amérique et d'Orient*.

Masmoudi, A., Khmekhem, M. E., Esteve, Y., Hadrich-Belguith, L., and Habash, N. (2014). A corpus and phonetic dictionary for tunisian arabic speech recognition. In *LREC*, pages 306–310.

Masmoudi, A., Habash, N., Ellouze, M., Estève, Y., and Hadrich-Belguith, L. (2015). Arabic transliteration of romanized tunisian dialect text: A preliminary investigation. In *Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 608–619. Springer-Verlag, 1.

Medhaffar, S., Bougares, F., Estève, Y., and Hadrich-Belguith, L. (2017). Sentiment analysis of Tunisian Dialects: Linguistic ressources and experiments. In *Proceedings of the third Arabic Natural Language Processing Workshop*, pages 55–61. Association for Computational Linguistics.

Meftouh, K., Harrat, S., and Smaïli, K. (2018). Padic: extension and new experiments. In *7th International Conference on Advanced Technologies ICAT*.

Mion, G. (2004). Osservazioni sul sistema verbale dell'arabo di tunisi. *Rivista degli studi orientali*, 78(Fasc. 1/2):243–255.

Schäfer, R. and Bildhauer, F. (2013). Web corpus construction. *Synthesis Lectures on Human Language Technologies*, 6(4):1–145.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of NIPS*, Cambridge, MA, USA. MIT Press.

Younes, J., Achour, H., Souissi, E., and Ferchichi, A. (2018a). Survey on corpora availability for the tunisian dialect automatic processing. In *2018 JCCO Joint International Conference on ICT in Education and Training, International Conference on Computing in Arabic, and International Conference on Geocomputing (JCCO: TICET-ICCA-GECO)*, pages 1–7. IEEE.

Younes, J., Souissi, E., Achour, H., and Ferchichi, A. (2018b). A sequence-to-sequence based approach for the double transliteration of tunisian dialect. *Procedia computer science*, 142:238–245.

Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Hadrich-Belguith, L., and Habash, N. (2014). 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, 2014*, pages 2355–2361.

Zribi, I., Ellouze, M., Hadrich-Belguith, L., and Blache, P. (2015). Spoken tunisian arabic corpus "stac": Transcription and annotation. *Research in Computing Science*, 90:123–135.

8 Language Resource References

Graja, Marwa; Jaoua, Maher; Hadrich-Belguith, Lamia. (2010). *Tunisian Dialect Corpus Interlocutor (TuDiCoI)*. MIRACL Laboratory, Sfax (Tunis), Arabic Natural Language Processing Research Group (ANLP), 1.0.

Masmoudi, Abir; Ellouze Khemakhem, Mariem; Estève, Yannick; Hadrich-Belguith, Lamia. (2014). *Tunisian Arabic Railway Interaction Corpus*. MIRACL Laboratory, Sfax (Tunis), Arabic Natural Language Processing Research Group (ANLP), 1.0.

Medhaffar, Salima and Bougares, Fethi and Estève, Yannick and Hadrich-Belguith, Lamia. (2017). *Tunisian Hadrich-Belguith Analysis Corpus (TSAC)*.

Mourad, Abbas. (2017). *Parallel Arabic Dialectal Corpus (PADIC)*. Scientific and Technical Research Center for the Development of Arabic Language, funded by the Algerian Ministry of Higher Education and Scientific Research, language resources, 2.0, ISLRN 429-053-323-228-4.

Zribi, Inés; Ellouze Khemakhem, Mariem; Hadrich-Belguith, Lamia; Blache, Philippe. (2015). *Spoken Tunisian Arabic corpus (STAC)*. MIRACL Laboratory, Sfax (Tunis), Arabic Natural Language Processing Research Group (ANLP), 1.0.