QUADERNI DEL CERM

ELISA GUGLIOTTA

# TUNISIAN ARABIZI

## LINGUISTIC ANALYSES AND CORPUS BUILDING USING NATURAL LANGUAGE PROCESSING

Ledizioni

Elisa Gugliotta

# Tunisian Arabizi: Linguistic Analyses and Corpus Building using Natural Language Processing

Ledizioni

This publication has been made possible thanks to the financial contribution of the Department of Theoretical and Applied Sciences (DISTA), Università degli Studi dell'Insubria, Italy.


Via Boselli 10 – 20136 Milan, Italy
www.ledizioni.it
info@ledizioni.it

Elisa Gugliotta, *Tunisian Arabizi: Linguistic Analyses and Corpus Building using Natural Language Processing*
First edition: September 2024

ISBN print 9791256002139
ISBN ePub 9791256002146
Graphic layout: Ledizioni
Cover image by Laura Gugliotta

Catalogue and reprints information:
www.ledizioni.it, www.ledipublishing.com

## Scientific Committee

# TABLE OF CONTENTS

# TRANSCRIPTION AND TRANSLITERATION TABLE

With regard to the system of transcription and transliteration chosen for this work, in the hopes that this work will reach a wide and varied readership, we have opted for an IPA-oriented transcription system. Moreover, when dealing with a writing system, we could not take for granted the phonetic realisation intended behind the graphemes that composed the texts we used. As pointed out by Durand (2012) and Mion (2010), the phonological representation for Tunisian flattens the phonetic richness of this variety, which presents, for example, a vowel system of six phonemes, i.e. three short and three long, as shown in the table below. However, the phonetic realisation of these phonemes varies depending on many factors, such as the diatopic, diastratic, the phonemic context, etc. Therefore, a phonological rendering through the graphemes is generally adopted, which will undoubtedly be reductive, not only with regards to the vocalic system. Therefore, for our work, an exact phonetic reconstruction would be impossible, which is why we opted for a mixed IPA-oriented transliteration system oriented toward the Tunisian *koiné*, which, as we will see below, represents the 'nationalc colloquial variety, even if in dialogue with the local varieties. For the same reason, we have reported short vowels as *schwa*, /ə/, with the exception of the short vowel [u]. In the tables below, we report the International Phonetic Alphabet symbols associated with the phonetic realisation of the grapheme following the Tunisian *koiné* realisation. Next to this, we report the employed transcription system (between // in the text), the grapheme name and its encoding in Arabic script. Arabizi texts in the examples will be in italics.

Regarding [g]* in the table below, it could be considered to be a qāf allophone, but it is also a loan phoneme from the Romance languages.

| IPA | Used Transcription | Grapheme Name | Graphemes |
|---|---|---|---|
| [b] | b | bāʔ | ب |
| [t] | t | tāʔ | ت |
| [θ] | θ | θāʔ | ث |
| [ʤ][ʒ] | ǧ, ž | ǧīm | ج |
| [ħ] | ħ | ħāʔ | ح |
| [χ] | ḫ | ḫāʔ | خ |
| [d] | d | dāl | د |
| [ð] | ð | ðāl | ذ |
| [r] | r | rāʔ | ر |
| [z] | z | zāy | ز |
| [s] | s | sīn | س |
| [ʃ] | š | šīn | ش |
| [sˤ] | ṣ | ṣād | ص |
| [dˤ] | ḍ | ḍād | ض |
| [tˤ] | ṭ | ṭāʔ | ط |
| [ðˤ] | ẓ | ẓāʔ | ظ |
| [ʕ] | ʕ | ʕayn | ع |
| [ʁ] | ɣ | ɣayn | غ |
| [f] | f | fāʔ | ف |
| [q] | q | qāf | ق |
| [g]* | g | gāf | ڤ |
| [k] | k | kāf | ك |
| [l] | l | lām | ل |
| [m] | m | mīm | م |
| [n] | n | nūn | ن |
| [h] | h | hāʔ | ه |
| [ʔ] | ʔ | hamza | ء |

*Table 1. Table of Transcription and Transliteration consonant system used.*

| Long Vowel Tr. Used | Grapheme Name | Graphemes | Short Vowel Tr. Used |
|---|---|---|---|
| *ā* | ʔalif | ا | ə |
| *ī* | yāʔ | ي | ə |
| *ū* | wāw | و | u |

*Table 2. Table of Transcription and Transliteration vowel system used.*

*If I were to dedicate this work to those who have supported and inspired me over time, I could go on indefinitely listing names, as some people do. I would prefer to avoid listing names at a specific point in time and space. However, I cannot avoid thanking Tunisia, which has given me much more than I could ask for, much more than I could give back.*

# INTRODUCTION

The object of this work is Tunisian Arabic, and more specifically its encoding in *Arabizi*. Arabizi is an orthographic encoding, which arose spontaneously in a diamesically influenced context, namely digital environments. The encoding we refer to uses the Latin alphabet, as well as some numbers, for Tunisian Arabic phonemes with no correspondence in the Latin script.

The development of the Internet, along with that of technology, transformed written communication, leading to the rise of its proper mode: electronic written communication. This new form of communication has continued to spread widely since the mid-1990s. Technology in general has had a major impact on our everyday practices, the way we read the news, our work environment, the way we socialise and how we express our identity. Until a few years ago, many introductory texts, when referring to the so-called Arabic dialect, henceforth *dārža*, i.e. the Tunisian autoglottonym, used to describe it as a primarily oral variety, being almost completely without a written tradition, with just a few rare exceptions. In terms of Tunisian Arabic, there are in fact some texts written in *dārža*. However, it has already been at least twenty years since Tunisian native speakers started writing on social media, every day, in *dārža*. This marked a major switch in language practices, which has been so significant that it might signal a break from the traditional definition of Tunisian Arabic as being 'predominantly oral'. Moreover, this is not the only manifestation of the writing in *dārža*. Reference can be made to comics, the practice of writing graffiti on walls, advertising panels, blogs, forums, and in general to Computer-Mediated Communication (CMC), which is the main means of expression for the so-called 'digital natives' and their wish to communicate by writing in *dārža*.

As such, it can be considered to be a very powerful identity marker. Therefore, we need to analyse how the Tunisian digital generation spontaneously organised itself to communicate in its own *dārža*.

Certainly, in the age of the Internet and CMC, writing and literacy rapidly takes on new forms and functions with a number of consequences for language and culture, as well as for economics and politics. Indeed, another example of how non-state agents have made use of new communication technologies with far-reaching consequences for both social and political events was the so-called Arab Springs of 2011. Regarding the vernacular writing practice, Auer (1997) relates it to the koineisation process. The latter results in the adoption of a national urban *koiné*, which becomes prevalent in public spaces, while the vernacular, which is specific to the place of origin, is confined to domestic interactions.

On the other hand, Auer (1997) defines this process as *destandardisation* explaining that it consists of different regional vernacular features converging and simultaneous standard variety decreasing of the standard status flowing into larger-scale regional *koinai*. When taking the wide diffusion of the practice of writing on social networks into consideration, as such a spontaneous and informal mode of communication, this allows us to imagine that an analysis of data gathered from the CMC context can reveal what kind of spontaneous linguistic dynamics of this type are currently taking place in Tunisia. Arabizi seems to be the most representative object of study in this context of computer-mediated writing. In fact, Androutsopoulos (2012) has already addressed a similar phenomenon, namely *Greeklish*, through both approaches, i.e. the autonomous one and the ideological one, specifying that an *autonomous approach* sees orthography as a *neutral technology for the representation of spoken language*, in contrast with the so-called *ideological approach* that views orthography as a *set of social practices in specific social and cultural contexts*.

Given that technology significantly impacts our language use and interpersonal interactions, research on this topic encompasses various disciplines, including communication sciences, psychology, and sociology. Methodological approaches to analyzing this phenomenon vary, from comparative studies of different discourse modes to investigations into specific linguistic phenomena.

We focus on the Arabizi system in use for writing Tunisian Arabic in order to observe the Tunisian Arabizi diachronic evolution on the one hand, but to also bear witness to the distribution of certain textual features depending on the more or less informal context. To this end, we consider a description of the Tunisian multilingual context to be an invaluable jumping-off point, in terms of diglossia and bilingualism. Therefore, with the purpose of introducing Tunisian

Arabic, we can delve deep into the multilingual and diglossic context that characterises it.

Specifically, our work begins by tracing back some of the historical linguistic milestones of Tunisian, which will constitute some of the tools necessary to depict the current linguistic complexity. We begin with the historical survey to consciously address questions regarding the status of Tunisian Arabic. We will then move into deeper linguistic issues, and we will see how Tunisian Arabic is classified in terms of traditional dialectological descriptions. We will also observe how this Tunisian multilingual society has been historically and modernly described by scholars, and what role Tunisian Arabic takes in society compared to that of Modern Standard Arabic (MSA) and French, in an attempt to describe all the nuances that exist between the different shades of the Tunisian linguistic ideology. We will finally examine the evolution of the Tunisian socio-linguistic landscape, which is a fairly common component part when studying the history of languages, namely the emergence of urban varieties with the status of *koiné*. As mentioned previously, this process is particularly relevant to the analysis of Tunisian CMC. In fact, it is in particular the *koiné* of Tunis that has had consistent diffusion through the early means of diffusion of oral texts, such as radio and television.

In fact, we asked ourselves several questions at the beginning of the project. The first concerns the kind of language we should expect to find in these digital contexts. Hary (1996) defines *multiglossia* as the linguistic situation in which different varieties coexist side by side in a language community, and that each one is used in different circumstances and has different functions. Thinking back to the principles of Hary's functions of communication, it comes to mind that a certain variations will be found depending on the contexts of communication and therefore on the channels. In our case, the channels are forums, blogs or platforms for simultaneous communication, such as Facebook.

At the same time, a series of questions arise regarding the hypothesis of a continuum of (in)formality, given the more or less spontaneous context of the communication. One of the objectives of the preliminary analyses set out in Chapter 4 is in fact to verify whether there is a basis for assuming the existence of this continuum between textual genres. In this case in which the continuum would go from a formal to an informal textual genre pole, the more formal one could be represented by genres such blogs, which are often narrative and do not involve simultaneous communication.

The more informal could be those marked by communication exchanges, which usually occur in sync, where real networks of social connection are created or reproduced, such as Facebook texts. An intermediate genre could be represented by those typical of certain platforms with characteristics shared between the two poles, such as forums. In fact, forums allow us to find real networks of social connection, despite not being in sync.

Another issue that we will explore through the analyses is the quasi-oral nature of this encoding system. Indeed, considering Arabizi as a non-standardised system, it is possible to assume that it allows users to be more independent from strong writing traditions, such as the one related to Arabic writing. This is one of the hypotheses about Arabizi that we will address. Furthermore, Arabizi presents a huge variety in its orthography, with many different possibilities for encoding one word. Our aim is to investigate precisely whether this freedom of graphic realisation is constrained by the fact that Arabizi belongs to a specific writing context. Before the invasion of Facebook, when one wrote in Tunisian, one did so in Arabic characters and according to some implicit but widely shared rules of 'correct encoding'. We would also like to consider the possibility that these rules are based on a transfer that took place through the Arabic alphabet when employed in Tunisian Arabic writing practices. Another perspective is that Arabizi enables users to circumvent certain rules typically enforced by Arabic spelling, with the focus shifting to 'correct encoding' in Tunisian Arabic. This possibility includes the concept of Arabizi being influenced by French orthography, being a Latin character encoding, and facilitating the use of French vocabulary during code-mixing.

A further preliminary analysis that we were interested in was related to the process of the settling of Arabizi writing practices, which we have defined as being a spontaneous and non-standardised system. However, each Arabizi national variety should have been subjected to a process of spontaneous settling since its inception. This is true also for Tunisian Arabizi.

In order to be able to carry out our analyses on this writing system, we must first describe the tools that enabled us to conduct our research. Indeed, our present work comes with the objective of building a set of tools to support the analysis of this language considering the lack of resources which can be used to conduct research on Tunisian Arabizi. A tool we built is called the *Tunisian Arabish Corpus* (TArC). It is a corpus that collects texts from different CMC

contexts, such as forums, blogs and social networks, amounting to 43,327 words. TArC has been provided with different levels of linguistic annotation in order to support different types of analysis. As for the name we chose for the corpus, the Arabizi encoding is known by many different names, depending on which Arabic country you are in. The name 'Arabish' was used, rather than 'Arabizi', because it is easily understandable even among non-Arabic speakers. Nevertheless, the term Arabish also allows for a direct reference to similar phenomena. In fact, many linguistic systems, which are formally encoded through Roman script, use the Latin alphabet for informal exchange in CMC contexts, among them are the previously mentioned *Greeklish* (which is blend of 'Greek' and 'English') or *Pinglish* (a merge between 'Farsi' or 'Persian' and 'English'). However, the term 'Arabizi' is now more widespread than 'Arabish', even if it is not so transparent, as it is the *portmanteau* of 'Arabic' and 'English' (/ʔinglīzī/), which is the Arabic term denoting the English language. For this reason, we decided to refer to this encoding, throughout this body of work, using the term Arabizi.

In order to perform the mentioned analyses, a corpus of a certain size was required, so that it could represent the different textual varieties, but could also support analyses in diachrony and diatopy, as well as the general observation of diastratic variations. We will explore what resources are available for the study of Tunisian and its encoding in Arabizi, but we had already anticipated the fact that, given the breadth of the issues we were interested in observing, the best solution for us was to create a corpus from scratch. We therefore decided to opt for a hybrid methodology that would allow us to build a corpus of Tunisian Arabizi of a sufficient size to contain a certain degree of variability and that would have levels of linguistic information with which to facilitate our analyses.

For many reasons, which are mentioned below, the adjective that best describes our approach to this work is 'hybrid'. First of all, the project is located in a space of intersection between different research fields: Arabic dialectology, corpus linguistics, and Natural Language Processing (NLP). Second, the texts collected in our corpus are encoded in a script that has hybrid characteristics of orality while also being a writing system. Once we opted for this hybrid method, we also realised that this would allow us to expand the pool of utility of the tool itself. In fact, we could integrate the construction procedure with semi-automatic corpora construction procedures. This choice has meant that the corpus can then be used in NLP research.

Finally, our motivations for building a corpus from scratch coincide with the utility we envisioned upon its release, both from a linguistic and an NLP perspective.

After defining the motivations that led us to build the Tunisian Arabish Corpus (TArC) from scratch and the methodology we adopted based on the state-of-the-art we will move on to describing the specific operations that led us, step by step, to the TArC realisation. The steps we will describe follow the path from the beginning, starting with the collection of data and then moving on to decisions regarding the selection and collection of metadata. We will also describe the stages of semi-automatic annotation of the corpus into its component layers: i.e. the transliteration of Arabizi words into Arabic script, the tokenisation of words in morphemes, Part-of-Speech tagging and lemmatisation. We will then review the steps and the experiments that led us to identify the best strategies with which to achieve our corpus building goals. Finally, we will describe the Multi-Task Sequence Prediction architecture that was built to produce the different annotation layers that make up TArC from the Arabizi texts. In the end, we will also describe the TArC structure, along with information about the amount of data and metadata.

Finally, before delving into the details of this research, we would like to make an anticipation in order to immediately highlight the boundaries of the research here presented. These boundaries therefore exist due to an *a priori* choice regarding the positioning of this study in the field. This research, in fact, aspiring to be placed in an interstitial point between different research fields, with the aim of combining tools and goals common to them, runs the risk of being perceived as atypical. In a way, this would not even be a completely incorrect perception, in the sense that placing this study in only one of the disciplines from which it draws would not be feasible. However, the *a priori* choice was precisely to position itself in an intermediate place between them, but in full communication with them. Here, however, factors other than those concerning the intentions and boundaries of the research intervene, namely the limits of the data. An ambitious challenge has certainly been to maintain a certain balance in the contributions that each research area has made to this study, without letting itself be carried away by the positivist enthusiasm that the use of certain tools, for better or worse, implies. This challenge encounters its own limitations in the analysis section, firstly because these are preliminary analyses, and secondly because the type of data collected in the TArC is indeed data that falls within the sphere of Arabic dialectology (precisely one of the

three fields of study involved), since it deals with texts in Tunisian Arabic, but it is not the prototypical data observed in dialectological analyses. However, we wanted to try to adapt some strategies of dialectological investigation to this type of data, precisely in order to delimit the range of linguistic research that can be carried out on this spontaneous writing system. In particular, we are dealing with analyses on the process of koineisation. This choice is motivated by the intention to orient the reader towards the possible interactions that a scholar working on *dārža* could have with this type of data, such as the use of the TArC as a collection of data that can support, at a quantitative level, certain hypotheses or qualitative analyses. Given the nature of the data that the TArC contains, it is naturally not always possible to exploit its contents. As will be explained in the analyses chapter and in the conclusions, some hypotheses for the interpretation of the data highlighted will be proposed, these hypotheses must therefore be taken in this perspective of preliminary observation aimed at marking not so much a point of arrival as a starting point for future research.

The analyses described in this work are reproducible, as each tool which was produced and employed has been made available on dedicated websites. These tools are the Tunisian Arabish Corpus (TArC), the architecture designed for the TArC multi-level annotation, and the query scripts conceived for our analyses.[1]

Regarding the organisation of the topics in the four chapters, Chapter 1 is dedicated to the historical and linguistic introduction of Tunisian Arabic (Section 1). It will explore the issues inherent to the processes of orthographic standardisation in order to reach the description of spontaneous varieties which have arisen (Section 2). Finally, the same chapter will also introduce the situation of the Tunisian CMC and the issue of *digraphia*, concluding with an introductory description of the main characteristics of the Tunisian Arabizi (Section 1). Chapter 2 deals with the methodology adopted for addressing CMC corpus building using deep learning techniques (Section 1). It will describe the corpus linguistic standards which have been our benchmark in making decisions about the structure of the corpus (Section 3). The same chapter will also introduce some fundamental concepts in order to better appreciate the deep learn-

[1] TArC and the query-tools are available at the following page: _urlhttps://github.com/eligugliotta/tarc. The architecture is available at the following link: `https://gricad-gitlab.univ-grenoble-alpes.fr/dinarelm/tarc-multi-task-system`.

ing techniques adopted in the corpus construction phases (Section 4). It will conclude with the state-of-the-art, in order to expose the path of Tunisian Arabic research in the Natural Language Processing (NLP) field. The state-of-the-art section also aims to highlight the lack of available tools for Tunisian Arabic processing (Section 5). Chapter 3 describes our journey, all the way from data collection (Section 1) to the corpus building phases. There are two main parts to this, an earlier one in which we started annotating the corpus (Section 2), and a second one in which we completed it using the multi-task system which was built for annotation purposes (Section 3). Finally, Chapter 4 will describe the query tools employed for our analyses (Section 1) and the observations we made on the TArC data. Our analyses aim to delineate the nature of the corpus itself and investigate the linguistic reality of Tunisian Arabizi. As briefly hinted at above, the analyses will follow three paths: the quasi-orality path (Section 2), the research on Arabizi spontaneous settling trends (Section 3) and the quest for evidence of the existence of a continuum of formality (Section 4).

# 1. TUNISIAN ARABIC IDEOLOGY

## 1. Historical and Linguistic Introduction

In its most elementary sense, 'linguistic ideology' is what determines how speakers of a given language interpret their language, and consequently how they use it. Linguistic ideology is a matter of beliefs and mental images involving societal concepts and their linguistic representations.[1] Linguistic ideology moves and marks the identity of the speaker, in association or disassociation with certain social groups. In general, it is not easy to describe the linguistic ideology of any culture. Moreover, Tunisian presents a particularly nuanced linguistic ideology, because of its historical, geographical and socio-political features.

In this section we will therefore retrace some of the historical linguistic stages of Tunisian, which will constitute a part of the toolkit necessary to portray the current linguistic complexity thereof. This begins with the historical survey in Section 1 in order to consciously approach the question regarding Tunisian's status as a so-called *Arabic dialect*. We will then move into deeper linguistic matters, and in Section 1, we will instead see how Tunisian Arabic is ranked in terms of traditional dialectological descriptions. Throughout Section 1, we will see how the Tunisian multilingual society has historically and modernly been described by scholars, and what role Tunisian

[1] For further discussion on topics surrounding linguistic ideology, see Woolard (1992).

Arabic assumes in society compared to that of Standard Arabic and French, in an effort to describe all the gradations that exist between the different shades of Tunisian linguistic ideology. In Section 1, we will finally examine the decline in the Tunisian socio-linguistic landscape, which is a fairly common process in the history of languages, namely the emergence of urban varieties with the status of *koiné*. This process is particularly relevant to the Tunisian Computer-Mediated Communication analysis, as we will see in later sections (namely 2).

### *Historical Framework*

In order to comprehend the Tunisian Arabic heritage of human contact, small fragments of both the linguistic mosaic that Tunisian Arabic constitutes today and the Tunisian linguistic ideology, it is necessary to trace back the history of this country, and consequently of its people. We must start by geographically placing Tunisia as a North African country halfway between the western and eastern North African countries. Its borders to the North-East coincide with the coast on the Mediterranean Sea, which is a very important factor in terms of the history of human and linguistic contacts of this country. To the southeast, it borders Libya, while its entire western and southwestern borders are shared with Algeria. Because of its location at the center of the Mediterranean, Tunisia has always been a land of passage, cultural exchanges and a basin of ancient civilisations, such as Carthage (Marçais, 1950).

Carthage was founded in the ninth century BC by the Phoenicians, a population that has long dominated the Mediterranean. The homeland of the Phoenicians corresponds to the current territory of Lebanon, where their ancestors have lived since the third century BC. Carthage rises on the shores of today's Gulf of Tunis. For the rising empire, its geographical position was a strategic crossroads, as it was the obligatory route for all the commercial ships travelling along the route between East and West. Upon arriving in Tunisia, the Phoenicians met the indigenous population, who spoke Lybian-Berber, which is still spoken by a good number of North Africa inhabitants, and of which many proper names of places and people are the only remains.[2] It is unclear whether the Berbers were

[2] Libyan-Berber is linked to the family of Afro-Asiatic languages.[3] The term *barbarus*, deriving from the greek *βάρβαρος*, was used by the Romans to name all non-Latin speaking populations, including Libyans.

the only 'pre-Tunisian' population, and the issue is fascinating, but still obscure (Picard, 1950).[4] The Carthaginian Empire was one of the greatest empires of antiquity, and in its brightest moments, the port city of Carthage was the symbol of its magnificence for all the Mediterranean peoples. For about a millennium, the Carthaginians dominated the Mediterranean coasts, and were known and respected by other peoples for the art of navigation and their aptitude for commerce. The greatest enemy of the Carthaginian Empire was the Roman Empire, and it is precisely because of Romans that not much remains from this great civilisation. The two had quietly coexisted for centuries, until the outbreak of the Punic Wars (264 BC). After Carthage's destruction (146 BC), Tunisia became a Roman province called *Africa*. It was only in 647 AD that the Arabic language made its first appearance in North Africa, adding itself to the already multicultural and multilingual landscape of Tunisia at the time. The Islamic Empire took half a century to conquer Byzantine Carthage (698 AD). To be exact, the first wave of diffusion of the Arabic language was conveyed by military campaigns that laid the foundations for Arabisation, encouraging the rise of the first urban centers, such as al-Qayrawān (670 AD). In these cities, people gradually began to speak the so-called Pre-Hilali Arabic (a theme we will return to in detail in Section 1), but outside the urban centers, people still spoke Berber (Mion, 2016, 2020). These troops settled in what were the great Greek-Roman-Carthaginian cities, also bringing with them the language of religion.

Between 1050 and 1052 AD the Bedouin troops of the Syrians Banū Hilāl and the Arabian Banū Sulaym invaded Tunisia again, in order to quell the Berber revolts against the caliphate's control once and for all.[5] From that moment on, Arabic became the predominant language in Tunisia, relegating all earlier language elements

[4] There is evidence of a Sicilian settlement at Cape Bon and in the region of Kef, where two inscriptions written in an alphabet analogous to Etruscan writing were found near Tunis.

[5] A third group was the Maʕqil from Yemen, as we will see in Section 1. Here we refer to the traditional classification while being aware that this narrative has been subject to recent revisions. In fact, as briefly mentioned in the next section, Benkato (2019) asserts that this classification is the product of a colonial view from which modern dialectology studies should break free. However, we believe it is useful for the purposes of this study to maintain the use of these traditional categories prevalent in modern dialectology textbooks.

(Punic, Latin and Greek) to *substrata* and *adstratum* (Berber).[6] The initial situation of Arabic-Berber bilingualism was soon replaced by the overwhelming predominance of Arabic over Berber. In fact, from that moment onward, the Arabisation of Tunisia underwent a substantial acceleration, penetrating the Tunisian countryside and through the almost total conversion of the Berbers to Islam. Today, the Berber language, namely Tamazight, is considered endangered, and survives only in some areas of Tunisia, particularly in Maṭmāṭa and Djerba (*Žərba*) (Baccouche, 1994; Camps, 1983; Daoud, 2001; Marçais, 1950; Ritt-Benmimoun, 2014; Taine-Cheikh, 2017).[7] Although we will explore these issues in more detail in the following sections, we can already note the coexistence of three 'kinds of Arabic' in Tunisia as the main linguistic strata: pre-Hilali Arabic, basically coinciding with the large urban centers, the Hilali Arabic which instead corresponds more closely to the Bedouin and somewhat rural varieties. In a diglossic continuum[8] with these so-called Arabic dialects (henceforth Neo-Arabic varieties, namely, the last strata of the Modern Arabic dialects (Blau, 1974)), we can find the Standard and Classical Arabic, crossing the whole country as an *acrolect*.[9]

As the last historical stages of this brief *excursus*, we must consider the contact that Tunisia had with European civilisation, starting with Spain, following the exodus of the Arab-Berber Moors from Spain as a result of the Spanish *Reconquista*. In the following centuries, Spaniards and Turks vied for control over trade in the Mediterranean. In the 16th century, Tunisia was annexed to the Ottoman Empire and remained under Turkish domination until the second half of the 19th century, until Italy and France began to compete for supremacy in the North African region in the 20th century. One of the few testimonies of the contact woven into the shared history of these Mediterranean cultures is the *Lingua Franca*, a pidgin language

6 A substratum is a language that has lost its power or prestige to another that has succeeded it, i.e., a superstratum. Both substratum and superstratum languages influence each other. An adstratum, on the other hand, is a language that is in contact with another language in a neighbouring population without having superior or inferior prestige.

7 According to Daoud (2001), the percentage of native speakers is less than 0.5% in Djerba and in some peripheral villages of the governorates of Mednīn and Taṭāwīn. For further information on the Djerba Berber variety, see Brugnatelli (1998).

8 For the definition of *diglossia*, see 1.

9 The most prestigious dialect or variety of a particular language.

of essentially Latin matrix (Bannour, 2000). There is no descriptive model for the linguistic erosion of the languages that merged to originate this language, but we know that the dominant languages were French, Italian and Spanish.[10] The Lingua Franca language was widely spoken, and it was through this language that trade relations (among other things) were brokered with the Mediterranean civilisations. For instance, Italy has a long history of linguistic and human contact with Tunisia because of its maritime activity, which was the basis of both the Italian and Tunisian economies. In addition, Tunis has been a safe harbor for numerous persecuted Jews throughout history. As Cohen (1964) reports, there were Sicilian Jews, Jews who were persecuted and fled the Iberian Peninsula, Neapolitans, and more generally Italians, particularly from Ancona and Livorno. The latter were known as Grana, a prestigious family, whose descendants still live in the Medina of Tunis (Cohen, 1964, 4-6).[11] This large community of Jews has coexisted with the Muslim community throughout history. Above all, starting in the 19th century, a large number of Italian laborers from the south, in particular from Sicily, poured into Tunisia. At the beginning of the 19th century, in fact, the number of Italians resident in Tunisia was at least 11,200 people.[12] In fact, Italy had great economic investments in Tunisia, to such an extent that an Italian company (*Compagnia Rubattino*) owned the rail line Tunis-Goletta. *La Goletta*, later Frenchified as *La Goulette*, or *Ḥalq ə-Wādi* in Tunisian, is a coastal town in Tunisia located 10 km away from the capital Tunis, and is the outer port of same. La Goulette

[10] Instead of Italian, it is more correct to discuss Venetian, Genoese and Florentine, even though historical documents simply refer to Italian. We also know that in Maghreb, to the west of Algiers, the most prominent was Spanish, particularly in Morocco. Other contributions came from Provençal, Portuguese, Arabic and Turkish, as well as from Greek and Armenian. However, in Tunis, there are also Greek and Turkish components. It was a very recognisable language, with a basic morphology, mainly comprising infinitive verbs. For further information see Cifoletti (2004).

[11] ɣ*rāna*, pl. of ɣ*ūrni* from *Leghorn* (Livorno), in opposition to the indigenous Jewish community which were simply referred to as *twānsa*, tunisian. The ɣrana family was also well known because of the market bearing their name: *sūq ə-l-*ɣ*rāna*, between Bāb Swīqa and Ḥafṣīya (Singer, 1984, 472).

[12] Whereas the French settlement was as small as 700 and 7,000 British, mostly Maltese. By 1906, when Tunis had already been a French protectorate for 25 years, Italians still accounted for the majority of foreigners (81,000) vs. the numbers of French (34,000) and Maltese (10,300).

was populated by Italians, and was thus nicknamed *La Petite Sicilie*.[13] As such, many Italian words are now integrated into Tunisian Neo-Arabic, particularly in the lexicon of construction, agriculture, shipping and the arts. However, within the same period, it was another European language that gained importance: the French language. With the collapse of the Ottoman Empire, French was imposed as a second language in light of Tunisia becoming a French protectorate in 1881.[14] At this point, Tunisia found itself living under bilingualism again. The effects of this *Frenchisation* process have lasted until today, with Tunisia experiencing a very particular linguistic situation in which French has become so ingrained in Tunisian culture as to be perceived by some as a threat to the Arab-Muslim identity (Daoud, 2001, 2011). This will be discussed in Section 1. Recently, English has also been added into the mix as the language of technology and of the globalised world. In concluding this section, we would like to quote the words of Mejri et al. (2009), according to whom:

> *'L'actuel dialecte n'est, en fait, que le fruit de différentes strates (substrats, adstrats et superstrats) : un beau syncrétisme qui traduit bien une évolution de la société tunisienne à travers les différentes époques historiques et qui reflète proportionnellement le poids des différentes cultures développées sur ce territoire.'*[15]

After quickly describing Tunisia's multilingual historical landscape and before addressing issues related to Tunisia's modern linguistic configuration, in the next section we will examine the debate among linguists regarding the formation of Arabic dialects and the possible resulting classifications.

---

[13] Just as an example of the cultural link between Italians and Tunisia, we report that La Goulette was the home place of Yolanda Greco, the mother of the Italian actress Claudia Cardinale, who was elected the most beautiful Italian woman in Tunis in the mid 50s.

[14] With the Treaty of Bardo, which was concluded on 12 May 1881. This was followed by the Convention of Marsa of 8 June 1883, which gave France the right to intervene in Tunisia's domestic affairs (Lewis, 2013).

[15] 'The current dialect is, in fact, only the fruit of different *strata* (*substrates*, *adstrates* and *superstrates*): a beautiful syncretism that reflects an evolution of the Tunisian society through the different historical eras and that proportionally reflects the weight of the different cultures developed on this territory.' My own translation.

### *Neo-Arabic Classification*

In the twentieth century, a number of scholars attempted to explain the presence of several linguistic features common to Neo-Arabic varieties with respect to Classical Arabic (or fuṣḥā). Ferguson (1959), for example, proposed the theory of monogenesis, which sees a single point of origin for the contemporary proto-dialect originating from military camps in Iraq, through the mixing of various pre-Islamic dialects spoken by soldiers. Such a proto-dialect, according to the theory of monogenesis, thus coincides with a military *koiné*, in which common features developed. Ferguson lists fourteen features, the first of which, for example, is the disappearance of the dual form in verbs, pronouns, and adjectives.[16] The differences between Neo-Arabic varieties are then explained as the result of a later process of divergence, probably due to the influence of the indigenous languages of the various Arabised regions. Critics of the monogenesis theory have objected to this, stating that the similarities could also be explained as the product of a general trend, or as the result of a subsequent process of convergence that leveled off the dialects in the various areas. They point, for example, to the fact that, like the Neo-Arabic varieties, some languages which are unrelated to Arabic have also lost the dual form. The main problem with the theory of a general trend is that the explanatory power of such a principle is minimal, since the mere fact that similar phenomena occur in different languages does not provide us with an explanation of the causes behind them. Versteegh (1984) proposed the pidgin theory, which saw the Arabic of the early conquests first *creolising* the indigenous languages into a new linguistic form, and then later reversing the trend through subsequent internal diversification producing the Neo-Arabic variation (Durand, 2009). According to Corriente's model, on the other hand, fuṣḥā is itself the end point of an evolutionary path, which is shared by other varieties of Arabic (Corriente, 1976). It thus follows that Neo-Arabic varieties represent a heritage as old as that of Old Arabic. From this perspective, the ancestors of Neo-Arabic varieties did not arise after the waves of Arabisation, but rather existed simultaneously and contributed to the development of Standard Arabic. In the opinion of certain scholars, Classical Arabic is derived from a common ancestor in a proto-Arabic language that has yet to be reconstructed (Embarki, 2008). According to Cohen (1970), Neo-Arabic varieties are the result of a progressive con-

[16] For a definition of *koiné*, see 1.

vergence of dialects spoken in Arab armies, composed of individuals belonging to different tribes. It is his opinion that the differences between the pre-Islamic dialects were leveled *a priori*, while the dialects in the conquered territories evolved subsequently and independently. He also sees this as a reason for the convergence of the pervasive spread of fuṣḥā and the adoption of linguistic innovations produced by cultural or political centers of influence, which speakers of language varieties perceived as languages of prestige. Cohen's theory thus deals with convergence by contact as the end point. However, one issue that cannot be explained through this theory is the presence of common features between varieties of regions of the Arabic-speaking world that have never come into contact with each other (Versteegh, 2014).

The typology that has been adhered to by many scholars classifies Neo-Arabic varieties into six major dialectal areas, from East (*Mashreq*) to West (*Maghreb*):

1. Arabian Peninsula Arabic.
2. Mesopotamian Arabic.
3. Near-Eastern Arabic (Levantine).
4. Egyptian Arabic.
5. North African Arabic (Maghrebi).[17]
6. Sub-Saharan Arabic

However, a number of features and variations within these seem to transcend regional boundaries and effectively escape this typological enterprise. Classification by geographical area is relatively recent compared to other types of classification, such as sociological classification. Indeed, linguists and other observers of Arabic-speaking countries have long since shown that both the smallest locality and the most extensive region are traversed by a division between ʕ*arab* (nomadic) vs. *ḥaḍar* (sedentary) types. As mentioned in the previous section (1), the historical-linguistic tradition narrates two waves of Arabisation in today's Arabic-speaking areas: a first sedentary wave and then a Bedouin wave. This already assumes a first line of separation between sedentary dialect and Bedouin dialects, ʕ*arab*, which are also referred to as *Badawī* dialects (meaning Bedouins). According to the same historical-linguistic tradition the second wave be-

[17] Tunisia belongs to the North-African Arabic class, together with the varieties from Libya, Algeria, Morocco and Mauritania, and the current Maltese language, which was historically a Libyan-Tunisian variety (Mion, 2010).

gan in the middle of the 11th century and was led by three different tribes, including the Banū Hilāl, to whom we owe the appellation 'Hilali' for the Bedouin Neo-Arabic varieties that were formed through this second wave of Arabisation. While the Neo-Arabic varieties derived from the first wave, i.e. those referred to as sedentary, are also referred to as 'pre-Hilali', the distinction between sedentary and Bedouin comes back to Marçais (1961), who not only separates same into the two phases of conquests, but also into two social contexts. For example, the *urban speech* that he defines as being the heir to the old urban *koiné*:

> '*la langue des villes est l'héritière de la vieille koinè citadine remontant à la première conquête*'[18]

and a *rural speech* or *villageois* based on the Bedouin prototype,

> '*la langue des campagnes et des steppes repose sur le prototype qu'apportèrent avec eux les envahisseurs nomades du XIe siècle, Hilāl et Solaïm*'[19].

Regarding the latter, Marçais (1961) also adds that:

> '*La langue qu'on y parlait était et est demeurée un arabe barbare. Il était issu de la déformation de la koinè citadine par des paysans berbères*'[20].

He thus highlights how the prestigious variety of the urban *koiné* was exerting its influence on the rural variety. Among the mentioned sedentary languages, one must also take an additional type of speech into consideration: the speech of Jewish city dwellers (*citadins juifs*), which was, on the whole, a sedentary population's speech, and thus opposed the speech of the Maghrebi Bedouin population (Grand'Henry, 1992). Following the above description, it is possible to outline the linguistic situation as follows:

---

18 'The urban language is the heir of the old urban *koiné* which can be traced back to the first waves of conquest'. My own translation.

19 'The rural language and language of the steppes are based on the prototype brought with them by the Bedouin invaders of the 11th century, Hilāl and Sulaym'. My own translation.

20 'The language spoken there was and has remained a barbaric Arabic. It came from the deformation of the urban *koinè* by Berber peasants'. My own translation.

1. **Pre-Hilali or sedentary, (**ħaḍar**)**, from the 7th century.
   (a) *Urban speeches*
      i.Muslim speech,
      ii.Judaic speech,
      iii.Christian speech.
   (b) *Rural speeches*, of the countryside.
2. **Hilali, Bedouin (**ʕarab**)**, from the middle of the 11th century.
   (a) *Hilali* Arabic, from Syria.
   (b) *Sulaymi* Arabic, from Arabia.
   (c) *Maʕqili* Arabic, from Yemen. (Taine-Cheikh, 2017)[21]

Regarding these classifications, it must be pointed out that it has recently come in for strong criticism. In particular, Benkato (2019) states that all Arabic dialectology to date is strongly influenced by Marçais linguists (William and Philippe), who nevertheless shaped the categories on which the field of research is based. The same author mentions this issue in a later work, where he addresses the description of Maghrebi Arabic specifically, introducing it in the following terms:

> '*In many cases, first-layer varieties* [Pre-Hilali] *of urban centers have been influenced by neighboring second-layer ones* [Hilali]*, leading to new dialects formed on the basis of inter-dialectal contact. It is important to note that North Africa is becoming increasingly urbanized and so not only is the traditional sedentary/nomadic distinction anachronistic (if it was ever completely accurate), but also that intensifying dialect contact accompanying urbanization means that new ways of thinking about Maghrebi dialects are necessary. It is also possible to speak of the recent but ongoing koinéization of multiple local varieties into supralocal or even roughly national varieties–thus one can speak, in a general way, of "Libyan Arabic" or "Moroccan Arabic"*' (Benkato, 2020, 198).

These assertions found approval from a number of experts in the field, who had previously expressed their doubts about the adaptability of the traditional classification to Tunisian varieties, and in particular the concept of village dialects. Among these researcher, Mion (2015) defines village languages as varieties whose mixed nature is caused by contact with Bedouin Arabic, from which they have taken on their fundamental characteristics. Also D'Anna (2020) stated that the category of village dialects should be contextualised

[21] As previously stated, we refer here to the traditional classification while being aware that this narrative has recently been challenged (see Benkato (2019)).

| | *OA features* | **sedentary v.** | **Bedouin v.** |
|---|---|---|---|
| 1. | */q/ | [-voiced] | [+voiced] |
| 2. | */θ ð ð̣/ | /t d ḍ/ | /θ ð ð̣/ |
| | | (*/ḍ ð̣/ ⟶ /ḍ/) | (*/ḍ ð̣/ ⟶ /ð̣/) |
| 3. | */ǧ/ [ʤ] | /ž/ [ʒ] | /ǧ/ |

*Table 1. Distinctive phonological features of sedentary and Bedouin varieties.*

in a language contact framework that takes into account the importance of local history.[22]

However, as far as this classification into sedentary and Bedouin varieties by Marçais linguists is concerned, not only does it remain the most widespread and established methodology for comparing Arabic varieties, but there is no alternative so far. In particular, this methodology aims to observe the preservation, neutralisation, mutations or innovations of some Old Arabic (OA) traits in Neo-Arabic varieties. Below are proposed tables that simplify the traditional axioms. The outline refers to Mion (2010), who presents classification criteria for Bedouin or sedentary Neo-Arabic varieties based on phonological, morpho-syntactic and lexical features. Regarding the lexical characteristics, since they are not central to this work, it is suggested to consult the studies of Mion (2010) and Marçais (1950).

With regard to the commonly accepted distinctive phonological features presented in Table 1, the first point concerns the phonetic realisation of the uvular plosive phoneme /q/, which is a voiceless /q/ in the sedentary varieties and is voiced /g/ in the Bedouin varieties. Regarding the second point, this is not the exact case in the Tunis *koiné* (Tūnsi), which is generally defined as a sedentary variety, but where the interdental phonemes are maintained (as in Bedouin varieties. This is in contrast to sedentary varieties (mostly for extrapeninsular Arabic), in which the fricatives /θ ð/ are realised as plosives /t d/ with the confluence of */ð̣/ in /ḍ/. Instead, in the Bedouin varieties, dental fricatives are generally maintained with the confluence of */ḍ/ in /ð̣/. Regarding the last point, the phoneme that presents an affricate realisation ([ʤ]) in OA seems to be conserved in Bedouin Neo-Arabic varieties along with the fricative realisation

[22] For further information about this new line of research see also Guerrero (2018) and Mion et al. (2014).

| | *OA features* | **sedentary v.** | **Bedouin v.** |
|---|---|---|---|
| 1. | 2pl. ≠ 3pl. (verbs, pron.) | 2pl. = 3pl. | 2pl. ≠ 3pl. |
| 2. | 3sg. pron. suff. */-hu/ | /-u/ | /-ah ∼ -ih/ |
| 3. | verbal time-aspect | innovative | conservative |
| 4. | passive verb (ablaut in CA) | innovative | conservative |
| 5. | dual forms | less maintained | more maintained |
| 6. | genitive phrase | analytical | synthetical |
| 7. | agreement with pl. heads[-human] | not maintained | maintained |
| 8. | gender and number agreement | not complete | complete |

*Table 2. Distinctive morphological features of sedentary and Bedouin varieties.*

([ʒ]). However, the latter appears to be the main realisation in sedentary varieties (Durand, 2007, 2009; Mion, 2010, 2018).

In contrast, an outline of the main distinctive characteristics of the morphological features employed to classify Neo-Arabic varieties in the Bedouin or sedentary type are reported in Table 2. The first point concerns the gender distinction in the second- and third-person plural of verbs and pronouns, which is maintained only in the Bedouin varieties. The second point concerns the different realisation of the suffix for the third-person singular pronoun; while the third point deals with the axiom that the verbal time-aspect systems display different behaviours between the sedentary and Bedouin varieties, wherein the sedentary varieties employ a set of preverbs and innovative strategies for expressing verbal time and aspect (Mion, 2004). In addition, the fourth point of the table concerns the difference in the verbal systems, and once again, it is the sedentary varieties, unlike the Bedouin ones, which exhibit an innovative system of passive verb realisation, which consists of employing prefixes instead of applying the vowel alternation recognised by the OA system (also known as ablaut or apophony). Both the dual form of nouns (the fifth point) and genitive phrases (the sixth point) illustrate the tendency of the Bedouin varieties to employ strategies which are also present in OA. Conversely, in the sedentary varieties, dual forms are mainly used in double body parts and dual forms which are already fixed in the lexicon, such as the Tunisian /ʔalfīn/, which has the literal meaning of 'two thousand', but which is used to express the amount of 'two Tunisian Dinars'. Regarding the genitive phrase, the sedentary varieties employ both the so-called construct state (the synthetical structure) and the pseudo-prepositional

system (the analytical solution). The pseudo-preposition employed for the latter in Tūnsi is /mtāʕ/ (where the etymology is */matāʕ/ 'property'). The last two points (the seventh and eighth points) concern the agreement, the first dealing with non-human heads, while the second deals with number and gender agreement. In the first case, the solution referred to in OA (of female singular agreement for non-human heads) is also employed only by Bedouin varieties. In the matter of the second agreement, the inflexions seem to be complete for Bedouin varieties, but are incomplete for sedentary varieties, with the constituent's segment showing $N^{pl.f}$+$ADJ^{pl.m}$ (Durand, 2009; Mion, 2010; Singer, 1984).

As previously discussed, dialect classification features do not produce sharp distinctions, either between regions or between sociological varieties. Few differences emerge within the same dialectal area, and regional marking is not very evident. As an example, let us take into account the study of Ritt-Benmimoun (2014) into the differences between the Tunisian Neo-Arabic which is traceable to the H type (which is Banū Hilāl Arabic), in comparison to the S type (Sulaymi Arabic), for example, the Kasserine (*Al-Qaṣrīn*) variety. She describes the affinity between the H and the sedentary varieties and cites the explanation given by Marçais (1957), i.e. that in regions where the H dialect is prevalent, the sedentary dialect had originally been spoken, but was subsequently overlaid by Bedouin (Sulaymi and Hilali) dialects, resulting in the Bedouin and sedentary varieties becoming merging. As Ritt-Benmimoun (2014) states, there are in fact transitional regions. As a result, no sociological *koiné* seems to have expanded across the entire Arabic-speaking area. However, legitimised by linguistic prestige, these urban areas were developing linguistic dynamics characterised by processes of homogenisation or differentiation and standardisation. It is therefore questionable as to whether a phonological trait can resist the centrifugal force of urban centers through the processes they mobilise, such as linguistic accommodation and dialectal levelling, and the strong pressure of the prestigious character of their local language (Mion, 2010). We will address this issue in terms of Tunisian Neo-Arabic in Section 1.

The next section will focus on an analysis of the current reality of the Tunisian linguistic situation, starting from the relationship between Arabic, as the official language of Tunisia, and Tunisian Neo-Arabic, as the Tunisians' mother tongue.

### *Diglossia & Multilingualism*

The *ante litteram* sociologist Ibn Khaldūn (*Tūnis* 1332 – *al-Qāhira* 1406) is well known for *The introduction* (*muqaddima*, 1377) to his universal history essay: kitābu l-ʕibar (lit. 'book of examples').[23] This introduction is an essay on social history, in which Ibn Khaldūn discusses the origins and development of civilisation, taking the dichotomy between Bedouin and sedentary life as a starting point. According to his view of history, civilisation gradually developed from a Bedouin way of life towards a sedentary one. Furthermore, Ibn Khaldūn includes, in his outline of the development of civilisation, a discussion regarding the origins of the discipline of linguistics. He says that the Bedouins spoke Arabic according to their 'natural disposition' and did not need grammarians to tell them how to speak. However, in sedentary civilisation things have changed: decadence was embraced and the language threatened to become *corrupted*. In Ibn Khaldūn's description, this process of corruption was connected to the 'invention' of grammar (Campanini, 2019; Embarki, 2008). Ibn Khaldūn's views on the historical development of the Arabic language are an important testament to how the Arabs of the time viewed the history of their language and how this is reflected in today's Arabic linguistic ideology. The history of the Arabic language begins in the period prior to Islam (the 5th-6$^{th}$ centuries), commonly referred to as *Ǧāhiliyya* 'the period of ignorance (of God)', when the Bedouins had not yet received the message of Islam.

According to the view of Ibn Khaldūn, the *pure* language of the Bedouins remained unchanged until the Arabs came into contact with other peoples during the period of the conquests. Despite some known differences among the spoken language varieties, these differences did not break down the essential unity of the language. The Arabic language was used in interactions between the Arabs and the conquered peoples, but because of the difficulty of the language and due to these people's mistakes when learning the complicated linguistic structure, the Arabic language began to be *corrupted*. The central theme of Ibn Khaldūn's historical reconstruction is this corruption of the language. This opposition between classical grammatical

[23] The full title is: *Kitābu l-ʕibari wa dīwāni l-mubtadaʔ wa-l-ḫabar fī ayyāmi l-ʕArab wa-l-ʕAǧam wa-l-Barbar*, meaning: 'Book of Lessons, Record of Beginnings and Events in the History of the Arabs and Non-Arabs and Berbers'.

language and ungrammatical colloquial variants is still present in the Arabic world today.

Classical Arabic was the same language that was later recorded in writing by medieval grammarians, following two centuries of oral transmission of the poetic texts, as were the sacred texts. However, this linguistic form, which was used for poetic-literary purposes in the 6th and 7th centuries in central Arabia, could represent the *crystallisation* of a precise linguistic stage in the evolution of the language originally spoken by the Bedouin tribes mentioned above (Durand, 2007; Mion, 2016). Focusing more on poetic-literary and sacred texts, in order to record their norms, grammarians left a large part of the linguistic reality, which was most likely already configured to be diglossic.

The word *diglossia*, from the Greek διγλωσσία, describes a sociolinguistic situation in which two main forms of a language coexist, in which a *high* form of language (H) is perceived as older, more prestigious, purer, more beautiful, and perhaps more appropriate for *high* educational contexts, literature, religion, and so forth. The *low* language (L) has no prestige and is devalued and even despised, even though it is the actual *mother tongue* of the population, having been learned in the first years of life and being used by all members for colloquial purposes (at home, on the streets, in bars, and for humour). However, summarising the multiple complexities of the Tunisian linguistic reality with the word diglossia is entirely too reductive, as can be seen over the course of this chapter.

### *Ferguson's diglossic model*

Ferguson (1959) is traditionally considered to be the academic father of the conceptualisation of diglossia, and his article represents a crucial step in explaining the functional coexistence and linguistic interrelationships between these varieties, a step that probably marks the beginning of Arabic socio-linguistics as an academic entity in its own right. In the article, Ferguson gave theoretical body and comparative generalisation to a situation that had already been recognised, for example by Psicharis (1888), who was the first to use the term diglossia in reference to the Greek linguistic situation. At the same time, the diglossic reality for the Arab world has been described by French linguists working in North Africa, first of all by William Marçais (1930), who described the coexistence of a *langue littéraire* that coincides with *l'arabe écrit* and a variety of *arabe parlé* in Alge-

ria (Benmamoun and Bassiouney, 2017). The notion of diglossia was later refined by Ferguson himself and many other scholars. In fact, Ferguson (1959) opened a Pandora's box full of extensive literature on the subject, which attests to the manifestation of diglossic situations and their widespread reach throughout languages across the world.[24] However, what has always been highlighted as a weak point of the Fergusonian model is a strict and fixed division that abstracts languages into linguistic models which are too *rigid* to coincide with reality. This has caused a move towards the search for further parameters to describe the contact and interactions between languages, such as the language continuum or the code-switching theories that will later be addressed here (Bassiouney, 2009; Lancioni, 2018). According to the ideal Fergusonian model, diglossia describes a stable socio-linguistic situation in which, in addition to a low variety *L* (coinciding with the colloquial variety of a language) which may include a standard or a regional standard, there is a high variety *H*, which is highly codified (and often grammatically more complex), *complementary* to the former, and acting as the medium for a vast and respected body of literature, from a previous period. This H variety, according to Ferguson, is learned largely through formal education and is used for most written and formal speech, but is not used in any sector of the community for ordinary conversation. Ferguson (1959) identifies a set of nine parameters which can be used to describe a typical diglossic situation:

1. **Function.** This concerns the rarely overlapping uses of the H and L variety. The respective written and oral languages serve different functions, where the former is preferred for formal contexts and the latter for informal contexts.[25]
2. **Prestige.** This refers to the superior status of the H variety over the L variety as perceived by the speech community.
3. **Literary heritage.** This describes the coincidence of the literary language type with the H variety.
4. **Acquisition.** This implicitly refers to the L variety being the native language of the speech community, describing the acquisition of the H variety through formal education.
5. **Standardization.** A process concluded only for the H variety.

[24] In India for example (Fishman, 1999).

[25] Later, Ferguson (1996) himself will explain that by 'variety' he means varieties of the same language and not varieties in terms of 'socio-linguistic variables' such as variations which are diatopic, diastratic, diamesic, etc.

6. **Stability.** Referring to to the diglossic situation as being stable over time.
7. **Grammar.** Alludes to structural divergences between the H and L varieties.
8. **Lexicon.** The two varieties share a certain amount of vocabulary and semantic content.
9. **Phonology.** Ferguson (1959) describes the divergence as follows:

> '[T]*he sound systems of H and L constitute a single phonological structure, of which the L phonology is the basic system, and the divergent features of the H phonology are either a subsystem or a parasystem* [...] *if pure H items have phonemes not found in pure L items, L phonemes are frequently substituted for these in oral use of H and regularly replace them in tatsamas*.'[26]

The appeal of Ferguson's model is that it provides an explanation for the maintenance of more or less structurally divergent forms in the same speech community. It is noteworthy that Ferguson's model incorporates spoken language, thus marking a departure from the Arabic philological tradition which was primarily oriented toward the interpretation of written texts. Notably, Ferguson not only described contemporary diglossia, but also projected its origins back to the time of the early seventh-century Arab diaspora. In addition, (Ferguson, 1997) argued that modern dialects developed from a *koiné* born in the military camps of the Arabs. This contact produced a certain simplification and levelling among the dialects, with a variety similar to Old Arabic continuing to be spoken, at least for a time, among Bedouins. This aspect will be returned to in more detail within Section 1.

As proposed by Badawi (1973) and then demonstrated by further studies (Bassiouney, 2009; Brustad, 2000), the Arabic language can be seen as a *continuum*, ranging from Classical Arabic (*fuṣḥā*) on the one hand as the highest and most formal linguistic variety, to *ʕāmmiyya*, or colloquial Arabic (the Neo-Arabic varieties also known as *Neuarabisch*), on the other. An intermediate variety known as Modern Standard Arabic (MSA), falls between the two forms, and is used mostly in news broadcasts and formal interactions, but is rarely spoken elsewhere and is not learned by anyone as a native language. The Egyptian linguist Badawī points out that the boundaries between these categories are fluid, and native Arabic speakers

[26] *Tatsamas* is a term, which is used by Ferguson, that means 'the same' and refers to the use of classicalisms appropriate to the phonological system of H in L varieties.

are able to adapt their language forms according to context and their communication needs.

### *Badawī's diglossic model*

Badawi (1973) was the first to propose the vision of a continuum of linguistic variation, proposing a socio-linguistic representation of the situation in the urban area of al-Qāhira. His work built on the idea of the high and low varieties of Ferguson (1959) by adding more levels, firstly identifying two levels by splitting the H variety into Classical Arabic (CA) and Modern Standard Arabic (MSA), and then by identifying three gradually less formal varieties, as shown in the list reflecting his pyramid model.

1. At the top, the two levels of *fuṣḥā*:
   (a) **at-turāθ**, the Classical Arabic of the pre-Islamic literary tradition (classical heritage).
   (b) **al-ʕaṣr**, the contemporary Arabic fuṣḥā better known as Modern Standard Arabic (MSA).
2. At the positions below, three levels of *ʕāmmiyyat*:
   (a) **al-muθaqqafīn**, which was spoken by highly educated native Arabs, comparable to what Mitchell (1986) defines as Educated Spoken Arabic (ESA), which will be further addressed later (the colloquial variety of the cultivated people).
   (b) **al-mutanawwirīn**, which was used by people who have received a generic education (the colloquial language of the basically educated).
   (c) **al-ʔummiyyīn**, the one spoken by people who have not received any kind of linguistic education (the colloquial language of the illiterates).

The crucial aspect of Badawī's model is the fact that his subdivision implies both a stylistic (regarding register) and socio-economic (education) hierarchy, which ends up being a more diastratic than diaphasic subdivision (Bassiouney, 2009). Specifically, as Durand (2009) explains, it is very difficult to consider that uneducated people can use the fuṣḥā linguistic registers according to the context needs, whereas an educated person can be expected to have a certain freedom in adapting their register to formal and informal situations.[27] As a result, this is a pyramidal model, especially when taking

[27] Durand (2009) also adds that this model, which is basically valid for the Mashreq countries, is not perfectly applicable to the Maghreb, not least

into consideration the amount of speakers belonging to each language class. However, the linguistic situation of Arab countries can also be described using a monosystemic model such as that of Hary (1996) discussed below.

*Mitchell*

In the mid-1970s, an alternative view of the study of linguistic variations in spoken Arabic emerged. Mitchell (1978) developed the concept of Educated Spoken Arabic (ESA), which lies on the continuum between the two poles of the acrolects (Standard Arabic and Literary Arabic) and the basilect (Vernacular Arabic). This ESA linguistic variety corresponds to Badawī's ʕāmmiyyat al-mutanawwirīn/muθaqqafīn being a form of spoken Arabic, which was only used by educated people. Educated Spoken Arabic has also been described by Meiseles (1980) as a national vernacular type *characterised by the aspirations of its speakers to get rid of local features through a process of koineisation and/or borrowings from literary Arabic*. A type of al-luɣa al-wusṭā, in this case meaning 'middle Arabic', is the informal variety used among educated Arabs and for inter-dialectal communication (Durand, 2009). Indeed, it consists of a hybrid form of standard and colloquial Arabic elements. However, according to Durand, Arabic is never 'pure', but is instead constantly subjected to mediation between the two poles, driven by the requirements for interaction. It is an intermediate style that varies from speaker to speaker, even according to their competence in Standard Arabic (Bassiouney, 2009). The main difference between the ESA described by Meiseles (1980) and that described by Mitchell (1986) is that the latter is not more committed to the idea of discrete levels of Arabic(s). In fact, Mitchell (1978) explains that a certain amount of overlap may occur between ESA and MSA or ESA and vernaculars, as for example in the SVO order, where in the second case, he describes the tendency to level two vernacular forms for the purpose of effective communication or to respect the formal stylistic code.[28]

---

because it does not take into account the important role that the French language plays in cultural and intellectual spheres.

[28] For studies on the Maghrebi ESA, see Benmayouf (2003) regarding Algeria, Taine-Cheikh (1978) regarding Mauritania and Youssi (1986) for the Moroccan situation. For an analysis of the markedness of SVO-VSO orders in MSA, see Lancioni (1996).

### *Hary*

Hary (1996) defines *multiglossia* as the linguistic situation in which different varieties coexist side by side in a language community and where each one is used in different circumstances and with different functions. Hary further describes the arrangement of such varieties as being spread along a context-dependent continuum where speakers (and writers) continually move from one system to another. Hary's proposal thus consists of a monosystemic model (rather than a dichotomous one with a sharp division between the H variety and the L variety), where the two ends of the language continuum coincide with the standard *acrolect* variety and the colloquial *basilect* variety. Along the continuum, between the two extremes, we can also find the *mesolect*, where innumerable varieties can be placed, which are used by native speakers in different circumstantial situations. Moreover, the native speakers themselves can make different combinations among the varieties, sometimes unconsciously, which is why the system outlined by Hary appears highly dynamic. Its dynamism is determined by a number of socio-linguistic and linguistic variables:

1. **Setting.** Formality or informality of the communication context.
2. **Topic.** The topic may require more or less seriousness.
3. **Speakers' skill in MSA.** Speaker competence in MSA is a necessary condition for the dynamism of discourse across different levels.
4. **Emotional state of the speakers.** According to Hary, a particularly emotional speaker (nervous, angry, etc.) tends to move toward the colloquial pole.
5. **Participants in the discussion.** The expertise of other participants in MSA is also a prerequisite for dynamism.
6. **Function of the discourse.** The function of speech is also the basis of this model, as it was for previous diglossic models.
7. **Personal relationship with the audience.** This is a component underlying all speech acts, i.e. the building of relationships with interlocutors, and the building of communicative solidarity between speakers.

It is also necessary to consider the continuous mutability of linguistic systems, such as the historical and political, economic and social transformations that the linguistic landscape of the Arab world has undergone in the last three decades. Among these trans-

formations, the advent of technology and the birth of Computer-Mediated Communication (CMC) has added further nuances to the already colorful picture, as seen in Section 2.

*The relationship between bilingualism and the concept of diglossia*

Regarding the relationship between diglossia and bilingualism, Fishman (1967) offers an overview of linguistic realities which feature both concepts, as in Tunisia, reorganising the two concepts. According to Fishman, the latter refers to the use of more than one language in a society, while the former refers to the functional distribution among them. Therefore, it is possible to have bilingual societies in which the two languages are used for parallel functions (without diglossia), as well as those in which different languages correspond to different communicative functions (this is the case, for example, of the Yiddish society in Germany). Conversely, there are diglossic communities that are also bilingual. Many of the communicative and symbolic functions performed by the L and H diglossic varieties may be performed by dialects and the standard language or by other languages unrelated to the H and L varieties. Mejdell (2017) makes an important remark about the different point of observation of the linguistic sciences by noticing that from a purely sociological or ethnographic perspective, the specificity of diglossia may not be relevant, whereas for a more linguistically oriented analysis of language contact, variation, and change, the linguistic analysis of diglossia that *opposes* two *related* and *highly divergent* varieties becomes central in order to study their differences. More precisely, what these distinctions consist of and how these develop alongside societal changes must be taken as the central theme.

*Switching theory instead of diglossic continuum*

An alternative view to that of the discrete levels or the diglossic continuum individuation began to be investigated just before the 1990s through contact linguistics. In particular, the code-switching theory, the most widely used model of which is known as the *Matrix Language Frame* (MLF), was developed by Myers-Scotton (1993) and later updated to the *4-M* model by Myers-Scotton and Jake (2000), as is explored in Section 2. The distinction between code-switching and code-mixing is far from clear. The most common understanding of these concepts is that code-mixing is the alternation of languages at a lower level of the sentence (or clause) and code-

switching is at a higher level, and the former is mainly attributed to grammatical and structural reasons, while the latter is accredited with a greater discursive and communicative importance. The term code-switching is used by Myers-Scotton (1993) to refer to alternations between linguistic varieties within the same conversation. According to the MLF model, every conversation between bilingual speakers presents a Matrix Language (ML) within which the elements of an Embedded Language (EL) are inserted. An ML can be identified on the basis of indicators of a predominantly morpho-syntactic nature: the type of syntactic structures used, the relative quantity and frequency of morphemes of one or the other language present in the language segment analysed, the order of the constituents, the type of morphemes of one or the other language used by the speakers, etc. Eid (1982, 1988) was the first to apply the code-switching model to the Egyptian diglossic situation. Regarding Tunisian Neo-Arabic, Attia (1966) identifies code-switching between MSA and Tunisian Neo-Arabic, including morphological and lexical features, such as the Tunisian relative marker and the preference of the Tunisian verb /nəžžəm/, 'I can' instead of the MSA /ʔastaṭiʕu/. Moreover, Boussofara-Omar (2006) argues that what is defined as *middle Arabic* by Hary (1989), *ESA* by Mitchell (1978) and Meiseles (1980) or *ʕāmmiyya l-muθaqqafīn* by Badawi (1973), is neither a level nor a section of the continuum as described by Hary (1989, 1996), but is instead the result of a switch between MSA and Tunisian Neo-Arabic, making diglossia a *particular case of language contact*. Tunisian-French code-switching has also been addressed for example by Baccouche (1994); Riahi (1970); Sayahi (2011), the first of which confirmed the extensive use, by Tunisian *educated* speakers, of what Baccouche (1994) later came to define as *franco-arabe*, as can be seen in the next section. Regarding Sayahi (2011), same confirms that Tunisian speakers with a higher education (university-level education) exhibited a much higher frequency of code-switching that reflects their higher degree of competence in the French language.

### *The Tunisian multilingual situation*

According to the Tunisian scholar Daoud (2011), Tunisian native speakers exhibit the use of five different levels of variables:[29]

[29] Here we report the terminology adopted by Daoud (2011). However, as noted by professor Francesco Grande, whom we thank for his observation, the taxonomy reported in this paragraph, presents an ambiguous reference

1. Classical Arabic (CA) + *written.*
2. Literary Arabic (LA)
3. Modern Standard Arabic (MSA)
4. Educated Arabic (EA)
5. Tunisian Arabic (TA) + *spoken.* This level includes all diatopic variations of TA.

Daoud (2011) thus organises the variables into levels along a vertical continuum, in a manner similar to Badawi (1973), asserting that the literature on the concept of diglossia diverges from the general perception in the Arab world of the Arabic language as a unicum, and that linguistic labels are only useful as descriptive terms of the function that particular variables play in society. He continues to explain that CA, the highest variety, coincides with the sacred texts of the *Qurʔān* and generally with *Ǧāhiliyya* period texts. In Tunisia, the diffusion of Classical Arabic has been sustained until today through a national network of lectures held in mosques (*kuttāb*), which continue to offer a valid, officially recognised alternative to kindergarten. CA is still maintained and used in the following religious practices today:

1. The reading or recitation of the Qurʔān, which is generally included in daily religious practices;
2. Prayer in the mosques, run by the Imams, who along with scholars of the religion (ʕulamāʔ) continue to be trained in the best classical tradition at the University of Zaytouna (*Žā*miʕat əz-*Zītūna*), which is a fully-fledged component of the higher education system of modern Tunisia;
3. The spread of religious books (mainly the exegesis of the Qurʔān and the Prophet's teachings).

Literary Arabic (LA) initially appears in the Arabic production of literary, rhetorical, philosophical and scientific works by Arab-Muslim authors of the early Islamic period until the fall of the Abbasid caliphate (and the beginning of the decline of Arabic-Islamic civilisation). This variety, which is described as a *manifestation of the CA of great works* by Daoud (2011), was later taken up by authors of the 19th and 20th centuries. It is still taught in Tunisian schools as part of the Arabic language and literature curriculum, but remains restricted to literary and academic circles.

---

to the Qur'anic Arabic and the later written literary usage, labelled Classical and Literary Arabic, respectively.

MSA is an intermediate functional register that has evolved into a lively written and oral variety. This is the language of the mass media, international politics, as well as modern plays, novels, literary magazines, and lectures. Among the 'high' varieties described by Daoud (2011), MSA is the first to have dynamic characteristics which have allowed it to evolve through the incorporation of terms which are necessary for the description of modern reality, such as scientific, technological and commercial vocabulary. As for the inclusion of these terms in MSA, sometimes they are loans from French or English, but they may also be neologisms based on the reworking of Arabic material.[30]

EA is the last intermediate variety before colloquial Tunisian Arabic (TA), with respect to which it ranks as a higher variant as it is characterised by its borrowings from CA, LA, and MSA. Daoud (2011) describes it as a strictly oral form, used by educated Tunisians in formal and semi-formal contexts such as meetings or negotiations, and in conversations involving *learned* topics and educated interlocutors, perhaps from other Arab countries.[31] Daoud also challenges the definition of EA by Walters (1996) as a *sociolect*. Walters defines EA as *a new social (i.e. class) dialect*, and adds:

> '*as children who live in urban communities, composed of Tunisian couples in which both wife and husband are well educated, grow up acquiring it as their native variety.*'

In contrast, when referring to TA, Walters says that it:

> '*indexes home and hearth, daily life, everyday logic and lived experience, and common sense.*'

However, according to Daoud, EA does not draw on one's social background, but only on one's level of education. However, an objection to this has arisen, namely that the level of education constitutes one (if not *the*) discriminating factor underlying social inequalities in a country that still has, to date, a wide socio-economic divide. In fact, Walters (1996) himself concludes his explanation of EA (which he refers to as ETA from Educated Tunisian Arabic) by saying that it is an intermediate variant between L and H, leaning toward the L pole of the diglossic continuum, and that it constitutes

30 With regards to the strategies applied in order to *update* MSA and in the meantime preserve its *purity*, see Mion (2016).

31 As we will see below, it is in fact not so strictly oral.

an emerging variety, which is connected to the social changes that have been set in motion since independence.

Regarding TA, i.e. the low variety, Daoud (2011) defines it as *the oral dialect of Tunisia*. He includes both local and regional varieties, characterised by phonological and lexical variations within the country, in the same level, in consideration of the fact that, despite the diatopic variation, TA remains distinguishable from the dialects of neighbouring countries, such as Algerian and Libyan, most likely due to the cumulative effect of contact with other languages since antiquity.

Daoud also emphasises the oral nature of this variety of Tunisian Arabic, specifying that there is a small archive written in Tunisian Arabic, but that it consists mainly of folk tales, collections of proverbs and poems, occasional contributions from the press, and some novels, such as Antoine de Saint-Exupéry's novel: *Le Petit Prince*, which was translated in 1997 by Hédi Balegh, a Tunisian intellectual known for his militancy in favor of recognition of الدارجة, *ə-dārža*, the North-African name for the *ʕāmmiyya* (Mion, 2007).

The first article of the Tunisian Constitution (approved on 26th January 2014) states:

> *'La Tunisie est un état libre, indépendant et souverain ; sa religion est l'Islam, sa langue l'arabe et son régime, la République.'*[32]

However, it is useful for the purposes of this study to keep in mind the distinction between:

1. The *national* TA, which is oriented towards the urban *koiné* of Muslim Tunis.
2. The *local* varieties of TA.

In fact, as later explored in this chapter (Section 1), there have been occasional calls to make national TA the official language of Tunisia, but these have never been taken seriously. Nevertheless, 'defenders of the purity of CA' have always perceived TA as a permanent threat to the high variety of Arabic because of its vitality. As Daoud mentioned back in 2011, when CMC was still in its infancy in Tunisia, the fears of the 'defenders of the purity of CA' were further exacerbated due to the use of TA for chat and text message communications, which did not even use the Arabic alphabet, but what Daoud

[32] 'Tunisia is a free, independent and sovereign state; its religion is Islam, its language is Arabic and its system of government is the Republic'. My own translation.

refers to as *the French alphabet*, explaining that the younger Tunisian generations 'resort to numbers to represent the distinctive (guttural) Arabic sounds'. This reference is to the graphemic system which is here referred to as *Arabizi*. However, before addressing the issue of *digraphia* (in Section 2) in the context of Tunisian social networks, it is necessary to first describe Tunisian bilingualism.

The systematic diffusion of the French language through Tunisia began with the protectorate (1881). However, the independence obtained by Tunisia in 1956 has done nothing to reduce the role of this language in Tunisian society. In fact, according to Baccouche (1994), its prestige has increased in the eyes of large proportion of the population for several reasons: first of all, the liberation from the 'colony complex', which is a source of hostility for the language of the oppressor; moreover, among the youngest generations, who have never known linguistic opposition of a nationalist nature, French has become a source of political and socio-economic privileges and a means of social progress. In fact, French extends across the formal spheres of education, business and administration to the informal spheres of everyday oral communication. Daoud (2011) analyses the French language in Tunisia in terms of diglossia, separating metropolitan French (the high variety of French) from the North African varieties of French (the low variety). The former, according to Daoud, is used by Tunisians who have had contact with the French language from France, having spent a period there for study or family reasons, or having attended *les écoles de la mission culturelle française* (French schools). These are opportunities for contact which are open only to a minor part of Tunisian society, since they require economic resources that most of the population does not have. Among the members of this elite with access to H-level French, as reported by Daoud (2011), it is not unusual to find Tunisians who are poorly proficient in MSA. French has thus become the language of high society, intellectuals, and senior executives. Speaking French, especially good French, is also a claim to prestige, an act of identity affirmation and a declaration of belonging to the Tunisian elite. This act may be the imitation of French pronunciation or the French accent, or in the 'French-like' pronunciation of the Tunisian itself, typically in urban society, particularly in the capital, and even more so in the *female genderlect*. Regarding the L-level French, Daoud (2011) expresses the idea that it is only *North African French* within which it is possible to find different shades of usage for different communicative functions. This seems to coincide with the variety that Baccouche (1994) defines as *le franco-arabe*, which, according to him,

consists of a much more complex *mélange d'arabe dialectal tunisien et de français*. In fact, this variety goes beyond the level of lexical interference to encompass syntactic interference in a complicated and curious linguistic tangle.

What Baccouche (1994) claims is 'paradoxical' is that such a *mélange* came about after independence. However, as illustrated above, there were other social motivations that explain the spread of French in Tunisia once freed from the colony complex. Another important point that Baccouche (1994) makes is as follows:

> '*Historiquement, un tel registre est appelé à disparaître tôt ou tard, une fois l'enseignement et l'administration arabisés.*'[33]

During the seventy-five years that Tunisia was under the French protectorate (1881-1956), the Arabisation of teaching and administration and public life was claimed unanimously and numerous times by Tunisian political parties and trade unions. For example, national legislation has not regulated the language of signage. Only an order given by the municipality of Tunis (August 6, 1957) obliges the owners of public establishments to arabise their names. The first article of this ordinance states (Ghoul, 2009):

> '*Toutes les enseignes commerciales, industrielles ou autres qui donnent sur la voie publique doivent être rédigées en langue arabe. Elles peuvent cependant être bilingues.*'[34]

Regarding teaching, in 1958, the *Educational Reform Law* was signed by the first Tunisian President: Ḥabīb Bourguiba (in office from 1956 to 1987). This educational reform, however, failed for structural and political reasons, which were analysed in detail by Daoud (2001). He offers a detailed analysis of Tunisian language planning starting immediate in the post-independence period. With regard to the 1958 reform, this was aimed at unifying the various school systems which were present in Tunisia (recalling the widespread role of mosques in dispensing first grade instruction) under a single bilingual system managed by the Ministry of National Education and offered free of charge. The reform also envisaged the adoption of the organisation and national curriculum of the French

[33] 'Historically, such a register is due to disappear sooner or later, once education and administration have been arabised.' My own translation.

[34] 'All commercial, industrial or other signboards that overlook the public highway must be written in Arabic. However, they may be bilingual.' My own translation.

system, complete with the *baccalauréat* as the final exam. An important point of the reform was the reinstatement of Arabic (mainly MSA) as the language of instruction, with the exception of elementary schools. However, in secondary schools and universities, the French language was relegated to a vehicle for teaching scientific and economic subjects. This reform marked the beginning of a campaign of Arabisation, which was implemented even more vigorously and systematically in 1976, but which did not bring about the desired results.[35] The biggest failure was the results of the 1986 *baccalauréat* exams, which President Bourguiba ascribed to students' weak proficiency in French, implying that Arabisation had been pushed too fast, or too far. Laroussi (2010) states that in Morocco, as well as in Algeria and Tunisia, Arabisation was presented as a legitimate operation with the aim of providing MSA with the necessary tools to assert its ability to convey all sorts of knowledge and to ensure equal opportunities to the Maghreb populations. However, in reality, according to Laroussi, the objective behind Arabisation was to eradicate the French language, considering it the language of colonialism. Indeed, both Daoud (2001, 2011) and Baccouche (1994) attribute the failure of Arabisation to the strong attachment of the elite to the French language and its cultural value system, which is why the political elite itself was unable to adequately promote an organised and planned process of Arabisation. The elite themselves were probably not even really interested in promoting it at the expense of bilingualism and biculturalism, let alone at a time when Islamic fundamentalism represented a realistic challenge to the leadership that took power in 1987.

> '[L]*es choix politiques en matière linguistique ont été un facteur déterminant dans la consolidation du bilinguisme et l'accentuation du prestige de la langue française en Tunisie, vu les multiples fonctions qu'elle n'a cessé de remplir dans la vie sociale, économique et culturelle.*' (Baccouche, 1994)[36]

---

35 The best result achieved by these reforms, according to Daoud (2001), seems to have been the acquisition of Tunisian students' passive competence in both MSA and French, such that they were able to understand texts and follow lectures in these languages, but not to improvise a spontaneous conversation.

36 'Political choices in linguistic matters have been a determining factor in the consolidation of bilingualism and the increase in the prestige of the French language in Tunisia, in view of the multiple functions that same has never ceased to fulfill in the social, economic and cultural life.' My own translation.

In 1988, several political parties and prominent national organisations signed the *Pacte National*, a document reaffirming the importance of the Arabic language in a global context. The call was aimed at reinforcing the Arabic language in order for it to become the language of communication, administration and education, and the national language, keeping Tunisia open to dialogue with other civilisations and languages.

> '[W]*e must strive, in this regard, to avoid the split between the elite and the popular masses.*'[37]

Among the decisions announced by President Zine El-Abidine Ben Ali (*Zīn ə-ʕĀbidīn ben ʕAlī*), in 1994, the following objectives were put forward:

1. Reinforce functional literacy in Arabic (the national language), namely MSA.
2. Improve functional competence in foreign languages, particularly French and English.
3. Generalize computer skills.

In October 1999, the Prime Minister sent a circular to public officials stating that all official acts of public administration should be converted into Arabic and outlining the regulation of same, for both internal and external documentation. The deadline for the Arabisation process was December 2000. The really short deadline was credible according to Daoud (2001), to whom we refer for further discussion.

The only real concrete result of the Arabisation process is the fact that French has remained the language of academic and professional contexts, and that only the upper-middle classes use it for everyday conversations, especially women (Daoud, 2001), despite countless socio-political debates on the topic of Arabisation and Francophilia. The Tunisian elite that led the struggle for independence from France and then took the reins of power made ambivalent choices on language policy and planning that, on the one hand, promoted the Arabic language as the anchor of Tunisian identity and, on the other hand, maintained French as the language required for gaining access to the scientific and technical and sociocultural knowledge of modernity. Daoud (2001, 2011) shows that these choices actually had deeper historical roots that concealed a consensus among Tunisia's elite, including political and union leaders, to

[37] Pacte National, 1988; in Daoud (2001).

institutionalise Arab-French bilingualism and biculturalism. These choices are still adhered to today. The sensitivity of socio-linguistic issues together with the tension and passion they generate, are still a current source of concern, particularly with regard to the so-called *langue mixte franco-arabe*, the Tunisian *koiné* of the Tunisian bilingual elite. A more in-depth exploration of some of the socio-linguistic issues related to this are included in the next section. Another consideration is that French is the native language of second and third generation Tunisians in France. Many Tunisians emigrated as political opponents to the Ben Ali dictatorship, before returning to Tunisia following the fall of the regime. This diaspora generated an impact on linguistic practices in Tunisia, as it revived the role of French as a common language, such that it no longer served as an emblem of the elite, and was instead seen as a language of international mobilisation and Tunisians from the diaspora, as well as Tunisians abroad (Allal and Geisser, 2018). Regarding the latter, if they are of the second or third generation, it is very rare for them to be competent in Arabic. Sometimes they can speak a little Tunisian, but they can hardly write or read it in Arabic characters, as also reported by Caubet (2019). For them, it was instead much easier to write in Arabizi, which will be explored in Section 2, as reported by Bashraheel (2008):

> '*Sara Ibrahim, a 25-year-old university graduate, finds Arabizi much faster when writing a message, a non-official e-mail or chatting. "I got used to typing in English. I find Arabizi a quick and fun way to write. Everyone knows it and everyone understands it. I don't like to write in Arabic; that's why I think Arabizi is brilliant," Ibrahim said.*'[38]

In recent years, the Tunisian language situation has been quietly changing and French is giving way, not to Arabic, but to English. Young Tunisians are increasingly choosing to invest time (and money) in achieving proficiency in this language, which is the current language of science and technology. An example of this phenomenon as seen in the new generation is represented by Amel Bedhyefi, a *booktuber* that publishes videos where she's speaking about English books in a mixture of English and Tunisian.[39] The motiva-

[38] Unfortunately, neither the origin nor the informant's country of residence was reported.

[39] Here is the link an interview with her developed by *Le Courrier de l'Atlas*: `https://www.youtube.com/watch?v=gDKoFm5DYko&ab_channel=LeCourrierdel%27Atlas`. Consulted on 4th April 2021.

tion for this can perhaps be the fact that the new generations no longer recognise themselves in the French cultural model in comparison with the American one, the language of which is instead experienced in a more modern and functional way in terms of the professional path.[40]

## *The Tunisian* κοινὴ

> '**koiné** /ˈkɔɪniː/ (*n.*) The spoken language of a locality which has become a STANDARD language or LINGUA FRANCA. The term was originally used with reference to the Greek language used throughout the eastern Mediterranean countries during the Hellenistic and Roman periods, but it is now applied to cases where a vernacular has come to be used throughout an area in which several languages or dialects are spoken, as in such notions as (for Old English) 'West Saxon literary *koiné*' or (for US-influenced British English) 'mid-Atlantic *koiné*'.[41]

Following the definition of the term '*koiné*' given in Crystal (2008), it can be seen that this term was originally applied to the Attic dialect, which quickly spread as an *official* language. Indeed, the original *koiné* was generally spoken by educated Greeks, although they still used their local dialect among themselves. This term has often been used to describe language varieties that were not always similar in either form or function to the original Greek *koiné*. The term 'koineisation' has more recently been applied to the process of *levelling* that can result in a *koiné*. In this context, the term 'levelling' refers to reduction or linguistic cleansing of all the marks which are characteristic of a specific variety; it represents a kind of flattening of the representational power of a specific micro-culture in favor of regularity, simplicity of use and national representational power, spread across regional boundaries.

As far as we know, the first to use the term 'koineisation' was Samarin (1971). Samarin describes this process as 'dialectal mixing':

> '*What characterises them* [koinai] *linguistically is the incorporation of features from several regional varieties of a single language. This kind of amalgamation (or dialect mixing) can lead to a certain amount of heterogeneity. That is, a* koiné, *caught at an early stage of its history, might consist of many kinds*

[40] For a more in-depth look at the rivalry between French and English in Tunisia, which, despite being a particularly interesting topic, is not central to this work, refer to the study developed by Daoud (2001, 2011).

[41] The scientific transcription given in Crystal (2008) refers to the English realisation, but concerning the Greek, it is: [kɔɪˈneː].

*of speech that are not easily correlated with non-linguistic factors like region, function, social status, etc.'*

However, he explains 'koineised colloquial Arabic', as previously defined, as being a type of 'levelling'. In his view, this levelling occurs in situations of 'interdialectal contact' when speakers attempt to suppress typical features of the local variety in favour of features that are simply more common, and more familiar. Indeed, he mentions the words of Blanc (1960):

*'In certain situations, usually interdialectal contact, the speaker may replace certain features of his native dialect with their equivalents in a dialect carrying higher prestige, not necessarily that of the interlocutor ... Moreover, levelling devices may be called into play without the speaker actually stepping out of his native dialect, but by selecting from among a number of equivalent features available to him those which are more general or more urban and suppressing those which sound local or rustic ... In general, levelling often takes place not so much in imitation of a specific dialect as in an attempt to suppress localisms in favour of features which are simply more common, more well known.'*

What can be deduced from Samarin's study is that, for him, the result of a koineisation process is an individual *koiné*, which is determined on contextual needs. Furthermore, Section 1 includes a reconstruction of the process of emergence and the diffusion of dialects, and it was noted that this is still an open and debated issue, not only in academia. In fact, it is such a lively issue that nowadays it remains at the center of heated discussions (even on social networks). Despite scholarly disagreement on the historical reconstruction of the emergence of dialects (the theory of monogenesis presented by Ferguson (1959), *koiné* as an endpoint by Cohen (1970), and the process of pidginisation by Versteegh (1984)), scholars seem to mostly be in agreement on the importance of contact between soldiers, as well as the idea of prestigious urban centers exerting their influence in the process of Arabisation and playing an important role in the formation and spread of 'modern' dialects. The influence of prestigious centres is a dynamic which has remained valid over the centuries. As pointed out by Miller (2004), in the post-colonial phase, i.e. since the 20$^{th}$ century, major city dialects are often emerging as national or regional standards in both the Maghreb and the Middle East. In fact, in French Tunisia, as seen in Section 1, political power was in the hands of foreign rulers and the Tunisian elite (regarding the latter, the elite remained so anchored to the political role it had at the time of the protectorate that it adopted the language of the occupiers as a distinctive feature (Daoud, 2011)). However,

following independence, Tunisian society was reorganized, leaving more space for political groups, but above all for local identities, or at least for politically prevalent ones. The history of the 20th century in Tunisia is explored in more depth below (Section 1).

In the same century, the phenomenon of urbanisation meant that the main cities, particularly Tunis, welcomed Tunisians from all over the country, a phenomenon particularly related to the process of koineisation leading to a rise in literacy levels (Holes, 1995). Miller (2004) describes the processes of koineisation that took place in the Arab world through five different modalities, regarding the modality that concerns the Tunisian process, she stated:

> *'Old urban centers with a declining urban elite and an important population renewal. The old urban dialect is no more prestigious enough to be acquired by new-comers and is even declining among the young generation of the original urban dwellers. The new-comers adopt instead the national urban koiné, and urban dwellers tend to speak this urban koiné in public space, keeping their own vernacular at best for family communication. This is the case in many North African cities like Fes, Tangier, Rabat, Tlemcen, and even Tunis, where the old urban vernacular tends to become more and more restricted to old women and is associated with an effeminate way of speaking.'*

Therefore, it is possible to affirm that the Tunisian *koiné* of Tunis (*Tūnsi*) does not coincide with the old dialect of Tunis' old city center (Tunis medina), but rather with an urban *koiné* which emerged in the 20th century. In particular, that which may be perceived as the 'national' variety, which takes up Blanc's description above, appears 'more general or more urban and suppresses those that sound local or rustic'. Regarding the differences between the old and the new Tunis-*koiné*, for example, Singer (1984) says:

> *'Sie unterscheidet sich von der allgemeinen stadttunisischen Koine der Gesamtregion in den Ohren der nicht zu ihr gehörenden Bevölkerungsgruppen durch eine gewisse prezïöse Artikulationsart (z. B. starke Imäla zuweilen auch bei Männern, die gelegentlich desgleichen die in der Frauensprache erhaltenen Diphtonge artikulieren) und eine ganze Anzahl ihr eigener Vokabeln, die außerhalb dieser engsten Gruppe nicht verwendet werden.'*[42]

[42] 'It differs from the general urban-Tunisian *koiné* of the whole region to the ears of population groups that do not belong to it due to a certain pretentious manner of articulation (e.g. a strong Imāla [the shifting from /a/ ~ /ā/ towards /i/ ~ /ī/, in a not-marked context, that will be addressed in Section 2] sometimes also used by men, who occasionally similarly articulate the diphthongs found in women's language) and a whole vocabulary of

In Section 2, we will examine how communication technology is supporting the spread of this national *koiné*. Before that, however, it is necessary to point out some not irrelevant aspects concerning the status of Tunisian contemporary *koiné*.

Based on the definitions of the koineisation process given by scholars, such as Samarin (1971), Blanc (1960) or scholars mentioned in Section 1, it appears that koineisation involves the mixing of features of different dialects, and leads to a new dialect of compromise, which is not yet a standard dialect (considered to be the result of a top-down regularisation, see Section 2), but is something similar to a Lingua Franca (see Section 1), being the result of an agreement among speakers of the individual contributing dialects, which may or may not be maintained over time. In other words, as expressed by Siegel (1985), koineisation is a process that involves contact between linguistic *subsystems*, even if that contact does not always bring about koineisation. There must be certain socio-linguistic conditions for a *koiné* to emerge from contact. These socio-linguistic conditions relate to the nature of the subsystems involved in contact. The scholar defines subsystems as different linguistic varieties, which are genetically related to the same linguistic system and therefore also typologically similar enough to be adequately reflected in the following two criteria:

1. Being mutually intelligible;
2. Sharing the same language system, national standard language or literary language as superior.

Although a widespread perception is that of the dialect being 'younger' than the classical language, as discussed above, the discourse on the diachronic evolution of modern dialects is still open, so instead of referring to a *genetically superior language*, we must instead refer to a language perceived as *superior in terms of diglossia*. The two parameters, along with the scholars' definition of the koineisation process, seem to be appropriate for both the interlanguage process of formation of the military camps in Iraq, on which some scholars agree, and the process described by Miller for the modern *koinai* of Tunisian Neo-Arabic.

As previously mentioned, however, koineisation is not a necessary process. In fact, speakers of different varieties can be in a situation involving stable contact, with well-defined social and functional roles,

---

their own, which is not used outside of this smaller group.' My own translation.

for long periods without ever reaching a levelling off. In addition to koineisation, other different processes can occur, such as diffusion for example, which involves the transfer of linguistic aspects beyond the traditional domains and regions of belonging. Other reasons for the levelling of the features of local variants also exist. Holes (1995) describes the linguistic practices following the urbanisation process in three Arabic capital-cities: Manama, the capital of Bahrein, Amman, the capital of Jordan, and Baghdad, the capital of Iraq. All three case studies exhibit a prevalence of features of the smaller urban community variety over the Bedouin variety. Regarding Bahrein, urbanisation and the consequent increased literacy resulted in level dialect differences. In Jordan, following the massive Palestinian immigration due to the Arab-Israeli wars, urbanisation was seen on a larger scale. Holes observed some general diastratic tendencies, such as different choices between young women and young men regarding whether they are Palestinian or Jordanian. The first group seems to be more inclined to follow phonetic patterns which are perceived as sophisticated (such as the phonetic [ʔ] urban variant of */q/), while the second group followed masculine models (i.e. preferring the typical Bedouin [g] for */q/). Palestinian men in general tended to avoid Palestinian identity language markers, such as [ʔ] or [k] by instead opting for [g] as an empathetic feed towards Jordanian interlocutors.[43] Another tendency highlighted by Holes (1995) is the general substitution of the rural [č] (*/k/), charged with negative evaluation, with both the urban and Bedouin [k]. Baghdad also exhibits practices connected to the perception of one particular variety as more prestigious than the others. In that case of opposition between sedentary *qəltu*-dialects and Bedouin *gəlet*-dialects, this broadly corresponds to the speech of religious communities, where the Muslims are *gəlet* speakers, while non-Muslims are *qəltu* speakers.[44] In Baghdad, the Muslim-Bedouin dialect is the one which is currently perceived as more prestigious because of the political prominence of the Muslim population since Iraq's independence in 1932. In brief,

[43] See also the study developed by Abd-El-Jawad (1987) regarding the Jordan linguistic association with socio-political values relating to the issue of Palestine. He analysed new prestigious local or regional varieties in dichotomous competition with MSA in informal settings.

[44] These names represent the different phonetic realisation of */q/ in the word qəlt, meaning 'I said'. In addition, in Tunisia, the speaker is aware of the two different way to speak, and they use *tkallәm b-әl-qāla*, as opposed to *tkallәm b-әl-gāla*, to distinguish between the /q/ and /g/ pronunciation. See Chapter 4 for a deeper analysis of this phenomenon.

simply in terms of diglossic realities, the process of koineisation of Neo-Arabic varieties has proven to be more complex than that which can be explained by a simple Fergusonian H-L dichotomy, because of contextual factors related to language use in various ways. However, in all the three cases reported by Holes (1995), a trend seems to be emerging: the dialects of the capitals have lost the connotations they previously had, i.e. being linked to specific ethnic-religious communities, and have instead taken on the role of the country's *koiné*.

Considering the distinctive macro-linguistic-features of sedentary and Bedouin dialects, i.e. those already schematised in tables 1 and 2 of Section 1, the levelling out of these characteristics can result in two possibilities:

1. A mix of the two varieties, or
2. The prevalence of one variety over the other.

According to Ghoul (2009), the urban *koiné* of Tunis is a variety of the Tunisian dialect, characterised by flexibility, inclusiveness and modularity, which Ghoul defines as *la langue in vivo* in comparison to *la langue in vitro*.[45] In other words, a kind of *spontaneous* language policy, based on individual initiative, that corresponds to the effective language practice of the speakers. Indeed, Ghoul (2009) states that:

> '*En raison de son caractère informel, cette langue paraît beaucoup moins contrainte que celle de l'affichage officiel, même si certains écrits peuvent parfaitement répondre aux exigences de l'écrit normé, notamment les enseignes de certains commerces ou de certains établissements scolaires privés. La différence fondamentale entre la langue in vitro et la langue in vivo est que cette dernière introduit, en plus de l'arabe standard et du français, l'arabe dialectal.*'[46]

It is now the dialect of reference and has been assimilated into the 'Standard' Tunisian Arabic, which is now perceived as the prestigious variety of Tunisian (Durand, 2009; Gibson, 2008; Mion, 2004; Tarquini, 2019). In particular, it has been proposed as a variety which should be elevated to a standard by a Tunisian association known as

[45] Corresponding to a language policy decided by the State.

[46] 'Because of its informal nature, this language seems much less constrained than that of official signs, even if some written material can perfectly meet the requirements of standard writing, such as the signs of certain businesses or private schools. The fundamental difference between *in vitro* and *in vivo* language is that the latter introduces, in addition to Standard Arabic and French, dialectal Arabic.' My own translation.

'Derja', which was created in 2016, for the promotion of the Tunisian dialect and its recognition as an official language of the country. The militant members of the association are mainly university professors, who meet regularly at 'Beit el Bennani', a private house in the medina of Tunis, which has become a kind of cultural space.[47]

According to some, however, the banner of Tunisian *koiné* as the language of the people is nothing but a political strategy, devised by the left of Bouguibian heritage, for dividing the people, in which the *koiné* of the elite is chosen as the national language, at the expense of the internal and southern regions, which are less developed and less permeated with French culture and especially bilingualism. In a similar way, Allal and Geisser (2018) state that the *langue-mixte-franco-tunisien* is the language of power, not the language of the people, and that same would consistently be difficult for some people to understand as it is less involved in French biculturalism and bilingualism. In fact, Myers-Scotton (2006) defines the ability to code-switch back and forth between the foreign language and the local one as a mark of the educated elite and thus as a typical *elite closure* practice. The same practice has also been registered by Miller (2004), as quoted above. Before presenting the dialect advocacy work of the Association Derja, it is however necessary to clarify a few questions about the concept of a national standard language.

## 2. Diffusion of spontaneous orthography

### *Standard Languages*

> '**standard** (*n.*) A term used in socio-linguistics to refer to a prestige VARIETY of a LANGUAGE used within a SPEECH COMMUNITY. 'Standard languages/dialects/varieties' cut across regional differences, providing a unified means of communication, and thus an institutionalised NORM which can be used in the mass media, in teaching the language to foreigners, and so on. Linguistic FORMS or DIALECTS which do not conform to this norm are then referred to as **substandard** or (with a less pejorative prefix) **non-standard** – though neither term is intended to suggest that other dialect forms 'lack standards' in any linguistic sense. The natural development of a standard language in a speech community (or an attempt by a community to impose one dialect as a standard) is known as **standardisation**' (Crystal, 2008).[48]

[47] See the Facebook page of Beit el Bennani: `https://www.facebook.com/Beit-el-Bennani-520104428117976/`. Consulted on 4th April 2021.

[48] Emboldened and capitalised words are taken directly from Crystal.

Let us imagine a native Tunisian student who, for example, moves to the capital Tunis. Based on what we saw in the previous section, it is most likely that, in order to communicate with his classmates, he will gradually adopt more and more of the Tunis-*koiné*, without however ceasing to use his own dialect of origin for communication with his family. Similarly, let us imagine a blogger who does not want to address a particular reader, but instead wants to communicate with the whole of Tunisia by making an inclusive linguistic choice. He might choose to level the features of his own dialect, in favour of a kind of *Middle Tunisian*. However, if a second blogger, regardless of whether they are writing in the same period or not, makes the same kind of linguistic choice, it is not certain that the linguistic product would be the same as the first one. The fruit of the spontaneous choices of each speaker would be a singular product shaped by the variables of the specific context, and is perhaps not even an idiolect, since it would not necessarily be maintained over time, even by the same writer. These exemplified processes are rather spontaneous and are operated at a more or less conscious level, and are of the same level as that which Tunisians already use when they find themselves sharing spaces, whether real or virtual, which are different from those of their origin. These are down-top processes, operated in order to facilitate inter-regional communication. This does not mean that Tunisian native speaker from different regions necessarily have great difficulties in terms of inter-comprehension. Despite this, some Tunisians would like to propose the Tunisian dialect as the official language of the country. In particular, following the Arab revolutions, the question of standardisation of Tunisian has begun to be raised as a matter of identity prominence. In order to standardise Tunisian, as we have seen in the definition of *standard*, institutionalised norms are required to regulate its use. However, before considering norms, it is imperative to figure out what linguistic material to operate on. It is from this point that at least two choices arise:

1. To elevate to a standard version of the *koinai* that we have seen emerging in the 20$^{th}$ century;
2. Levelling the traits of the various rising varieties within the country by creating a rather artificial seeming Middle Tunisian.

In the first case, it would be a choice which should be made on the basis of the most prestigious *koiné* or the most widespread through the national media, without taking for granted that same is probably the *koiné* of Tunis. In the second case, it would be a rather complicated operation because each individuals' perception of what

the most representative features of the Tunisian identity are, which should be kept alive in the national language, so that it represents the whole nation. However, choosing either of the first type or of the second type, in favour of a national language, would certainly destroy local micro-identities, despite the fact that having a national standard does not necessarily presuppose the suppression of local varieties, which could still be maintained in private contexts or in contexts in which all the participants in the communication belong to the same linguistic micro-community. In any case, regardless of whether one wants to proceed with the first option or prefers the second, the first step is orthographic standardisation.

The officially recognized languages in Tunisia today are French and Standard Arabic, so at the level of standard orthographies, both the Arabic and Latin orthographies exist. On the, albeit infrequent, occasions when Tunisian Arabic is written, it has been encoded using the Arabic orthographic system,[49] with the exception of electronic communication and messaging contexts, where Tunisian is also written in Latin characters, as seen in Section 2. However, the use of the Arabic orthographic system for the encoding of Tunisian presents a number of problems, which could be solved through the creation of a new encoding system modulated on an approximation of Tunisian phonology. It is probably also for this reason that the recognition of Tunisian is perceived by some as a replacement for Standard Arabic, which was a vehicle for a rich literary tradition, as well as a relative of the Classical Arabic holy language. The Association Derja proposed both an Arabic-based and a Latin-based input system for Tunisian encoding, as explained later in this chapter (Section 2). At the same time, the encoding of Tunisian in Latin characters evokes a historical precedent, i.e. that of the Maltese language (the only Arabic dialect that reached the status of a language (Miller, 2012)). Azzopardi-Alexander and Borg (2013) report that the Arabic vernacular spoken in Malta is the result of linguistic contact with Muslim Sicily, but that its origin seems to have been Tunisia. In fact, Maltese exhibits some areal traits typical of Maghrebi Arabic, although it has diverged from Tunisian Neo-Arabic during the last eight hundred years of independent evolution. Both historical and ideological reasons, among others, have led to the orthographic standardisation of Maltese in Latin characters.[50] Let us then exam-

[49] See Langone and Mion (2019) for an example.

[50] In 1924, the orthographic code devised by the *Għaqda tal-Kittieba tal-Malti* (Association of Maltese Writers) was adopted as the official alphabet of the Maltese language (Cassola, 2014; Vanhove, 2003).

ine the motivations, implications, and the canonical path of the orthographic standardisation of a language.

*Orthographic Standardisation*

There are many reasons for supporting the use of a single regularised orthography for a country's language or dialects. The main reason is certainly the potential to generate a shared social and linguistic identity for communities that use different spoken varieties of a language. The use of a single orthography offers a way to overcome regional boundaries that represent a division not so much of territory, but more importantly of identity (Simons, 1977). For example, in the case of the Chinese logographic system, the shared writing system has had the desired effect, quantitatively extending the possibility of reading and reducing the difficulties of learning a large number of Chinese characters.[51] Regardless of the advantages of a shared writing system, there are also potential difficulties, such as the issue of 'artificial systems' that arise during the construction of a multi-dialect orthography, which ends up representing none of the varieties that were intended to be synthesised. This is a risk particularly in divergent phonological systems where a shared orthography cannot faithfully adhere to the phoneme-grapheme correspondence of each variety which it seeks to represent. According to Simons (1977), there are two basic options in this situation:

1. A specific dialect may be selected as the basis for the standard and used as a writing model,
2. Choices are made regarding the development of an orthography calibrated so that the set of its features is representative of the set of dialects sublimated into the variety which is elevated to the standard.

The resulting spelling will not represent the most psychologically real spelling for each individual dialect, but can still be easily learned and used for all of them. Simons argues for the second option in all circumstances, because the first requires speakers of non-standard

[51] A logogram can be used in a number of languages to represent words with similar meanings. The same is true for alphabets, but the degree to which they can share identical representations for words with various pronunciations is much more limited. Arabic numerals are also logograms.

dialects to learn at least some aspect of the orthography by rote memorisation.[52]

In addition, there are proposals that aim to normalise not only spelling, but also grammar and vocabulary, such as Corraine's standardisation proposal for Sardinian, which included grammar, orthography and lexicon standardisation. In this case, the proposal was refused, as it was mostly based on *Logudorese Sardinian*, one of the two main varieties that the Sardianian's traditional classification takes into consideration, together with *Campidanese Sardinian*. As a result, language speakers claimed the right to be informed and involved in the decision-making process (Lai, 2018). As also highlighted by Simons (1977), direct involvement of the people is in fact necessary for the development of a valid standard, because the most important factor for the success of language planning measures are their acceptance by the speakers of the language(s) involved. Regarding Sardinian, with its high level of linguistic variation, it would be possible for one variety to gain ground on another if speakers gave it greater prestige (Lai, 2018). One of the most common reasons for promoting a written standard is the desire for political or cultural unity in favor of the common development of the country. A single standardised writing system makes communication easier in a number of ways, such as in terms of education, given that instruction in Tunisia is taught in Tunisian, sometimes even in high school. Having a standard orthography can increase the functional domains of a language's use, which in turn increases its status within the community and reinforces community values (Schiffman, 1998). Even though standardisation has certain indisputable benefits, it is not without social consequences. One of the most obvious is the development of consciousness and beliefs about what is correct and incorrect. Prescriptive ratings of language forms are introduced with the written language form; native speakers tend to have smaller fixed notions of accuracy before a language is written. As such, an orthography represents a regulatory idea that does not have counterparts in the linguistic reality of the speech community (Coulmas, 1999). The main concern about the standardisation process is that it actually leads to the disappearance of linguistic variation. Standardisation has been accused of potentially contributing to the loss of linguistic diversity, since a written standard inhibits the variability allowed in

[52] In order to show how an orthography can be created by drawing from multiple dialects, Simons discusses the Dani language of Irian Jaya (Western New Guinea), based on data in Bromley (1961).

the language and thus inevitably causes the disappearance of features which diverge from the standard form of the language. In oral communication, a great array of dialectal variation can be maintained, but literacy necessarily favors standardised languages and discourages variation. Furthermore, standardised conventions for local languages are often explicitly shaped on the writing conventions of prestigious languages or languages of more widespread communication in order to facilitate the acquisition of the standard, which on the other hand may facilitate the loss of the original languages. Another reason why standardisation might contribute to the loss of variation lies in the status that is given to one variety over others as the former is elected as the standard. Writing certainly has this power to elevate the status of a language variety, to the detriment of other surrounding languages or varieties, which, not having a written tradition, may be perceived as lacking in prestige and therefore to be avoided, even in oral use (Coulmas, 1999). However, the notion of standardisation in written languages is itself a convention, being an abstraction from *spoken* varieties (Shortis, 2007). This is particularly true for informal writing systems (Thurlow and Poff, 2013).

### *Process of Orthographic Standardisation*

Among all the factors that must be taken into consideration during the standardisation process, one of the most relevant is sociopolitical factors. In fact, there are often unspoken political implications in spelling choices that must be taken into consideration. Creating an orthography that is highly distinct from what is already perceived as the national writing system can also be seen as subversive or challenging to the goal of national political unity. Moreover, if a language is used in different countries, or different regions of the same country, it should always be coded with the same spelling system (Baker, 1997). In addition to the political implications for countries that share the same language system, as mentioned in the previous subsection, one must also consider the consequences that the imposition of a spelling standard will have on linguistic variation in the target country. Grenoble and Whaley (2005) list a number of issues to be considered:

1. People's attitudes toward different varieties;
2. Which variety has the largest number of native speakers;
3. Which variety is most widely understood;
4. Which varieties are mutually intelligible;

5. If people already consider one variety to be more prestigious;
6. If they consider one to be more 'pure' or closer to the 'original' language;
7. Where the varieties are spoken (especially if one variety is spoken in an urban center);
8. Which varieties are used for religious or administrative purposes.

Many of these considerations are related to the perceptions of native speakers, and are insights that can only be held by members of one or more local communities. Therefore, the ideal solution is for the standardisation committee(s) to consist of, or include, representatives from these local communities, so that they can be a part of the decision-making process. If this is not taken into account, committees may be inefficient, and may function differently than expected. The same is true with regard to particular cultural taboos. After determining which variety will be elevated to the national standard, members of the language planning committee should consider which aspects of spelling should be standardised. In addition to standardising the choice of writing systems and particular symbols within a system, it is also important to determine which other conventions should be introduced, such as whether, in the case of the Tunisian's Romanisation, capital letters should be included or not, since the Arabic alphabet does not have these, or how the punctuation system should be organised. In making these decisions, one should take into account the conventions used in the national and broader languages of communication, which in the case of Tunisia are Standard Arabic and French. The standard chosen may have to reflect one or both of these, or deviate from them in order to be accepted by the Tunisian language community. On the other hand, with reference to what we have seen in Section 1, in the Tunisian diglossic and bilingual context, there are both standard languages and Tunisian Neo-Arabic, which Daoud (2011) split into two types, i.e. national and local. Encoding a standardised Tunisian Neo-Arabic should not distort the state of functional organisation of this continuum. It should again be reiterated that this is a very delicate task, and that there are several issues to be taken into account when doing so. Finally, potentially the most delicate moment in the process of creating a standardised spelling comes in the testing phase. In fact, the spelling should be tested to evaluate its effectiveness and to address any problems. Depending on the accuracy of the work of the committee, a number of issues may arise, perhaps due to particu-

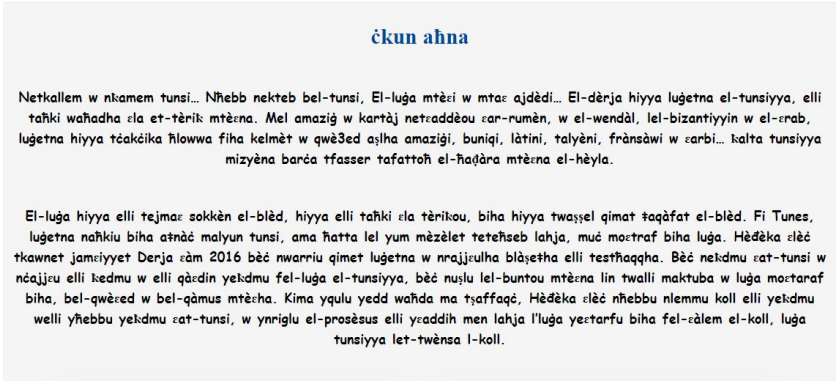



*Figure 1 Association Derja Presentation*

lar symbols or diacritics that may be ambiguous or difficult to use (Grenoble and Whaley, 2005).

A group that declares itself to be a promoter of the Tunisian dialect, known as the Association Derja, has already been mentioned above. The following section presents a brief overview of how they present themselves and the motivations that drive their work.

### *Association Derja*

I speak Tunisian and think in Tunisian [henceforth *Tūnsi*]… I want to write in Tūnsi, it is my language and that of my grandparents… *Dārža* is our Tunisian language, that speaks for itself about our history. From Amazigh and Carthage we passed from the Romans and the Vandals to the Byzantines and the Arabs, our language is a sweet mix [containing] in it words and rules whose origins are Amazighi, Punic, Latin, Italian, French and Arabic… a very beautiful Tunisian mix explains the openness of our beautiful civilisation.

That language is the one that brings together the inhabitants of the country, it is the one that talks about its history, through it the culture of the country takes on value. In Tunisia, we speak our language with 12 million Tunisians, but until today it is still considered a dialect, it is not recognised as a language. For this reason, the Derja Association was founded in the year 2016 to show the value of our language and allow it to reach the place that it deserves. We will work on Tūnsi and encourage those who have worked on Tūnsi and those who are working on it, to reach our goal, so that it becomes written and recognised, with its own rules and dictionary. As it is said, a hand does not clap alone, so we want to gather all those who work and want to work on Tūnsi,

and fix the process that makes it go from dialect to the language that the whole world can recognise, Tunisian language for all Tunisians.[53]

This is the Association Derja Manifesto, which appears on their website homepage.[54] As we can see, the association has several goals: to codify Tūnsi, to develop a manual of the Tunisian language, and to create a library of works written in Tūnsi. Derja wants to be a place of debate and a catalyst for ideas. As per their manifesto, Tūnsi is perceived to be the true mother tongue of Tunisians. According to the members of the association, Tunisian *dārža* experiences the paradox of being the most widely used language, but not being recognised as a 'language' in its own right, as it does not exist as a means of written knowledge and information. In fact, few books, articles and other documents are published in this language, although Tunisian does appear in radio and television programs, and sometimes takes precedence over Modern Standard Arabic (MSA). The main reason that same has not already being recognised as the official language of Tunisia seems to lie in the absence of and lack of consensus on standardised spelling and written grammar. However, there have been many projects to promote Tunisian *dārža* in recent years, including various literary or artistic projects, for example, such as that pioneered by the writer Hager Ben Ammar, who has published traditional children's stories in Tunisian, which have been very successful commercially and were also a success, both culturally and educationally. Reporting the words of a member of the association, during a conference held on 27th February 2020:

> *Dārža is the language in which Tunisians recognise themselves. It is one of the foundations of our identity, but in which we can also easily express ourselves.* [And] *Today, in the context of democratic transition, our mother tongue is more than essential for social and cultural cohesion because of its origins in the collective popular memory of Tunisians.*[55]

Such is the message of the Association Derja, and the reason why the association wants to develop a manual of the Tunisian language and create a library of books in Tunisian.

---

53 This is my translation. I have tried to respect the structure of the original text.

54 The link to the Association Derja is as follows: `http://www.bettounsi.com/`. Consulted on 4th April 2021. All the contents are written in both encodings: the Latin script and the Arabic script.

55 My own translation.

The most obvious demonstration of what has just been stated is the wave of rebellions and demonstrations that began in the Arab world at the end of December 2010, starting in Tunisia. This wave relied so heavily on Internet communications and social networking tools, which are said to have contributed to these uprisings on two distinct levels:

1. Mobilisation,
2. Deconstruction.

Regarding the first case, digital platforms had a political impact, being instrumental in mobilising protesters and coordinating their activities. Indeed, young Internet-savvy people used social networking sites to disseminate information and organise rallies, repeatedly circumventing government attempts to shut down websites and servers.[56] Regarding the second level, the way in which Internet communication impacted the so-called 'Arab Springs' was the release by WikiLeaks of a series of protected information that deconstructed the image of Arab leaders, revealing their diplomatic agreements with the United States. Reposted across other types of media platforms, these revelations fueled resentment against Arab governments for the masses.

The suicide of Bouazizi in December 2010, is generally considered to be the *casus belli* of the Arab revolutions, but the truth is that the discontent had already been rife for some time. In 2008, due to a rigged public contest, a revolt that is considered to have been the forerunner to that of 2010 broke out in the town of Redeyef (*Gafṣa*). Despite the fact that the clashes between the striking miners and the army were very violent, almost no news reports about the incident were leaked. There are those who argue that the Gafsa uprising had much more potential than that of 2010, but that its retreat was also determined by the low resonance of same across the country at the time, due to the embryonic stage of the media. In Tunisia, until 2011, several sharing platforms such as Youtube and Flickr, various Facebook and Twitter pages, specific blogs and information sites, including the Wikipedia page on censorship in Tunisia, were blacked out.

As far back as in the 90's, some pioneer 'cyberactivists' were fighting censorship through counter-information activities, among them *Anonymus* and *Takriz*.[57] In May 2010, there was even a

[56] Particularly in the uprising that began on 25 January 2011 in Egypt.

[57] There was the particularly sad case of a well-known Takriz activist Zouhair Yahyaoui, creator of the online magazine TUNeZINE, who was

campaign against web censorship known as *sayyəb ṣālaħ*.[58] These facts were further fuelling a groundswell of discontent. As shown by the chronology of events reported by Bettaïeb (2011), the revolution's outbreak was strongly supported by media outreach. During the months following Bouazizi's suicide (on 17th December 2010), the protests quickly spread to the neighbouring city of Sidi Bouzid. The news spread just as rapidly, through both photos and hashtags, such as *#sidibouzid*, spread on social networks, through which the demonstrations were organised in the squares. Moreover, cyberactivists spread the images to satellite channels such as *Aljazeera* and *France24*.

Many Tunisians today describe the revolution as a failed revolution, mainly because of the serious economic and political crisis that has been going on for ten years. The great conquest of the revolutions was freedom of expression, and this seemed to be exactly the way in which to express all of one's frustrations. The main channel for such expressions have been, since the beginning, the social networks.

Regarding Ben Ali's last speech (13th January), every time he had previously made a speech, he had always read prepared notes in fuṣḥā and pronounced them respecting the phonetics of the prestigious variety.[59] On the other hand, Boussofara (2006) argues that Ben Ali, by imposing fuṣḥā as the only language of official political discourse, made a calculated choice aimed at legitimising his authority and breaking away from past traditions. The implicit devaluation of *dārža* on the official level, however, never prevented the latter from conquering more and more areas of action according to the previously mentioned trends. The linguistic issue was immediately highlighted as being politically prominent with the

sentenced to two years in prison for spreading false news and fraudulent use of the media.

[58] Note the Arabizi spelling used for the slogan *Sayeb Sala7*, as reported by the article published on *Nawat* blog at the following link: `https://nawaat.org/2010/05/29/anti-censorship-movement-in-tunisia-creativity-courage-and-hope/`. Consulted on 4th April 2021.

[59] Unlike his predecessor Bourguiba, who did not disdain the use of *dārža* or French. As Sayahi (2014) reports, Bourguiba often expressed himself in 'Middle Arabic', making use of frequent code-switching from one language to another.

first post-dictatorship Republic President, Moncef Marzouki,[60] who considered the Arabic language to be the main means of pan-Arabic identity and inter-communication. He rejected the practice of introducing the French lexicon into Tunisian, considering this *mixed language* to be a political expression of a specific social group, i.e. the decadent elite of Tunis, that would marginalise the internal regions of Tunisia. However, his successor, President Beji Caid Essebsi, expressed himself in Tūnsi, supporting the idea that the language of the urban elite of Tunis coincides with the language of the people (Allal and Geisser, 2018). A new linguistic inversion was represented by the President Kaïs Saïed, who has shown a preference for what, as in Section 1, has been defined as LA, or the language used in the literary heritage. The President's choice to express himself in LA during public speech has been largely criticised by the online Tunisian journal *webdo.tn*:

> *'Qu'il se mette à la place des nombreux Tunisiens qui attendaient son allocution et ont été frustrés de ne pas comprendre ce que disait leur président.'*[61]

However, the situation regarding diglossia and bilingualism in Tunisia has already been discussed (Section 1), and, on the contrary, we have seen that perhaps it is the so-called *langue mixte* that is not widely understood by the Tunisian people outside of the area under the influence of Tūnsi.

### *Tunisian Computer-Mediated Communication*

Intentional communication is narrowly defined as any action or actions that a person consciously uses to affect another person's behaviours (Miller, 1966), while Computer-Mediated Communication (CMC) is a branch of study that deals with how information technologies, together with computers, make peculiar forms of communication possible between people at a distance from one another

[60] Historical opponent of Ben Ali. Exponent of the Islamic party Ennahda.

[61] 'Let him put himself in the shoes of the many Tunisians who were waiting for his speech and were frustrated by being unable to understand what their president was saying.' My own translation. Article available at: `https://www.webdo.tn/2020/03/21/et-la-derja-tunisienne-pourquoi-kais-saied-senferme-t-il-dans-cette-langue-rigide/`. Consulted on 4th April 2021.

(Miller, 2012). As early as 1992, Walther and Burgoon (1992) stated that:

> '*Computer-Mediated Communication is no longer a novelty but a communication channel through which much of our business and social interaction takes place, and this transformation is expected to continue.*'

One new phenomenon is the ability of human beings to adapt their language to the medium of communication. With the advent of CMC before web applications, such as personal e-mails, chat channels and forums, a new type of discourse was born, which is defined as the Electronic or Computer-Mediated Discourse or Digital Networked Writing (DNW). This is known to be a type of writing characterised by linguistic and extra-linguistic peculiarities specific to technological networks (Androutsopoulos, 2011; Crystal, 2004; Panckhurst, 2006, 2009). Messaging and the use of social networks has become the emblem of a globalised society. DNW has distinctive characteristics that differs from in-person oral communication. As Morel and Doehler (2013) noted:

> '*Dans la conversation orale, les personnes s'entendent, perçoivent timbre et intonation, souvent aussi gestes, mimiques et déplacements. Ils peuvent réagir rapidement, interrompre l'interlocuteur, intégrer ou refuser d'autres interlocuteurs, etc. Dans la communication par SMS ou WhatsApp tout cela est bien différent. Parmi les caractéristiques les plus saillantes de la communication médiée par téléphone portable, on compte le recours au code écrit, le régime temporel différé à échéance variable et une disposition spatiale où l'absence physique de l'autre est dominante.*'[62]

Following Androutsopoulos (2011), a typical DNW is organised around four main conditions:

1. It is vernacular;
2. It is focused on interpersonal relations;
3. It is spontaneous;
4. It is interaction-oriented.

[62] 'In oral conversation, people hear each other, perceive timbre and intonation, and often also gestures, facial expressions and movements. They can react quickly, interrupt the interlocutor, integrate or refuse other interlocutors, and so on. In communication via SMS or WhatsApp, all of this is quite different. Some of the most salient features of cell phone mediated communication include the use of written code, a time-delayed regime with a variable time frame and a spatial arrangement where the physical absence of the other is dominant.' My own translation.

There are many scholars, who have, in recent decades, devoted themselves to the study of messaging and communication via social networks. For a detailed analysis of the characteristics of CMC and DNW, it is imperative to refer to Androutsopoulos (2011); Baron (1984, 2010); Crystal (2004); Herring and Stoerger (2014); Panckhurst (2006, 2009). With regard to the Andoutsopoulos, he argues that DNW has the potential to expand the vernacular writing practice, in addition to diversifying writing styles and setting up new writing norms. He states that:

> '[t]*he expansion of digital literacy practices afford vernacular written usage more space, visibility and status than ever before, and vernacular usage itself is diversified in what we might call 'old vernaculars', representing locally bound ways of speaking that traditionally didn't find their way into (public) writing, and 'new vernaculars' - new patterns of differentiation from written standards, indexing practices and networks of digital culture.*'

This statement seems very much in line with the view of Ong (1986), who describes the potential of writing to entrust the word to space by expanding its power. Indeed, Androutsopoulos (2011) coined the expression 'secondary orality' to depict DNW, employing three keywords to portray the essence of DNW: *orality*, *compensation* and *economy*. With regard to the second of these, same refers to the compensatory-devices employed in DNW to offset the lack of prosody, due to non-present communication, which will be further explored in the following section. Concerning the last theme, same refers to the strategic apparatus for achieving both effective and concise communication. These key words also fit the context of Arab DNW. One of the first to analyse the Arab CMC landscape was Mimouna (2013), who dedicated herself to the study of Algerian users of e-mail communications. The initially concern was the study of the different linguistic features of e-mail communication in order to discover the various attitudes towards the impact of the e-mail language on the standards of the traditional written language. She also offers a pedagogical dimension to the study of e-mail communication among young university students. Among these early scholars of Arabic CMC was Albirini (2016), who offers a broad vision on the uses of Arabic on the Internet from a socio-linguistic perspective. In his analysis, the socio-linguistic landscape of digital media is characterised by multilingualism and the widespread use of code-switching. In particular, it shows that young network users are not monolingual and that they are aware of the communicative context in which they interact. For this reason, they use an informal register

that does not aim to reproduce the rules of standard language. In comparison, Palfreyman and Khalil (2003) focused on Gulf Arabic instant messaging, while Warschauer et al. (2002) focused on Egyptian CMC, analysing it in terms of broader global trends of language, identity, and globalisation.

With regards to the Tunisian situation, an explanation has already been provided regarding the fact that since the revolution there has been a multitude of opportunities, which were previously almost completely absent, for informal public political communication, such as press conferences, interviews and televised political debates. The language of public political debate, therefore, has given up more and more ground to *dārža* (Sayahi, 2014). Since the revolution, the rise of *dārža* in public Maghrebi spaces, both virtual and real, is also due to the pride in one's 'Tunisian-ness', which arose after Tunisia initiated the uprisings in the Arab world: if fuṣḥā is considered to be the 'real' Arab language, *dārža* is the symbol of the Tunisian national identity as opposed to the rest of the Arab world (Sayahi, 2014). Caubet (2018) reports that a Moroccan blogger called Mohamed Sokrate, who used to write in MSA, decided, after being released from prison detention, to start writing in *dārža* in 2014, explaining this choice as a way to get closer to all Moroccans. However, there are also different attitudes, such as that of young people who intentionally write long posts on Facebook in fuṣḥā and in Arabic characters. Thurlow (2006) also reports the attitude of 'language puritans', who see texting practices as having a negative influence on standard writing. This all forms part of the identity debate that has exploded in the wake of the revolution, and in this sense, depending on one's points of view, even the use of fuṣḥā can be a way to regain possession of part of one's own identity, i.e. the Arab part. However, *dārža* not only represents this sense of identity, but is also a way to encourage the diffusion of culture, making sure that high culture does not necessarily mean the use of French, as expressed in the Association Derja presentation in Section 2.[63] Laroussi (2010) makes an interesting observation about the Maghreb language policies, saying that these never corresponded to reality, because there is no democracy in public life and the process of language man-

[63] As in 1, Francophonie, in public debates since the revolution, is proudly flaunted by some and rejected by others, in a picture that is, as always, contradictory. Expressing oneself in French instead of Arabic, or vice versa, can be traced back to ideological choices, e.g. Islamists prefer Arabic, the Tunisian elite prefer French, etc.

agement was not based on the consensus of social groups, dialogue and inclusion. Today, perhaps, this change is finally starting to take place. Therefore, in the case of Tunisia, social change consists of both the revolution and the advent of the Internet and social networks, two phenomena that are seen to be strongly interconnected. Both phenomena are an engine for linguistic change, and in this regard, Crystal (2004) states that:

> '[T]*he differing expectations, interests, and abilities of users, the rapid changes in computer technology and availability, and the rate at which language change seems to be taking place across the Internet (much faster than at any previous time in linguistic history) means that it is difficult to be definitive about the variety's characteristics.*'

Similarly, Heath (2018), speaking about English orthography on social media, argues that:

> '*With the popularisation of social media and the sheer number of users participating in online conversation daily, social media has become a mecca of rapid language change and standardisation.*'

This phenomenon can also be seen in Tunisian Neo-Arabic, where traditionally spoken language varieties are being written in similar ways because they are used in online spaces. Moreover, this is also a mass daily practice. Crystal (1994) had already posed the question of how to categorise this new type of communication featuring hybridised text including an informal style of speech combined with features of written texts. Baron (1998), for example, consider e-mail writing to be a creolising blend of written and spoken, being the product of new linguistic modalities. Androutsopoulos (2011); Androutsopoulos and Schmidt (2002); Jaffe et al. (2012), when discussing *greeklish*, which is the Greek representation in Latin script employed in CMC, use the term 'neography', which had already been coined by the French linguist Jacques Anis (2007). This hybridisation can also be seen in the way that Arabizi tends to mirror Tunisian Neo-Arabic phonology by its approximation, while also including features, such as numbers, selected by analogy with the graphemes of the Arabic alphabet. Regarding this particular *dārža* encoding, i.e. the Arabizi system, Younes and Souissi (2014) collected a corpus of 85,000 Tunisian messages (approximately 37,000 were Facebook posts). 43,222 messages were in *dārža* or written in Arabizi, equating to more then the half. Regarding Facebook posts, 81% were written in Arabizi. It goes without saying that this spelling system

needs to be a part of the subject of this study, and as such, the next section being dedicated to it.

### *Digraphia*

The term digraphia is used to refer to the use of two different graphical systems to encode the same language. Androutsopoulos (2012) explains that he addresses the *greeklish* phenomenon through both an autonomous and ideological approach, specifying that an *autonomous approach* sees orthography as a *neutral technology for the representation of spoken language*, in contrast with the so-called *ideological approach* that views orthography as a *set of social practices in specific social and cultural contexts.* Besides orthographic choices for specific symbols, script choices can also be motivated by identity distinctions (Grenoble and Whaley, 2005; Miller, 2017).[64] The invention of writing, despite occurring relatively recently in the history of the human species, revolutionised the way language can be used. As introduced by the linguist Ferdinand de Saussure in the early days of modern linguistics, a linguistic sign is a link between a concept (signified, from French *signifié*) and a sound pattern (signifier, from French *signifiant*), which is based on socio-cultural arbitrary conventions, and while it is perhaps impossible to deliberately change the laws of sound, other aspects of language are open to deliberate modification and innovation. One must also consider the well-known distinction between semasiographic, representing ideas, i.e. the mathematics language, and glottographic systems, representing elements of a specific spoken language (Sampson, 1985, 26-45).[65] Finally, within the category of glottographic writing systems we can identify logographic (i.e. based on morphemic or polymorphemic units, such as words) and phonographic writing systems and/or strategies (that reproduce a phonic segment, e.g. a phoneme). In brief, writing is a social practice and a mode of communication in its own right (Coulmas, 2013). Linguistic behavior is in all circumstances a matter of choice, just like the role one takes within a specific situation. As such, it is necessary to take the

[64] For example, Coulmas (1999) points to the fact that so many groups have developed their own scripts, syllabaries in particular, as evidence of the importance of a script as a marker of identity.

[65] Among the semasiographic systems we can consider also the programming languages.

power of written language as a political tool into consideration, for example, the Rosetta Stone[66] or the Code of Hammurabi, the former of which contained a Ptolemaic decree and the latter a series of dispositions of the king of Babylon.

Both codes are, moreover, engraved on stone, a further affirmation of the strength and power of the written code in a socio-political context, and the object of re-appropriation by the masses in modern times through the practice of graffiti on walls, including public expression of political opinions and criticisms of society, normally prohibited by law (Akbar, 2019). The practice of writing on walls, in the globalised era, is flanked by writing on a type of virtual wall, the visibility of which assumes an unquestionably greater scope, that is, the Facebook wall, which is referred to in English as 'wall', in Spanish as 'muro', in French as 'mur', and in Arabic as لوحة الحائط, 'wall table' or 'wall frame'.[67] Another important factor to note is that writing produces changes in both language and society, making, among other things, diachronic variation visible and thus bringing it to the attention of researchers (Coulmas, 2013) and to the sight of the native speaker himself, who, by observing his own product, has more opportunity to develop linguistic thoughts, which is a preliminary and indispensable step for the transition to the orthographic standardisation of a system which was previously only oral (Mion, 2017a). In fact, it is no coincidence that in recent years, in the context of Arabic Natural Language Processing (ANLP), Tunisian has acquired a continually increasing visibility. Indeed, the enhanced availability of written texts has made it easier to collect data for lin-

[66] Interestingly, the Rosetta Stone shows an inscription in three different scripts: hieroglyphics, demotic and ancient Greek. The first and second of these are two different encodings of the same language, with ancient Egyptian that was encoded in hieroglyphics being limited to priestly functions or inscriptions, while the demotic was more widely understood and used.

[67] As with the practice of graffiti, the use of Facebook can also be subject to control, either among peers or from above, especially in dictatorial contexts. Regarding the latter, sometimes control is 'limited' to the prohibition of expressing an opinion contrary to that of the dictator, as in pre-revolution Tunisia. In other contexts, Facebook is downright illegal, as in the People's Republic of China.

guistic analysis.[68] As seen above, even in the choices made by the Association Derja, there are two spontaneous encoding possibilities for Tunisian; one in Arabic characters and one in Latin characters. As previously mentioned, this phenomenon is also known in other languages, the orthographic systems of which are not Latin-based, such as Greek (*Greeklish* (Androutsopoulos and Schmidt, 2002)), Farsi (*Pinglish* (Babaei, 2022)) and Serbian (*Latinica* instead of the Cyrillic alphabet (Ivković, 2015)). However, in terms of Tunisian, neither of these systems was created for encoding Tunisian, and the first was in fact created for encoding Arabic. Consequently, graphically representing Tunisian Neo-Arabic with either system presents some difficulties. The most obvious difficulty in using an Arabic-like spelling for Tunisian is that of code-switching, where two possibilities are given to the native speaker:

1. Change the input system only for code-switching elements;
2. Encode code-switching elements in Arabic characters.

The first choice presupposes a type of challenge for the author of the text. In fact, anyone who has ever written even a single Arabic word in electronic text that involved both writing systems knows that the formatting of the text quickly goes out of control. The second possibility instead represents a reality which is actually practiced through the approximate transliteration of foreign words into Arabic characters.

This phenomenon opens the doors to curious hybrid forms (henceforth *hybridisms*) due to morphological integration of a lexical transfer (following Regis (2005) point of view) or to mechanisms oriented towards a loan adaptation, that Poplack and Meechan (1998) define as *nonce borrowing*. Hybridisms are words made up of a lexical morpheme from an X language and an inflectional morpheme from a Y language (Regis, 2010, 622). As an example, the French word 'restaurants' can adapt to the morphological rules of Tunisian becoming رستورانات, /rəstūrānāt/, with the Tunisian plural feminine suffix. However, for some words, the understanding may not be immediate. The mechanism behind these formations, and their definition, is subject of an open debate, that can see this phenomenon as coming close to the phenomena of code-switching

[68] Younes and Souissi (2014) showed that the Facebook usage rate in Tunisia was around 97%. YouTube monopolised second position (1.3%), while Twitter took third (1.01%).

(CS) or loan adaptations, without however coinciding completely with either. In fact, unlike Myers-Scotton (1993),[69] Cerruti and Regis (2005) and Berruto (1995) consider that it is better to keep hybridisms separate from CS phenomena, involving only the superficial linguistic system (i.e. words, morphemes and phonemes) and not the discourse, although they are non-institutionalised manifestations of contact in use as much as CS (Cerruti and Regis, 2005, 193-194).

Another of the reasons that favors the input system in Latin characters instead is that, for reasons that will become apparent in the next section, native Arabic speakers are faster at writing in Latin characters (at least on computer keyboards). This consideration is supported by the emergence of Arabizi-based text entry support systems.[70] These systems are still used today, despite the fact that there are now highly developed Arabic character input systems.

Definitions for CMC-related spontaneous spelling systems, such as *Greeklish* and Arabizi, have been proposed by scholars, among which, as reported by Androutsopoulos (2011), there is *neography*, which was coined by the French linguist Jacques Anis (2007), *graphostylistics* and *respelling*. The following section examines Arabizi in this context.

### *Arabizi encoding*

The phenomenon of written Arabizi was born spontaneously at the end of the 1990s, following the arrival of the first electronic devices, to compensate for the lack of Arabic keyboards or input systems that allowed typing in Arabic characters. The name Arabizi seems today to be the most popular, together with 'Arabish', which comes from a mixture of the words 'Arabic' and 'English', which is the second most popular (Bianchi, 2013). The former probably comes from the word 'Arabic' and the Arabic word for 'English': /ʔinglīzī/ as reported by Alghamdi and Petraki (2018), or from the merge of the words 'Arabic' and 'easy' as suggested by Caubet (2019), who also mentions the names 'e-darija', '3aransiya', 'franco' or 'franco-arabe' for Moroccan

[69] See Section 2 for a description of the Matrix Language Frame model developed by Myers-Scotton, who consider these forms as inter-sentential code-switching.

[70] As we can see from this blog which sponsors the use of Yamli, one of the most popular of these input systems. `https://alaashaker.wordpress.com/2009/03/14/arabic-text-driving-you-crazy/`. Consulted on 4th April 2021.

Arabizi, which are mixtures of the Arabizi encoding for the 'Arabic' and 'French' languages: '3arabiya' and 'faransiya' (Alghamdi and Petraki, 2018; Caubet, 2012, 2018; Yaghan, 2008). Caubet retraces the progressive and massive passage from *dārža* to the written 'Do It Yourself' practice, which she describes as:

> '*une action spontanée et collective* [...] *d'acquisition de la lecture et de l'écriture d'une langue non-codifiée* (Caubet, 2019, p. 391).'[71]

The name Arabizi was mostly used in eastern Arab countries, but in the wake of the release of the *Arabizi* movie, directed by Dalia Al-Kury, in 2006, became widespread everywhere. Recently, Fourati et al. (2020) proposed a specific name for Tunisian Arabizi, i.e. *Tunizi*, coinciding with the dataset they collected for Tunisian Arabizi sentiment analysis. There have been a few *stigmatisers* of Arabizi encoding, with some considering it to be a *deviant* type of encoding that could compromise expertise in Arabic-character encoding (Alghamdi and Petraki, 2018). A number of tools have also been created to facilitate online Arabic writing, such as Yamli[72] or Microsoft Maren,[73] which are smart Latin character keyboard input systems for Arabic, that allow for the conversion of text into Arabic characters through the user's selection of the desired word from a range of possibilities listed in a drop-down menu. Today it is possible to install the Arabic keyboard on any 'smart' device. However, there is still a large number of Tunisian social network users who still prefer Arabizi. The most likely reason for this is that Arabizi is easier, faster and more flexible than Arabic encoding, but also because it is considered cool and stylish by young CMC users in Saudi Arabia, according to interviews by Alghamdi and Petraki (2018) during their sociological survey on Arabizi beliefs.[74] Another reason is the already acquired familiarity with the Latin keyboard, as Bou Tanios (2016) reports with regard to the Lebanese context. Despite Facebook's meteoric rise in

---

71 'a spontaneous and collective action [...] to acquire the reading and writing of a non-coded language'. My own translation.

72 Yamli's Smart Arabic Keyboard was launched in November 2007: `https://www.yamli.com/fr/`. Consulted on 4th April 2021.

73 Here is a link for Microsoft Maren's tutorial: `https://www.youtube.com/watch?v=IWbYGTxy5x4&ab_channel=WaelKabli`. Consulted on 4th April 2021.

74 It would be worth investigating the motivations behind this sociolinguistic connotation of the Arabizi script. Certainly, the positive evaluation associated with this writing system is related to the youth environment, but it would be interesting to investigate the specific implications.

popularity since 2006, and the arrival of Twitter in the same year, the Arabic version of Facebook did not arrive until 2009, with the Twitter version being completed in 2012 (Alghamdi and Petraki, 2018). As shown in Salem (2017), Facebook has had a huge impact on Tunisian society in comparison to the rest of the social networks, such as Twitter, Linkedin and Google+. Tunisia is, in fact, the third most active Arabic country on Facebook, considering it a daily activity. Twitter is less widely used, reaching only 2%.[75]

In conclusion, the data shows that the preferred solution for Tunisian nowadays is still the Arabizi system, which is considered to be a *neography* of Tunisian, that can be isolated in *diamesia*. Regarding the transliteration of terminology, Gorgis (2010) distinguishes the romanisation from latinisation stating that transliteration, as the conversion of one writing system into another, is a generic term, while romanisation is more specific, being the representation of a writing language system in a Romance-language script. However, in contrast, latinisation also concerns the use of Latin vocabulary. Concerning transcription in the Roman system, the source language pronunciation is represented by finding the closest equivalent in the Roman system, while Arabic romanisation is more a transcription process considering that, for example, short vowels are also encoded.[76] This is particularly true in terms of Neo-Arabic romanisation, which is the encoding of oral systems into Roman-based conventional encoding. As such, we will henceforth use the term romanisation to refer to the Arabic-romanisation system. Finally, Arabizi seems to be a process which falls somewhere between romanisation and creative transliteration, basically being a non-standard encoding of Tunisian phones through Latin graphemes, where for Tunisian phones which do not match the Roman alphabet, the substitution is graphical-analogical, i.e. a graphemic substitution based on an iconic similarity between Arabic letters and Arabic numbers (such as '7' for the Tunisian phoneme /ħ/).[77]

---

[75] The same tendency is also confirmed by Caubet (2019) with regards to Moroccan society.

[76] When in Arabic script, short vowels correspond to diacritics.

[77] Blau (1981) reports that, as in Judeo-Arabic script, *[a]s far as possible, Arabic letters were marked by the phonetically corresponding Hebrew letters, including letters denoting allophones which phonetically resemble Arabic phonemes. When, however, no correspondence between Arabic and Hebrew existed, perforce the Arabic orthographic method is applied.*

| *Transcription* | *Tunisian* | *Arabizi* | *Transcription* | *Tunisian* | *Arabizi* |
|---|---|---|---|---|---|
| /a/ | ة | a, e, h | /ā/ | ا ,ى | a, e, é, è |
| /ʔ/ | ء | 2 | /ð̣/ | ظ , ض | dh, th, d |
| /b/ | ب | b, p | /ṭ/ | ط | 6, t |
| /t/ | ت | t | /ʕ/ | ع | 3, a |
| /θ/ | ث | th | /ɣ/ | غ | 4, gh |
| /ž/ | ج | j | /f/ | ف | f |
| /ḥ/ | ح | 7, h | /q/ | ق | 9, q |
| /ḫ/ | خ | 5, kh | /g/ | ڤ | g |
| /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 |
| /š/ | ش | ch, (sh) | /ū w/ | و | ou, w |
| /ṣ/ | ص | s | /ī y/ | ي | i, y |

*Table 3. Arabizi code-system for Tunisian Neo-Arabic*

In some infrequent cases, we can see complex graphemic substitutions, such as @ employed in the word 'bouss@', *būsāt*, 'kisses', where the *at* sign replaces the feminine-plural-ending '-āt', based on the analogy between the English name of the symbol @ and the phonetic realisation of the morpheme '-āt'. Tunisian Arabizi is therefore a spontaneous encoding, which operates according to some more or less shared principles that involve the use of French spelling where possible, such as in *ch* for ش, or *ou* for و and *ss* for س (Akbar, 2019; Caubet, 2018).

*Some preliminary remarks*

In order to start gaining insight into the Arabizi encoding system, we must first examine some of its specific traits, which will then be the subject of a more detailed study in Chapter 4 through a quantitative analysis of the corpus data.

1. **Representation of Arabizi vowels**
   Tunisian Neo-Arabic presents a wide variety of vowel phones, which are usually ascribed to the phonemes of the tripartite sys-

tem of Arabic /a/, /i/, /u/ and the corresponding long ones (Gibson, 2008). However, with regards to short vowels, further studies have take into account the possibility that Tunisian is a mixed Bedouin-sedentary type, considering the initial phonological distribution, namely the pre-hilalic /ə/ < **a i* ≠ /u/ < **u*, which was subsequently exposed to a process of partial *re-phonologisation* under the Arabic tripartite system.[78] These considerations are supported by the presence in Tunisian Neo-Arabic of both the convergence of /i u/ in [ə] in opposition to /a/ (Caubet, 2000), which is a Bedouin trait, and the sedentary convergence of /a i/ in [ə], in opposition to /u/ (Durand, 2007, 2012; Mion, 2008b). Regarding long vowels, these do not undergo any major change, except in atonal syllables. Moreover, the general rule is that final long vowels, in an atonal syllable, are shortened, while the length is maintained for vowels in a tonic syllable.[79]

One of the main differences between the representation of Tunisian in Arabizi and in Arabic characters lies in the encoding of short vowels, since the latter provides only three diacritics which are rarely used in informal writing. Regarding long vowels, these are better encoded in Arabic-script, even if with a certain inconsistency, as also recorded by Caubet (2019) in the case of Moroccan. In contrast, in Arabizi, same are occasionally encoded through the repetition of the grapheme, for example, *aa*, which can encode both /ā/ and /ʕ/. Previous studies on Arabizi systems of other Arabic dialects presented an inconsistent representation of the vowel system, such as the study on the Kuwaiti dialect developed by Akbar (2019). Akbar also signaled a tendency to delete unstressed vowels, which is also found in the Tunisian system. Some phenomena that are instead visible through encoding in Arabizi, unlike the encoding in Arabic characters, are as follows:

(a) Palatalisation.
(b) Vowel metathesis.
(c) Monophthongisation.

The first of these is intended to be the ʔ*imāla*, namely the shift from /a/ ∼ /ā/ towards /i/ ∼ /ī/, in a not-marked context. The

[78] The neutralisation of short vowels is the process of short vowels gradual flattening through their confluence in schwa [ə]. This phenomenon is characteristic of Maghrebi dialects, especially the Moroccan dialect (Durand, 2012).

[79] The length is restored to abbreviated long final vowels if a suffix is added to the word.

Tunisian /ā/ with ʔimāla sounds like [æː ~ eː ~ ɛː], which is the same for /a/, except for the length.[80] Tunisian Neo-Arabic is characterised by a spontaneous ʔimāla of medium intensity. There are in fact different degrees of /a/ palatalisation in Tunisian. Depending on the phonological context, the speaker provenience and also their gender, ʔimāla can be light (ḫafīfa, towards [æ]), medium (*mutawassiṭa*, such as [e] in the ending of monosyllabic words as /smā/, [ˈsmeː], 'sky', or [ɛ] as in /bāb/, [ˈbɛːb], 'door', where same is particularly heard, being both long and tonic), or strong (*šadīda*), heavily tending towards /i/ or more rarely with the ʔimāla breaking in the diphthong *ie*.[81] This phenomenon is partially evident in Tunisian encoded in Arabizi, considering that palatalised /a/ is encoded through the grapheme 'e', with or without an accent, instead of 'a'. However, unless palatalised /a/ is encoded through a specific grapheme, it is impossible to determine its degree of intensity (Durand, 2007; Mion, 2008b). In Example 1 below, it could be assumed that the grapheme 'è' is unconsciously used to differentiate the only /ā/ present in the sentence, which we can suppose as being a realisation of [ɛː], considering that the same grapheme is used for the same phone in the French vowel system. Concerning the examples, these are taken from the corpus subject of this research and are organized as follows: a sample of the Arabizi encoding is given in italics, followed by a romanisation of the same sentence between slashes, preceding the English translation, which is between quotation marks.

(1) *t'hezzni w tgoud feyya lel nessyèn (malla 9assida)*,
/thəzz-ni w tqūd fə-ya lə-n-nəsyān (malla qaṣīda)/,
'You bring me and lead me through the oblivion (what a poem[!])'.

What is also interesting to note in Example 1 is the use of the apostrophe to separate the *t* from the *h*, disambiguating the possible match of same with the digraph *th*, which is used to encode the phoneme /θ/.[82] This is indeed a relatively common use of the apostrophe in Tunisian Arabizi, diverging from other Arabizi advanced-systems, where the apostrophe is used to represent the

80 ʔimāla in Arabic means *inclination*.
81 For further analyses on ʔimāla, see Mion (2008b).
82 Unlike many Maghrebi and Mashreqi dialects, in Tunisia, and particularly in Tunis, the dental fricatives are preserved (Durand, 2007; Mion, 2010).

dot of Arabic graphemes, such as ɣayn's dot in *3'* (Akbar, 2019; Allehaiby, 2013).

Depending on the phonological context, the ʔimāla phenomenon can be blocked by the *emphatic* consonants (Ḍ)[83] /ṣ/, /ð̣/, /ṭ/, /ḍ/ (Mion, 2010) and the guttural consonants (G), such as the uvular /q/, /ḫ/ and /ɣ/, and the pharyngeal /ħ/ and /ʕ/ in the following phonological situations:

(a) **/Ḍā/** such as in: /ṣāħb-i/, 'my friend', /ð̣āyaʕ/, 'lost, confused' or /sbīṭār/, 'hospital';
(b) **/Gā/** such as in: /qāl, 'he said', /ḫāyb/, 'bad', /ɣāli/, 'dear, expensive', /ħāža/, 'thing', /ʕāyla/, 'family';
(c) **/āḌ/** such as in: /blāṣa/, 'square', /ħāð̣ra/, 'ready' and /šāṭ/, 'to kick'.

It can be observed that Arabizi tends towards the use of a faithful phonetic representation of the /a/ phoneme in these contexts, as shown in the following example.

(2) *Oussama* [...] *sahbii l 4ali rak*
/ūsāma [...] ṣāħb-i l-ɣāli ṛāk/,
'Oussama you are my dear friend'.

With regards to /ī/ and /ū/ vowels followed by pharyngeal phonemes (/ħ/ and /ʕ/) at the end of the word, Mion (2008a,b) reports the appearance of a very short [a], which can be explained as a supporting vowel to facilitate the articulation of the phonetics. The scholar traces this vowel manifestation to a phenomenon which is also present in other languages, such as Hebrew and Aramaic, where the phenomenon is traditionally known as *pataħ furtif*. This phenomenon does not seem to be found in the Arabizi data collected during this research, suggesting that this ultrashort vowel has not undergone a phonologising process. With regards to epenthetic vowels, Durand (2012) asserts that Tunisian follows a Libyan and Near Eastern type pattern, interrupting the /CCC/ sequence to simplify it into /C$^{v}$CC/, a phenomenon which is also found in Arabizi encoding. In the following example, it is possible to identify the apical schwa (ᵊ) as an epenthetic element, whose function is to interrupt the continuity of the consonantal

[83] Defined pharyngeal consonants are also articulated by retracting the root of the tongue towards the pharynx. In Tunis, the pharyngalisation, or *tafḫīm*, is very light (Durand, 2012). The influence of the emphatic on the adjacent vowel in Arabic terminology is called *iṭbāq*.

accumulation between the word boundaries of /kīfā-š/ and the verb that follows it.

(3) *Kifech enajmou nt7aslou 3lik ?*,
/kīfā-šᵊn-nəžžm-u n-tḥaṣl-u ʕalī-k?/,
'How can we reach you?'.

Vowel metathesis and diphtongs, in comparison to the Arabic-script system, are easier to note in Arabizi, even if it is not possible to determine the vowel length. Indeed, /ay/ and /aw/ are *monophtongised* in the intermediate /ē/ and /ō/ in the Hilali dialects and mainly in /ī/ and /ū/ in the pre-Hilali ones. Mion (2008b) reports that in Tunisia, the Hilali confluence, which appeared in the rural speeches of *Sāḥel*, on the eastern coast of Tunisia and in southern Bedouin speech, is maintained by the female speech of Tunis, but has nowadays almost disappeared.

2. **Representation of Arabizi consonants**

Tunisian has a conservative consonant system which is expanded by some loan phonemes, such as /p/ and /v/, from the Romance languages, which, in the graphic system in Arabic characters, are represented as ب and ف, as for the loanword بروفة /brūva/, from the Italian 'prova', meaning 'rehearsal' (i.e. theater rehearsal). The same phonemes in Arabizi tend to be encoded through the same Latin letters, as in the Arabizi word *parfanet* 'perfumes', which is a result of code-switching with morphological adjustment into Tunisian, as in the case of *resturanet*, as seen in Section 2. The phoneme /g/, as well as being an allophone of /q/, albeit with rare examples of minimal pairs, which is typical of Bedouin dialects (as opposed to a /q/ realisation typical of urban dialects), is also a loan phoneme as in the following example.[84]

(4) *7ell gaz wraja3ou yet9lei*,
/ḥəll l-gaẓ w-ṛažžʕ-u yətqla/,
'Turn on the gas and let it fry'.

In Arabic-script encoding the loan /g/ is usually encoded through the غ grapheme, i.e. غاز for /gaẓ/ of the above Example (4), while the /q/ allophone is encoded through ڤ, i.e. بڤرة for /ḅagṛa/, a

[84] It should be specified that there is no such thing as purely /g/ or /q/ speech, but the realisation also depends on the semantic scope or the context of the realisation, as seen in Section 1.

cognate of the MSA word for 'cow'.[85] The glottal stop Hamza /ʔ/ in Tunisian is almost completely absent, with the exception of isolated words and some loanwords from Standard Arabic.[86] In Arabizi, the hamza is usually encoded by the number 2, as in the following example:

(5) *9bal kol chey me jewebni 7ad 3ala sou2eli*,
/qbal kull šəy ma-žāwəb-ni ħadd ʕalə suʔāl-i/,
'First of all nobody answered my question'.

As detailed below, the phenomenon of encoding the hamza using the number 2 in Arabizi is not negligible from a quantitative point of view, and in Romanisation, the realisation of the hamza has been reported, even if it is uncertain. In fact, the encoding of this consonant is rather unstable. As expressed above, it can disappear causing a lengthening, or can even be replaced by another consonant, the /h/ (Durand, 2007; Ouerhani, 2006). Encoding hamza as a '2' in Arabizi could be a form of hyper-correction to a more widespread and 'conventionalised' graphic form of this consonant. Other likely forms of hyper-correction will be expressed in Example 15.
The general trend in Arabizi encoding seems to be partly oriented towards phonetic representation and partly driven by the need to make the text comprehensible by reconstructing an orthography that is in some way faithful to the root of the word. This is also true for the consonantal changes that frequently occur in Tunisian words, which are caused by phonemes overlapping and their mutual influence, such in assimilation or emphasis, or emphasis blockage. As an example, Stumme (1893) states that '/ž/ and /z/ never fit together in the same word'.[87] The same happens when /ž/ meets /z/, or another /ž/ or /s/. This is the reason why, for example, in Arabizi we can find the mixed yoghurt of the well

[85] However, this is not to be intended as a regular enconding practice.
[86] The hamza vanishing generally produces vowel lengthening of the corresponding vowel or consonant gemination. With regard to the hamza vanishing from the end of the word, this results in compensation lengthening which will also carry the word's accent.
[87] 'ž *und z vertragen sieb nie in demselben Worte'* (Stumme, 1893).

known brand *Danone*: 'Délice Mamzouj' written as in the following example:[88]

(6) *delice mamzouz*,
/delice məmz*ūz*/,
'delice mixed'.

However, as shown in Table 3, the main issue in a representation of consonants which is faithful to the phonetics concerns the Arabizi encoding of the emphatic phonemes, such as /ð̣/, /ṣ/, where for example, the Arabizi out-of-context-word *sa7a* could encode both the word /ṣaħħa/, 'health', or the word /sāħa/, 'square'. Regarding /ṭ/, when Arabizi began to spread, it was generally encoded with the number '6' to differentiate it from the voiceless dental plosive /t/.[89] As will later be explored in Chapter 4, nowadays, only the numbers 5, 3, 7 and 9 have a stable presence in this type of encoding, which respectively encode /ẖ/, /ʕ/, /ħ/ and /q/. Regarding the latter, as explored below, same is rarely replaced by the 'q' grapheme, which is instead associated with the /k/ phoneme. Indeed, Arabizi seems to be more oriented towards a phonetic representation of Tunisian, especially considering the following phenomena:

a. Consonant gemination *(Tašdīd)*;
In Tunisian, all consonants can be either simple or tense (geminated). Gemination can occur if the consonant in question is preceded by a short vowel and followed by a long or short vowel. In phonetic realisation however, if the tense consonant is preceded by a long vowel, the duplication often disappears, for example, in the active plural participle of the verb هزّ, to take: هازّين /hāzzīn/ can be pronounced as /hāzīn/. Gemination can be also missed in rapid phonetic realisation (Stumme, 1893). This is particularly true if the geminated phoneme is not followed by a vowel. Such cases are also frequently found in Arabizi representation, as in the following example:

---

88 With regard to the /ž/ phoneme, Tunisian, consistently with Maghrebi dialects, presents a fricative [ʒ] and not affricate [dʒ] realisation (Durand, 2007).

89 The use of the number six to encode the pharyngeal /ṭ/ can be observed in the reference name /baṭṭāl/, i.e. 'unemployed', for the following Tunisian Facebook page: `https://www.facebook.com/ba66al/?ref=page_internal`. Consulted on 4th April 2021.

(7) *hatha el kol mel 7oub hhhh*,
/hāða l-kull m-əl-ħubb hhhh/,
'All this because of love ahah'.

b. Consonant assimilation;
Assimilation is a phenomenon that occurs when a phonological segment modifies the preceding segment or the following segment. This phenomenon may involve the dental phoneme /t/ followed by other dental or sibilant phonemes. The groups [/t/+dentals] or [/t/+sibilants] are definitely not well tolerated in Tunisian, which has been resolved by the assimilation of the /t/, as in ما تتذكّرش؟, 'don't you remeber?', which has (slowly) been realised as: /ma-tððəkkər-š/. If gemination, resulting from an assimilation, is found at the beginning of the word, it is not heard in the phonetic realisation unless the /t/ is supported by the insertion of an epenthetic vowel. This is also the case in Example 8, where the passive morpheme /t-/ should not be heard due to it being at the beginning of the verb /ðkər/, 'to mention'; however, it is represented in Arabizi. There are also other types of not-represented assimilations in final or intermediate positions of the word, as in the case of the Arabizi word 'jedti', which is realised as [ˈʒətːi], 'my grandmother', in the first word of Example 9.

(8) *ethika leblassa elli t'thakret fi souret el ka7f*,
/hāðīka l-blāṣa əlli t-ðəkrət fī sūrət-əl-kahf/,
'This is the place mentioned in the Surat Al-Kahf'.

(9) *jadti mel om tbajal wled wledha 3lina*,
/žədt-i m-əl-ūmm tbəžžəl wlād wlād-ha ʕalī-na/,
'My maternal grandmother prefers her children's children to us'.

c. Consonant assimilation: the definite article.
With regard to article assimilation, this normally occurs when followed by a coronal consonant, which are consonants articulated with the tongue blade raised towards the dental, alveolar, or palato-alveolar region.[90] In Arabizi, this is easily identifiable (Gugliotta et al., forthcoming), as in the following example:

[90] In Tunisian Neo-Arabic, these are /t/, / θ/, /d/, /ð/, /r/, /z/, /s/, /š/, /ṣ/, /ð̣/, /ṭ/, /l/, /n/ and /ž/.

(10) *Inchalah cycle ejjay wala eli ba3dou*,
/nšāl̪la *cycle* əž-žāy walla əlli baʕd-u/,
'God willing next time, or the time after that'.

As can be seen in Example 10, in Tunisian, as in other Maghrebi dialects, there are new consonants which are subject to the *taf̪ḫīm* phenomenon, whose pharyngalised (or emphatic) realisation results in /l̪/, /m̪/, /r̪/, /b̪/, /z̪/. Coming back to the topic of article assimilation, this may not appears in Arabizi encoding, mostly in the cases where the article and the defined name (starting with a coronal phoneme) do not form a single graphical compound and there is white space between them. These cases generally occur when the article is graphically joined to the preposition, as in the following examples:

(11) *wa9t yji nghar9ou bil dlel*,
/waqt yži nɣarq-u b-əd-dlāl/,
'When he comes I drown him with vices'.

(12) *entouma yelzemkom derss fil jografia*,
/ntūma yəlzəmkum dərs fī-ž-žuɣr̪āfya/,
'You(.pl) need a geography lesson'.

This mode of writing hinders the representation of assimilation, because the article, when separated from the noun it defines and linked to the preposition that precedes it, is always represented as 'l'. In addition, Tunisian Arabizi seems to prefer the formation of a compound between prepositions and articles separated from the noun, rather than a compound formed by the three (see Chapter 4). Conversely, the article tends to appears in its full form as a unique token when the determined name does not begin with the coronal phonemes that generate assimilation. Two examples are given below, the first one (13) shows a preposition, the article and a noun not beginning with a coronal phoneme, and all of these elements are graphically isolated. While, the second one (14) shows a similar case, this time not involving a prepositional sentence, but only a nominal phrase, where the noun starts with a loan phoneme (/p/), being the noun a hybridism from the English word 'penalty'.

(13) *nebki min il far7a*,
/nəbki mən əl farħa/,
'I cry for happiness'.

(14) *ki zayech dhaya3 el pilanti*,
/ki zayyāš ḍayyaʕ əl-pīlānti/,
'When Ziyech missed the penalty'.

This is a hyper-correction practice, which can be connected to the Tunisian ideological question of the perception of the writing system, with regards to the Arabic-script-encoding of 'CCC' clusters or words starting with a '#CC' pattern, which is not allowed in Standard Arabic. Since the Arabic script is the encoding system used for the acrolect, using it seems to activate in the writer a form of *respect* for the prestigious Arabic's norm. Such hyper-corrective tendencies bring us back to the earlier discussions on diglossia (Section 1), and in particular to the perception, in this case, of one orthographic system as being more prestigious than another as a system which is habitually dedicated to the encoding of the prestigious variety. Perhaps the same cognitive process that led to the writing of medieval texts in that form of Arabic mixed with elements of Standard Arabic and Neo-Arabic and pseudo-correct features, which Blau (1981) defines as Middle Arabic, in this context, leads the author of the written text to produce an encoding which is not so much of transcriptive type, but is more transliterative, having the orthography of Arabic characters as a reference source system. The hyper-correction, in this case, consists in representing the cluster below with opening epenthetic vowels (codified through the ʔalif grapheme), in order to transform it in a different cluster (#VCC) which is easier to realise. This phenomenon is much less commonly found in Arabizi, but is still present, as in the following example, where an epenthethic vowel can be seen at the beginning of the word *tkūn*:[91]

(15) *t'es dans une ces regions donc yelzmek etkoun men mdina men ces regions alors que t'es mehdwi*,
/*t'es dans une ces regions donc* yəlzmək tkūn mən mdīna mən *ces regions alors que t'es* məhdwi/,
'You are in one of these regions so you should come from

[91] This phenomenon has already been encountered in Example 14 in the Arabizi word *entouma*, which can be traced back to the Tunisian-Arabic-script form انتوما.

a city of these regions, that should make you a Mehdia citizen'.

## 3. **Arabizi code-switching**

This chapter has documented how the variant of the Tunisian dialect, which has emerged as the national variety, coincides with that of the capital. It has also documented how this urban variety represents, at the diastratic level, a high social stratum with access to a French cultural and linguistic background, being rich in French code-switching (CS) elements. The definition of code-switching adopted here is that of Myers-Scotton (1993), who states that the term is used to refer to alternations of linguistic varieties within the same sentence (intrasentential).

> '*CS is the selection by bilinguals/multilinguals of forms from two or more linguistic varieties in the same conversation.* [...] *Stretches of CS material may be inter-sentential (switches from one language to the other between sentences) or intrasentential (within the same sentence, from the single morpheme level to higher levels).*'

Considering the two kinds of *asymmetry* involved in code-switching, i.e. the structural and the content asymmetry, Myers-Scotton (2006) explains that, in a bilingual context, one language supplies the main grammatical framework for a clause containing words from both languages. The scholar defines this structural-driver-language as the *Matrix Language* (ML), while the content-driven-language is the *Embedded Language* (EL). As mentioned earlier, from the perspective of Myers-Scotton (1993) this can be seen as in the previous example of Tunisian *re-morphologisation* of the French word 'restaurants', *rəstūrānāt*, where the ML coincides with Tunisian Neo-Arabic, while the EL coincides with the French language.[92] The ML not only governs the form of a word selected from the EL lexicon, but also governs the structural relationships between words within the sentence. A code-switching model for bilingual speech is the Matrix Language Frame Model (MLF), while a model specialised in the morpheme types is the 4-

[92] We considered this to be Tunisian Neo-Arabic and not Standard Arabic for two main reasons: the first one being the specific lexicon, as Standard Arabic has its own widely-used term to express the concept of restaurants; the second reason is the *ʔimāla* applied to the female-ending morpheme, which, even if it also applied to Standard Arabic, is mainly a characteristic of non-Standard Arabic phonetic realisation.

M model (Myers-Scotton, 2006). At the same time, other scholars, such as Berruto (1995); Poplack and Meechan (1998); Regis (2005), consider these hybrid forms of words as phenomena more analogous to lexical borrowing. With regards to MLF, it makes a distinction between *content morphemes* and *system morphemess*, where the first are those which assign or receive thematic roles,[93] while the second of the two are basically functional words, even if the two categories do not overlap completely. The 4-M model refines the MLF model by dividing system morphemes into three types:

* *Early system morphemes*, morphemes which are conceptually activated by a speaker's pre-linguistic intentions. As an example, the Tunisian determiner for the French word 'famille' was chosen in the following sentence, and this phenomenon is very common (Tarquini, 2019).

(16) *hatta el famille walit j evite tout le monde*,
/ħatta l-*famille* wallīt *j'evite tout le monde*/,
'Also the family I started to avoid everybody'.

The other two systems are defined as being 'late', that is to say, they are not activated until a first production level.

* *Bridge late system morphemes*, which are activated to fully form the relationship between elements of a syntagm. As an example, the following sentence, as well as showing the Tunisian determiner for the French proper name Cap Bon,[94] shows the pseudo-preposition /mtāʕ/, which is used for the possessive analytical construction, followed by a French indefinite nominal syntagm *une zone*. It should be noted that the indefinite French article is used, instead of the Tunisian one, which is ø.[95]

(17) *S7i7, ama essa7el fi tounes hiya tasmya mta3 une zone, kima el cap bon...*,
/ṣħīħ, āma əs-sāħəl fi tūnəs həya təsmya mtāʕ *une zone*, kīma əl-*cap bon...*/,

[93] Also called θ-roles, which are mainly carriers of semantic cores.

[94] Cap Bon in Arabic, كاب بون, or *Rās əd-Dār*, which was known in antiquity as the Cape of Mercury, or *Promontorium Mercurii*, is a cape and peninsula at the northeastern tip of Tunisia, i.e. the northwestern end of the Gulf of Tunis (Shaw, 1757).

[95] As seen in Table 2, the use of the analytical construction of genitive phrase is an innovative pre-Hilali trait.

'That's right, but the Sahel in Tunisia, is a nomenclature of a zone, as the Cap Bon'.

* *Outsider late system morphemes*, which are activated to indicate relationships within the clause, outside of its immediate constituent. The example features the agreement of the verb *ṛāpa*, from the French *rapper*, with the first person singular subject.

(18) *Nrapi manich conscient f les chansons Njib zéro vues ama naj7 b mention*,
/nṛāpi ma-nī-š *conscient* fi *les chansons* nžīb *zéro vues* āma nāžaḥ bi *mention*/,
'I do rap without being aware of the songs, I get zero views, but I obtained a mention'.

Further phenomena of contact with the French language at the orthographic level can be observed in terms of the '-e' which has been added at the end of french loanwords, as in the word /ʕars/ in the following example (Durand, 2012).

(19) *pour tous vos èvènement lamet enfaset 5outba 3arsse anniversaire je vous propose des petits fours*,
/*pour tous vos événements* lammāt nfāsāt ḫuṭba ʕars *anniversaire je vous propose des petits fours*/,
'For all your events: meetings, puerperia, reunions, engagements, weddings, birthdays, I propose you small pastries'.

## 4. **Arabizi extralinguistic features**

Typical characteristics of a Digital Networked Writing (DNW), according to Thurlow and Poff (2013), are *Typographical-cum-linguistic devices*, such as onomatopoeia (i.e. *hahah*) or exaggerated use of spelling and punctuation (i.e. *no moooore!!!!*), pragmatic use of capitalization, such as *STOP IT.* (Heath, 2018), spacing, and special symbols for adding the missing prosody and emphasis impact which is obviously missing in an DNW text. All of these are also found in Arabizi. The only *devices* less represented in Tunisian Arabizi DNW are abbreviations, which are usually consonant clusters (i.e. *thx*, meaning *thanks*), or clusters of numbers and letters defined by Androutsopoulos (2011) as *rebus-like spellings* (i.e. *2l8* having the meaning *too late*), and acronyms, which are words made from the initial letters of other words, such as *IRL*, meaning *In Real Life*, or *rotfl*, meaning *rolling on the floor laughing* (Androutsopoulos, 2011; Crystal, 2004). Shortis (2007) suggests that the non-standard orthography of texting almost certainly expresses the generally creative, playful and friendly tone intended by texters.

This can also be detected in Tunsian DNW, where the playful tone is often realised by wordplay or the use of spelling and punctuation. Regarding the lack of abbreviations, these would probably hinder the comprehension of the written text, which, despite not being yet standardised, already presents a certain degree of difficulty on the level of comprehension. Indeed, it is also important to notice that CMC users have not only developed an individual writing style, but that different web-based channels also require different codes and metapragmatic awareness of the choice of written style.[96]

Arabizi is an interesting phenomenon in itself, but at the same time can also be a useful tool for the study of Tunisian in general, allowing for a study of a large amount of data. For this reason, it was chosen as the encoding for the Neo-Arabic Tunisian texts collected in the Tunisian Arabish Corpus, the building methodology of which will be described in Chapter 2.

[96] In fact, one of the future studies we would like to carry out, on the advice of Professor Carmela Perta, whom we thank, could be a study of Arabizi encoding according to a functional pragmatic approach.

# 2. A TUNISIAN DIGITAL NETWORKED WRITING CORPUS

## 1. Introduction

The development and diffusion of the Internet, together with that of technology, has transformed written communication, leading to the emergence of a new specific mode: electronic written communication. This new form of communication, which has spread rapidly since the mid-1990s, has also raised many concerns. Technology in general has had a great impact on our daily habits, including the way we read the news, the job market and careers, the way we socialise and the way we express our identity. Technology has also had a major impact on language use and the way we relate to each other. In fact, research which deals with this phenomenon ranges from discourse analysis with approaches from the communication sciences to psychology and sociology, and the methodological way in which same is analysed is comparative with that of other discourse modalities or that focused on specific linguistic phenomena. This second chapter addresses this particular topic, which has carved out its own space within linguistic studies over the last three decades. Section 2 will offer a general definition of Digital Networked Writing, while the following sections are intended to describe the methodological choices made with the aim of building a representative and balanced corpus of Tunisian Arabizi, in both the fields of Corpus Linguistics (Section 3) and Deep Learning (Section 4). Regarding these

fields, the main strategies and principles that govern the art of building linguistic corpora and the art of building artificial intelligence tools with Deep Learning techniques in order to allow machines to handle linguistic problems will be discussed. Finally, the last section (Section 5) presents the state of studies dedicated to Arabic Natural Language Processing (ANLP), in particular regarding Tunisian in both scripts, i.e. Arabizi and Arabic characters.

## 2. Defining Digital Networked Writing

Electronic communication refers to a form of communicative exchange where messages are conveyed by electronic systems, i.e. based on the combination of informatics and telecommunications, from mobile phones or computers, through the Internet. Electronic communication is therefore a generic term that encompasses many different types of communication situations, whether oral or written, private, public or semi-public, messaging or computer-mediated, synchronous or asynchronous, etc. Regarding written communication which is mediated by computer, this is often defined as Computer-Mediated Communication (CMC). Nowadays, given the spread of smartphones, Communication Mediated by Telephone (CMT) is also included within CMC, despite previously consisting of a distinct category. In fact, the distinction is no longer very significant today in light of the arrival of web 2.0 (in the late 90s), and electronic communication can instead be subdivided according to the different channels and scopes of communication, and, in the case of these research aims, by writing. The mode of communication of blogs, compared with applications designed for discussion, such as forums or those for instant messaging, which may also include Facebook, certainly has its peculiarities. A blog is a kind of diary which is shared on the Internet, in which texts, of no predetermined length, are ordered chronologically and can be enriched by multimedia elements, such as photos, videos, and audio. It is possible to interact with them, but it is an asynchronous form of communication. Forums are instead generally organised by topics and require registration. Users are not required to provide their true identity and may use pseudonyms. However, it is not unusual to share a sense of belonging and friendship with forum users, who often find themselves connected simultaneously or discussing shared issues. This possibility has lately been denied to Facebook users, who for security reasons are now being asked to prove their

real identity (Ziamari et al., 2020). Communication on the Facebook platform mainly falls under one of two types, asynchronous and synchronous, and is carried out either through long posts, or through comments on others posts, which can be public or private, and addressed to users who are part of your circle of friends, to a specific group or to anyone who is part of the Facebook community (in the case of completely public profiles). With regards to the name Digital Networked Writing, this refers to writing practices in digital environments. Schmitz (2001) distinguish four different kinds of online written communication:

1. Monologic;
2. Dialogic;
3. Non-linear;
4. Interactive.

Here, non-linear communication refers to hypertextuality, the networking function of new media that allows a large quantity of information to freely move around within a series of interconnected nodes in the network. A number of terms has been created to highlight the specificity of different electronic communication practices, such as *electronic communication* by Murray (1989), who examines turn-allocation techniques in non-linear organisations of CMC discourse. Panckhurst (2006) uses two French expressions, based on the neologism *médier*: *communication médiée par ordinateur* and *discours électronique médié*,[1] explaining that they were adopted from reflections on the role of the instrument during the communication process, referring to the studies of Vygotsky (Bronckart and Schneuwly, 1985; Vygotsky, 1985). She indeed justifies the neologism by explaining that the verb *médier* is more appropriate than the verb *médiatiser* because electronic communication is truly mediated (modified) in the Vygotskyian sense of being influenced by the extra-linguistic context, and not merely being broadcast through a medium. Baron (2010) and Anis (2003) also chose similar solutions, respectively: *electronically-mediated communication* and *communication électronique scripturale*. In order to distinguish between the different characteristics of monologic and dialogic communication, scholars have also adopted specific terminology for the latter, such as *computer conversation* by Murray (1989), or *écrit interactive et dialogique* by Anis (1998). However, the more popular name for this

[1] These French expressions respectively mean *computer-mediated communication* and *mediated electronic discourse*.

type of communication in a very broad sense remains *Computer-Mediated Communication* (CMC), which focuses on the medium itself (Herring, 1996). *Netspeak* is the solution proposed by Crystal (2004), as an alternative to *Netlish*, which stresses the concept of an everyday more multilingual *Net*, or to the terms *Internet language* and *cyberspeak*, which focus more on the technological side of CMC. He underlines the fact that:

> *'As a name, Netspeak is succinct, and functional enough, as long as we remember that 'speak' here involves writing as well as talking, and that any 'speak' suffix also has a receptive element, including 'listening and reading'.'*

With regards to classification, Anis (2003) proposed a division of electronic communication types based on instrument properties, such as the available space for the text, the temporal processing of communication (synchronous, asynchronous), the type of readers to whom the message is addressed (individual, group) and indeed the type of electronic support (computer, phones). Through these parameters, Anis (2003) identifies six types of electronic communications, namely: e-mails, mailing lists, forums, instant messaging, chats and SMS (Short-Message Service). With regards to the type of reader, some platforms, such as Facebook, offer multiple modalities, as mentioned above. The offered modalities include the possibility of having a private communication with a specific person, or sharing content with a group of people. The first modality is usually synchronous, while the second one it is not necessarily synchronous, instead resembling the forums or even the blog writing genre. Marcoccia (2016) observed that the asynchronous texts (e-mails, mailing lists, blog, forums) are usually the most formal ones, making use of a writing style which is much closer to the standard style, for reasons that he traces back to the fact that texts of this type are more permanent because they can be archived. On the other hand, synchronous conversations make use of writing styles closer to the oral communication style, due to same not being persistent. Herring (2007) classifies computer-mediated discourse in a scheme based on features, which she refers to as *facets*, applying the Hymes (1974) S.P.E.A.K.I.N.G. taxonomy.[2] According to Herring, the most important medium factors are:

[2] Setting and Scene, Participants, Ends, Act Sequence, Key (tone, manner, or spirit of the speech act), Instrumentalities (channels), Norms, and Genres.

| | |
|---|---|
| M1 - Synchronicity; | M2 - Message transmission (1-way vs. 2-way); |
| M3 - Persistence of transcript; | M4 - Size of message buffer; |
| M5 - Channels of communication; | M6 - Anonymous messaging; |
| M7 - Private messaging; | M8 - Filtering; |
| M9 - Quoting; | M10 - Message format. |

*Table 4. Herring's medium factors in computer-mediated discourse.*

| | |
|---|---|
| S1 - Participation structure; | S2 - Participant characteristics; |
| S3 - Purpose; | S4 - Topic or Theme; |
| S5 - Tone; | S6 - Activity; |
| S7 - Norms; | S8 - Code. |

*Table 5. Herring's situational factors in computer-mediated discourse.*

Regarding the M2 parameter, Herring (2007) explains that it depends on the granularity of the units that are transmitted through the technological device, character-by-character (2-way) or word-by-word, or even line-by-line (1-way), considering the implication that these different modalities have on simultaneous feedback availability during conversations. With respect to the M10 parameter, the scholar states that the order in which messages appear, or are organised, in the conversation structure, is not always chronological. He also affirms that this can influence the communicative practices on the platform. Herring (2007) also observed social and situational factors, which are shown in Table 5.

The S1 parameter refers to the intended participation modality, which can be one-to-one, one-to-many, many-to-many or a closed group, public or private, anonymous or using nicknames. The participant characteristics parameter describes participants' profiles, which documents them as belonging to specific cultures, having their real-life status or role, their skills in CMC, as well as ethnographic factors, such as age, gender, etc.[3] The S3 parameter refers to the group's interaction aims, while S4 is about the specific topics, if there are any. The tone parameter focus on the formality or standard attitude within the group, while S6 concerns the main activities of

[3] For the role of corpus linguistics in gender in sociolinguistics studies, such as studies on gendered languages, see Mori (2019).

group users. Finally, the norms parameter refers to the group rules, concerning organisation, social behaviour, and languages, while S8 focuses on the latter, namely the language variety, the font and the writing system. This last parameter is particularly relevant to our research, which concerns the construction of a Tunisian Arabizi corpus, where, as we saw in Chapter 1, Arabizi is an encoding system which was created specifically for electronic writing. In general, the rapid and consistent diffusion of electronic writings made available to social sciences and language scholars has created a large repository of data, which is easily accessible, albeit with certain limitations. In general, there is a positive aspect to studying linguistic phenomena through a corpus, due to the zoom lens effect that allows the observer to easily detect the phenomena that he is looking for (Duchet et al., 2008). At the same time, it has been proven that, in order to observe certain specific structures of the language, such as collocations, a quantitative overview is needed (Kraif and Tutin, 2020). However, the positive aspect of using DNW data for linguists' work is the fact that DNW data collection allows the researcher to avoid the *observer's paradox* as described by Labov (1972):

> '[T]*he aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain these data by systematic observation.*'

This is not the only problem to exist in DNW data collection. There are also the ethical considerations connected to what is considered to be public or private, and how to protect sensitive user data when collecting his texts. These topics will be further analysed within the following section, in particular in Section 3.

On the basis of Herring's studies it is therefore possible to state that the DNW does not constitute a single textual genre, but is rather a collection of textual micro-genres brought together by the common characteristic of being forms of writing which are supported and mediated by the use of electronic devices. In fact, quantitative approaches, such as that of Panckhurst (2006),[4] may reveal divergent trends among different DNWs, i.e. chat, forums, and email. As an example, Anis (2004) shows the formality traits of a typical mailing list register. As previously mentioned, a productive discussion about DNW is the assignment of its subgroups to the written vs. oral or hybrid class. A number of scholars have been working on

[4] In this specific case, the different trends relate to the different amount of nouns, verbs, and adverbs employed.

DNW feature classification of the written or oral type, with the most rational conclusion that can be drawn being that DNW represents the hybrid nature of a graphic-speech that contains both written-like and speech-like characteristics (Anis, 1998; Gadet, 2008; Georgakopoulou, 2011). As early as the 1980s, the view of a continuum of spoken and written language manifestations had started gaining ground, a proposal which was advanced by Tannen et al. (1982) from the observation that, for example, speech can follow a script and writing can be colloquial. In fact, since the early days of CMC research, the blending of the written and oral types in employing oral-like strategies for electronic writing has been evident (Gadet, 1996), as reported also by Hert (1999):

> *'La forme du débat électronique induit certes une continuité dans le style avec d'autres formes d'écritures (note de synthèse, note de lecture, texte d'une conférence, etc.). Cependant, ce mode d'écriture fait exister des formes d'interaction qui se rapprochent des propriétés de l'échange oral. A travers les pratiques de reprises, de citations, de synthèses, à travers l'utilisation du contexte des échanges, à travers également le déroulement temporel du débat en parallèle avec d'autres événements, se construit cette quasi-oralité de l'écriture* [...] *Elle consiste dans ce travail sur l'écriture visant à y introduire des propriétés de l'échange oral, comme la participation collective à ce qui s'énonce, la définition en situation du contexte de l'échange, un sentiment de communauté* [...].'[5]

There is no doubt that electronic *writing* is primarily a form of *writing*. Panckhurst (2006) reports a number of features of written discourse, such as interrogative forms, circumfixed French negation etc., which are also present in DNW texts, particularly with respect to such phenomena in the typology of asynchronous DNW. On the other hand, as Hert points out, electronic writing lies along a continuum between written and oral, where it can draw on the char-

[5] 'The form of the electronic debate certainly induces a continuity in style with other forms of writing (summary notes, reading notes, conference texts, etc.). However, this mode of writing brings into existence forms of interaction that are close to the properties of oral exchanges. Through the practices of repetition, quotations, and synthesis, through the use of the context of the exchanges, through the temporal unfolding of the debate in parallel with other events, this quasi-orality of writing is constructed [...] It consists, in this work, on writing with the aim to introduce properties of oral exchange, such as collective participation in what is being said, the definition in situation of the context of the exchange, a feeling of community [...].' My own translation.

acteristics of both forms. For example, oral discourse is generally improvised and spontaneous, which lumps it in with chatroom conversations. It has already been stated that some types of DNW are more formal than others, namely the asynchronous ones, such as e-mails, mailing lists, blogs, and forums, in comparison with the synchronous ones, such as text messaging and instantaneous chatting services. Imagining the extremes of the above continuum as constituted by the written and oral poles, it is therefore possible to assume that it is the register (formal/informal) that motivates the use of traits which are more inclined to the written-organised-formal-pole or the oral-spontaneous-informal pole (Panckhurst, 2006). Marcoccia (2016) in fact supports the same thesis and also includes the binomial monological/dialogical in the subdivision between the two poles. Section 2 has already explored some typical characteristics of DNW, which Thurlow and Poff (2013) defines as *typographical-cum-linguistic devices*, such as onomatopoeia or exaggerated use of spelling and punctuation (with more of an expressive than a syntactic function (Anis, 1994)), pragmatic use of capitalisation, spacing, and special symbols for adding the missing prosody and emphasis impact which are obviously missing in a DNW text. According to Gadet (2008), these devices can be traced back to the logic of fictitious and performed orality, that we often encounter in the DNW of chats (and which is almost absent in blogging, for example), as brands of familiarisation with other participants through a kind of mimicry of the prosody and emphasis of pretended-oral-discourse. According to Crystal (2004), these are mainly devices which are used in order to fill the pragmatic gap due to being physically non-present during the conversation, which leads the user to respect the four conversational maxims of the philosopher H. P. Grice for an efficient co-operative use of language, namely: *The Maxims of Quality*, *The Maxim of Relevance*, *The Maxim of Quantity* and *The Maxim of Manner*. Marcoccia (2016) distinguishes five levels of DNW characteristics:

## 3. Corpus Linguistic Standards for DNW Corpora

> '*A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research'* (Sinclair, 2004, 19).

This is the definition of a corpus given by Sinclair (2004). Before describing the building of the Tunisian Arabish Corpus (TArC), it is necessary to give a general overview of Corpus Linguistic (CL)

| | |
|---|---|
| 1 - **Typographical:** | *Integration of non-alphabetic symbols in text (such as @), non-standard capitalization, emoticons.* |
| 2 - **Orthographic:** | *Use of phonetic spelling (phonetic-based substitutions, i.e. 'koi' for French 'quoi').* |
| 3 - **Morphological:** | *Use of abbreviations, word truncation, use of acronyms, etc.* |
| 4 - **Lexical:** | *Use of specific vocabulary associated with youth and CMC vocabulary.* |
| 5 - **Syntactical:** | *Use of telegraphic or fragmented syntax, with the loss of some sentence components, such as articles. The frequent use of 1st and 2nd person pronouns vs. the 3rd, as well as the extensive use of modal verbs (Yates, 1996).* |

*Table 6. DNW characteristics.*

standards, starting from the external criteria mentioned in the above definition of a corpus.

### *General Principles and Criteria in Data Collection*

The guiding principles for corpora building are criteria that are not strictly definable, but are based on the common sense and good thinking of those involved in the corpus building. Feedback from users is also of great value in order to improve the structure of the corpus. The concepts generally associated with a good corpus structure are those of *representativeness* and *balance*. According to Sinclair (2004), in order to achieve the most representative corpus possible, structural criteria, which are few in number and easy to respect, should be decided *a priori*. Following the structural criteria, the corpus builder (CB) should identify the text genres that meet them in order. With regards to the aim of balance, the CB should avoid texts which are too specialized, unless those coincide with the specific target texts of the corpus. Other common structural criteria are documented as follows:

According to Sinclair (2004), the following table outlines a general strategy for corpus building:

With regards to the first and second points, these concern any kind of corpus building, while the third point mainly concerns inbuilt contrastive corpora, such as the parallel ones (involving more than one language), which is not the case of the Tunisian Arabish Corpus (TArC). This can also be considered in the case of a diachronically

| | |
|---|---|
| **1. Text's mode:** | speech, writing, DNW, etc.; |
| **2. Text's genre:** | romance, journal, blog, etc.; |
| **3. Text's domain:** | academic, street language, etc.; |
| **4. Text's language:** | language or variety; |
| **5. Text's location:** | the place of origin; |
| **6. Text's date:** | date or period. |

*Table 7. Corpus Structural Criteria.*

| | |
|---|---|
| **1.** | *Collect texts according to their communicative function and not their linguistic contents.* |
| **2.** | *Commit toward the highest possible degree of representativeness.* |
| **3** | *Avoid basing contrastive analyses on contrastive structural components of the corpus.* |
| **4.** | *Choose criteria a priori; these should be few, well separated from each other, and efficient as a set of reference points.* |
| **5.** | *Separately store information about the texts.* |
| **6.** | *Collect samples of the language which are as complete as possible.* |
| **7.** | *Document corpus design steps.* |
| **8.** | *Be guided by representativeness and balance, even if these are not precisely definable concepts.* |
| **9.** | *Exert control on the topic of texts driven only by external criteria, not internal ones.* |
| **10.** | *Aim for homogeneity, avoiding unusual texts.* |

*Table 8. Corpus Building Strategy.*

contrastive corpus. According to the fourth point, criteria for TArC building have been chosen *a priori* and are outlined in Table 9. With regards to the fifth point, this concerns information on the structure of texts and explains that texts should be separated from plain text. The information can be about the provenience of the text or about the language contained in it, in that case, usually: annotations. The annotation structure of TArC will be described in Chapter 3. Concerning the ninth point, no control was exerted over the topic of the text. However, in order to detect DNW Tunisian websites and in order to build a corpus which is as representative as possible of the

linguistic system, it would be useful to identify wide thematic categories that could represent the most common topics of daily conversations on DNW. In this regard, two instruments with a similar thematic organisation have been employed:

1. 'A Frequency Dictionary of Arabic' and in particular its 'Thematic Vocabulary List' (TVL) (Buckwalter and Parkinson, 2014);
2. The 'Loanword Typology Meaning List', which is a list of 1460 meanings (LTML) (Haspelmath and Tadmor, 2009).

The TVL consists of 30 groups of frequent words, each one represented by a thematic word. The 'Loanword Typology Meaning List' is the result of a joint project by Uri Tadmor and Martin Haspelmath: the 'Loanword Typology Project' (LWT), which was launched in 2004 and ended in 2008. The LTML consists of 23 groups of basic meanings sorted by a representative heading word. Considering that the boundaries between some categories are very blurred, some categories have been merged, such as 'Body' and 'Health'. Some others were not taken into consideration due to their irrelevance to the purposes of our research, e.g. 'Colors', 'Opposites', 'Male names'. In the end, the focus of this work will be 15 macro-categories. With the aim of easily detecting texts and their respective URLs, without introducing relevant query biases, it was decided that the category names should not be used as query keywords (Schäfer and Bildhauer, 2013). Therefore, we associated a set of Tunisian Arabizi keywords belonging to basic Tunisian vocabulary to each category. It was found that three meanings belonging to the semantic category were enough to obtain a sufficient number of keywords and URLs for each category. For example, for the category 'Family', the meanings 'son', 'wedding', 'divorce' were associated with all their Arabizi variants, obtaining a set of 11 keywords (Gugliotta and Dinarelli, 2020a,b).

As shown in Table 9, which details the criteria adopted to build TArC, in addition to the first three textual genres, which are related to DNW context, the decision was made to also collect text of rap lyrics, which were spontaneously written on dedicated forums by their users. This choice was motivated by the hypothesis that these texts could represent a useful tool for orthographic comparison (between the two systems of Arabizi and Arabic-characters script), since these texts use a mixture of both systems. The slightly different nature from the target of the other textual genres in TArC is not considered to invalidate the consistency of the contents of TArC, consider-

| | |
|---|---|
| **1. Text's mode:** | DNW informal written texts. |
| **2. Text's genre:** | blogs, forums, social network posts and related comments. Rap lyrics. |
| **3. Text's domain:** | Computer Mediated Communication (in particular DNW). |
| **4. Text's language:** | Tunisian (encoded in Arabizi). |
| **5. Text's location:** | users' origins. |
| **6. Text's date:** | publication date. |

*Table 9. TArC Structural Criteria.*

ing that TArC is structured in such a way as to be flexible in isolating each textual genre, so that, in the case where users of TArC want to carry out analyses on the texts of TArC, except for those of rap, they can easily extract the latter, working only on the other three genres. With regards to the location criterion, this coincides with metadata, which is collected only if made available by the users themselves. We decided not to record the exact hometown of users, instead recording the governorate to which it belongs, in order to comply with the users' right to privacy, a topic that will be discussed in Section 3. The 24 governorates of Tunisia (*wilāyāt*) represent the first level of territorial subdivisions of the country, and have been reported in TArC under their French names.[6]

### *Context and Metadata*

The DNW data collection presents a methodological problem which requires attention. This concerns the link between observed phenomena and the contexts, in other words, the possibility of being phenomena context-dependent variables. Sinclair (2004) explains that ideally a corpus should be designed by a researcher who is knowledgeable about the language context and the speech community whose language is to be represented by the corpus. He further explains that semantic content should be of secondary importance during text selection, and that texts should be selected regard-

[6] Ariana, Béja, Sousse, Bizerte, Gabès, Nabeul, Jendouba, Kairouan, Zaghouan, Kebili, Le Kef, Mahdia, Manouba, Medenine, Monastir, Gafsa, Sfax, Sidi Bouzid, Siliana, Ben Arous, Tataouine, Tozeur, Tunis and Kasserine.

less of content. It will be difficult for the observer, due to the lack of the context, to understand whether the observed phenomenon is context-dependent or not. According to Marcoccia (2016), there are two strategies for addressing this problem, a text-centered solution and context reconstruction. The first of these takes into account the fact that the context cannot be reconstructed, such that the observer acts as if it doesn't exist, dealing with the texts as decontextualised texts by default. The second of these instead tries to reconstruct the contexts considering that the same texts provide all the necessary information to contextualise it. These two positions are quite radical, considering that in some occasions it is really easy to extract the context from a conversation, and in this case, dealing with the text as de-contextualised by default may mean that the observer loses part of the information. On some other occasions, it is not really possible to re-construct the context, and for that reason, a strong effort in this direction could cause the observer to misrepresent it. With regard to the understanding of whether a phenomenon is context-dependent or not, a good number of diversified texts can ensure that the observer has the right point of view.

Concerning the corpus building stage, a good practice to use when recording contextual information is the recording of metadata, namely data about data (Burnard, 2004). Linguistic corpora should be designed to support different types of analyses and this is the reason for providing a metadata framework which is as complete as possible. CL acknowledges different strategies for providing texts with information, such as the Extensible Markup Language (XML) and in particular, the Text Encoding Initiative (TEI), the Open Language Archive Community (OLAC), and the ISLE Metadata Initiative (IMDI) (Burnard, 2004). Depending on the aims behind corpus collection, it is not strictly necessary to encode information through this type of formalism, which is commonly criticised for being too wide and complex (Lancioni, 2011). In fact, computer systems are now able to extract information from texts organised in *comma* or *tab separated values* or *tabular* format (CSV, TSV or TAB). One strong point of these formats is that corpora encoded in CSV or TAB can easily be converted into other formats, such as the Excel format (XLSX), which makes these formats readable through other

| **Text metadata** | **User metadata**. |
|---|---|
| Genre (Forum, Blog, Social Network, Rap). | Gender: Male (M) and Female (F). |
| Publication date. | Age range: [-25], [25-35], [35-50], [50+]. |
| | Governorate of belonging. |

*Table 10. TArC Metadata.*

software or tools, such as *Sketch Engine*.[7] At the same time, in the case of corpora designed with a specific use in terms of linguistic analysis, the type of metadata that must be recorded can be very complex (e.g. annotations of turns of phrase and general discourse sub-units). Not only is it difficult to automatically extract this type of metadata, requiring manual work, but it also requires a format that allows for deep levels of annotation, such as XML or TEI formats, unlike other formats that are more operational for NLP (CSV or TAB), but which are structurally poorer. In general, it is better to avoid building a corpus in commercial word-processing programs, such as Microsoft Word, because it cannot be processed by a corpus analysis tool (Wynne, 2004). In addition, if the corpus is stored in CSV or TAB, it will be available regardless of whether a specific software is valid or if it has been updated. TArC is stored on GitHub, a Git repository hosting service.[8] While Git is a command line tool, GitHub provides a Web-based graphical interface. It also provides access control and several collaboration features. TArC files are made available, sorted by genre in both TAB and XLSX, and are continuously updated. However, every update, even small ones, is recorded by the system, so it is possible to go back to previous versions. Concerning the metadata collected in TArC, as shown in Table 10, this is metadata about the texts and metadata about the users.

### *Data annotation*

> *'Corpus-builders do not in general have the leisure to read and manually tag the majority of their materials; detailed distinctions must therefore be made either*

[7] The following is the link to this linguists and lexicographers tool for language analysis: `https://www.sketchengine.eu/`.Consulted on 7th April 2021.

[8] Git is a distributed version control software usable from a command line interface, created by Linus Torvalds in 2005. TArC is available at the following link `https://github.com/eligugliotta/tarc`.

> *automatically or not at all* [...] *In the simplest case, a corpus builder may be able reliably to encode only the visually salient features of a written text such as its use of italic font or emphasis, or by applying probabilistic rules derived from other surface features such as capitalisation or white space usage*.' (Burnard, 2004).

Annotating a corpus refers to adding interpretative, linguistic information to the corpus texts, such as marking a word in a text as corresponding to a particular Part-of-Speech (POS tagging), based on both its definition and its context, or reducing an inflected form of a word to its lemma (lemmatisation) (Leech et al., 1997). Depending on the purpose of the corpus, annotations can be of different types, and in some cases, CBs prefer not to annotate the corpus at all in order to investigate it in its *pure* form, or to avoid same being affected during analysis by any annotation errors or problems due to not highlighting phenomena compared to others which are marked by a tag. Regarding the latter, this is the case described by Sinclair (2004) concerning the fact that if the data in the corpus can only be observed through the tags, anything that the tags are not sensitive to will not be noticed. In other cases, annotation is seen as enriching the text and supporting its use (Leech, 2004; Leech et al., 1997). With respect to TArC, levels of annotation of texts and words are provided, due to the fact that in an under-researched language such as Tunisian, which is encoded in a non-standardised orthographic system, text annotation could be considered to be of added value to the corpus. There are two main reasons for this: to support linguistic analyses on TArC and to support the development of tools for automatic processing or as a benchmark for developing web-based interfaces for Tunisian learners (Granger et al., 2007; Gugliotta et al., 2020). As will be discussed in Section 4, with regard to automatic processing, the tools that Natural Language Processing have made available today allow us to expand the possibilities of an annotator, who, as we will see, contrary to what Burnard says in the quotation above, does not have to manually annotate an enormous amount of data, but must instead post-edit the product of an automatic processing of data, thus reducing the amount of errors in the corpus. On the other hand, although the support of NLP tools decreases the rate of annotation errors if combined with manual quality control, it is also true that the result, even in the case of completely manual annotation, can never be 100% correct (Leech, 2004). In fact, Artificial Intelligence systems mirror human behaviour and, in the specific case of NLP systems, reflect human *language*, which is a strong

signal of individual behaviour depending on the *context*. However, context is latent information in corpora, and a lack of attention to context while processing texts automatically could lead to issues in representativeness issues, such as the overgeneralisation of human behaviour (Hovy and Spruit, 2016). Leech (2004) states that the information recorded through the corpus annotation levels can be also extracted with wider aims, such as building dictionaries or, in the case of POS tagging annotation, building an automatic syntactic parser. Indeed, annotation should aim to be multi-functional, especially considering that

> '*future uses are always more variable than the originator of the corpus could have imagined! The same is true of an annotated corpus: the annotations themselves spark off a whole new range of uses which would not have been practicable unless the corpus had been annotated.*' (Leech, 2004, 23).

For the same reason, annotation should always be easily separable from the text, transparent to human readers, and explicitly documented together with the pre-annotation practices, such as text segmentation. As stated by Leech (2004), annotation should comply with two quality criteria: the category realism of the annotation, and its accuracy and consistency. Category realism refers to the tag's capability to represent linguistic elements well. Achieving a high quality of representation in terms of tagsets is a very difficult, if not impossible, task, especially when it comes to tagsets designed to describe multiple languages. Accuracy concerns the percentage of tokens which have been correctly annotated. The information integrated in TArC is therefore as follows:

1. The metadata
   (a) The publication date
   (b) The users' metadata:
      i. Governorate of origin
      ii. Age range
      iii. Sexual gender
2. The structural partition in subcorpora:
   (a) Social network corpus
   (b) Blog corpus
   (c) Forum corpus
   (d) Rap lyrics corpus
3. The annotation levels at the word level in
   (a) Classification in
      i. Arabizi
      ii. Foreign

iii.Emotag
(b) Transliteration in Arabic characters (according to the CODA* convention)
(c) Tokenisation
(c) POS-tagging
(d) Lemmatisation

TArC annotations will be outlined in the next chapter, together with TArC's annotation semi-automatic processing.

### *Access & Ethics*

One of the aims of corpus building is sharing the corpus with the scientific community, so that it can serve as a useful tool for various types of research. As already mentioned, it is important to ensure the ongoing availability of the resource, i.e. providing it though a reliable repository. In the case of TArC, as previously noted (Section 3), the GitHub platform was chosen in this case. Freely accessible corpora in Open Access are not readily available for two reasons:[9]

1. A corpus is a CBs intellectual property, considering their intellectual work on texts.
2. A corpus is a collection of texts produced by other people, who are the intellectual owners of this texts.

As mentioned previously, the corpus collection of DNW texts leads to ethical issues, the solution to which is not always evident, a reminder of which is provided by the *National Committee for Research Ethics in the Social Sciences and the Humanities* (NESH) of the Norwegian National Research Ethics Committees.[10]

> '*Technological development is advancing rapidly, and this raises new challenges for research ethics.* [...] *Internet based social networks such as Facebook*

[9] Collecting someone's text and putting them into a corpus can be a breach of copyright, even if the corpus is not distributed. However, in general, CBs prefer at least to be the first to be able to perform analyses on corpora, before they make them available to others. This also allows them to test them and to provide an evaluation of their results, which can then be reduplicated, once the corpus is made available to others (Wynne, 2004).

[10] A guidelines for ethic research based on Internet data is available at `https://www.forskningsetikk.no/en/guidelines/social-sciences-humanities-law-and-theology/a-guide-to-internet-research-ethics/`. Consulted on 7th April 2021.

> *became widespread around 2005, and in all the sharing of information it is sometimes unclear what is public and what is private. Smartphones and mobile Internet connections (3G, 4G, Wifi) appeared around 2008, with various apps that register data on health and location, raising issues pertaining to data storage and surveillance. Furthermore, since 2010, development in the areas of digitalisation and automatisation has led to production, dissemination and storage of huge amounts of data at an increasing pace, often in real time. As a result of this rapid development, the criteria for what constitutes good and justifiable research are not always obvious.* [...] *Researchers are personally responsible for ensuring that the protection of the interests and rights of individuals, based on the respect for human dignity and the requirement to protect privacy, are protected.*'

Europe has implemented the General Data Protection Regulation (GDPR), a European Union regulation on personal data processing and privacy, which was adopted on 27 April 2016, and operational as of 25 May 2018. With this regulation, the European Commission aimed to strengthen the protection of personal data of European Union (EU) citizens and EU residents, both within and outside EU borders, by giving citizens back control of their personal data, simplifying the regulatory environment affecting international affairs, and unifying and homogenising privacy laws within the EU. However, there are no regulations dedicated to the practice of data collection for research; the responsibility for this lies in the hands of the researcher. In order to exhibit ethical behaviour, the first question the researcher must ask oneself is whether the data he or she intends to collect is public or private. It turns out that the distinction between the two is not as sharp as we might expect at first glance. In addition to the fact that some sites (usually forums) require a membership that regulates access, limiting the visibility of conversations, the expectations of the authors for their texts must also be taken into account, that is, if the user has written a text, whether they imagine their audience to be limited to users of that environment or to be open to the public without any limits, as NESH further reports:

> '*As a main rule, researchers ought to proceed with greater caution the stronger the restrictions on access. On the other hand, the purpose of technical restrictions on access and 'private' groups may be to protect statements that in principle are public. For example, Facebook-groups with thousands of members could be regarded as public, despite any technical settings indicating that the group is 'private' or only for 'friends'. The larger the group, the more public the information.*'

Marcoccia (2016) states that forums are usually considered public, even if the publication of a new post requires an inscription, and

that easy access doesn't automatically translate to public. The researcher should also consider the content, and if it reveals personal and sensitive information in open online forums, the researcher should exercise due care and take personal responsibility to safeguard the integrity and interests of the individual, including respect for their privacy and family life. The best solution for ethical data collection is to seek written permission from users, which, in the case of large numbers of users, is not always feasible. In order to protect the users' identity, the researcher must then adopt strategies such as the anonymisation and obscuration of sensitive data.

As far as TArC is concerned, the following measures were taken: the anonymisation of all texts and the obscuration of sensitive data. Consent from the authors of the blog was also requested, with the owners signing a document, granting us the possibility to publish the corpus, with one of the two non-commercial Creative Commons licenses which we intended to use. The text of the agreement form states that:

> 'By signing this form, the author of the blog: *Blogger name*, agrees to the use that *Corpus Builders names* put forth, for academic research purposes. The uses will consist of:
>
> 1. Blog text collection into a corpus;
> 2. Text analysis;
> 3. Mentions of the corpus in academic publications (such as papers and theses);
> 4. Publishing the corpus containing the blog with a non-commercial Creative Commons license (CC BY-NC or CC BY-NC-SA) in order to support researchers from the scientific community with their research on Tunisian dialects.'

In the end, we opted for license *Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)*[11].

## 4. Deep Learning Techniques

### *General Background*

Deep Learning (DL) is a sub-set of Machine Learning (ML), which is a sub-set of Artificial Intelligence (AI). As noticed by Oberlander (2005),

[11] The link to the licence is as follows: `https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode`. Consulted on 7th April 2021.

> '[m]*ost researchers in AI aim to develop computer programs that help a machine exhibit behaviour that, if it were the behaviour of a human, would be called intelligent.*'

Moreover, Chowdhary (2020) stated that *intelligence* could be defined as the result of *perception* + *analysis* + *reaction*. The first of these is committed to the concept of learning through real experience, while the second is committed to the capacity for identifying its structures, and the third to the ability to solve new problems.

DL, being a type of ML, deals with making machines learn specific tasks through models trained on large amounts of data. What distinguishes DL from ML models, in general based on probability calculations, is mainly the neural modeling which is typical of DL. Artificial Neural Networks (ANNs) are inspired by the structure of the nervous tissue of biological brains (in fact, Deep Learning is also part of the bio-inspired intelligence field). ANNs today are used for many different and disparate tasks, such as voice recognition, customer support, and image recognition, for example for medical care (MR image analysis). Regarding tasks which are more familiar in NLP, we can mention machine translation, language modelling, coreference resolution, sequence tagging, syntactic analysis, etc..

Artificial intelligence can be traced back to the 1940s, when mathematicians formalised the neuron as a small processor. In particular, McCulloch & Pitts (1943) observed that networks of binary neurons could perform logical operations, by activating or not activating connections between them, like Boolean circuits. The conclusion of their observations was that the brain is a logical inference machine, because neurons are binary, so a neuron computes a weighted sum of its inputs and then compares the weighted sum to a threshold (Russel et al., 2009).

The idea that the brain learns by modifying the strength of connections between neurons, which are called synapses, is formalised in the Hebbian theory of synaptic plasticity, which owes its name to Donald Hebb, a Canadian psychologist. Donald Hebb demonstrated, in 1949, a simple update rule for changing the connecting forces between neurons. His rule, now referred to as *Hebbian learning*, still remains an influential model these days. Russel et al. (2009) state that there were several early examples of work that could have been characterised as AI, but Alan Turing's vision was perhaps the most influential. Of particular note is his 1950 article 'Computing Machinery and Intelligence', in which he introduced the Turing test, on machine learning, genetic algorithms, and reinforcement learn-

ing. One of the most important figures in the study of biological and mechanical control systems and their connection to cognition was the mathematician Norbert Wiener. In the late 1940s, Wiener, along with Warren McCulloch, Walter Pitts, and John von Neumann, organised a series of influential conferences exploring new mathematical and computational models of cognition. Wiener's book 'Cybernetics' (1948) became a bestseller and awakened the public to the possibility of artificially intelligent machines. Finally, in 1958 Frank Rosenblatt proposed the 'perceptron' as an entity with an input layer and an output layer and a learning rule, based on the minimisation of an error, i.e. the so-called *error back-propagation function*. This function, which is still used today, alters the weights of the connections (synapses), taking into account the difference between the actual output and the desired one.[12] The enthusiasm was great, but after Minsky and Papert (1969) demonstrated the limits of the perceptron, i.e. its ability to recognise, after appropriate training, only linearly separable data, interest waned rapidly. In fact, a multi-level network of perceptrons could solve more complex problems, but the increasing computational complexity of training made this road impractical. It was only in the following decade that the usefulness of this operational entity began to be considered again, especially in terms of multi-layer neural networks. However, if artificial intelligence began around the 1940s, it is only now that it is undergoing exponential acceleration thanks to the power of current technological tools compared to those of the past. Being more powerful, they can analyse more data at the same time, and the resulting programs can also be more effective. The explosion of data and computing power in the 2000s brought these techniques back to the forefront, due to the fact that it is now possible to build much deeper networks with many layers, hence *deep* networks. An important milestone in NLP was the work of Bengio et al. (2001, 2003). In fact, they definitely made real improvements by proposing a neural probabilistic language model which can simultaneously learn a distributed word representation (word feature vector) and the probability function for word sequences (in terms of word feature vectors of the sequence words).[13] A few years later, Schwenk (2007) extended the previous

[12] The algorithm which is used in supervised learning to implement *error back-propagation* consists of an application of the *gradient descent* method, where the contribution of each parameter to the model error is given by the partial derivative of the loss function with respect to that parameter.

[13] This topic will be explored again in next section.

work to language modeling for large vocabulary continuous speech recognition. Indeed, NLP research during this period experienced a huge surge. Android has been using neural net speech recognition systems since 2012. However, 2013 was the year in which computer vision started to uses neural nets and gradually abandoned all other techniques. In addition, Google's and Facebook's current NLP applications are completely built around Deep Learning.

Neural Networks structure can be described, in terms of structure, as a model consisting of at least two layers, an input layer and an output layer, and there are usually also additional intermediate layers (hidden layers). Each layer of the network contains a number of specialised artificial neurons. Information in the form of patterns or signals is transferred to neurons in the input layer, where it is then processed. Each neuron is assigned a weight, so that each one receives a different relevance. The weight, along with a transfer function, determines where the information is then forwarded. In the next step, an activation function using a threshold value calculates and weights the output value of the neuron. Depending on the information evaluation and weighting, other neurons are connected and activated to a greater or lesser extent. By means of these processes, a model is set up to produce an output for each input. With each training process, the weighting, and thus the algorithm's parameters, is modified so that the network yields more accurate and better results. Different neural network structures are used depending on the learning method used and the purpose of the application, such as Feed-Forward Networks (FFNs), Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs). Learning can occur either in a supervised or unsupervised form. The difference between the two is that supervised learning allows the model to make inferences from data that has already been labeled and is used for classification and regression tasks. Unsupervised learning occurs when the model learns from an unlabeled training set, and thus requires a self-organisation of the net parameters, based on training data set characteristics.[14]

[14] Reinforcement learning also exists, which aims at learning to select an action in order to maximise the output.

### *Overview of Natural Language Processing Approaches*

> '*Natural Language processing* [...] *is the sub-field of artificial intelligence (AI) focused on modelling natural languages to build applications such as speech recognition and synthesis, machine translation, optical character recognition (OCR), sentiment analysis (SA)* [...] *etc. NLP is a highly interdisciplinary field with connections to computer science, linguistics, cognitive science, psychology, mathematics and others*' (Darwish et al., 2021).

The neural network revolution and the birth of Deep Learning was hinted in the previous section. This revolution has also disrupted the field of Natural Language Processing (NLP). In fact, nowadays, NLP is embracing AI techniques, and can thus be considered to be an AI sub-field, as stated by Darwish et al. (2021) in the above quote. However, before the DL upheaval, NLP was also addressed using other available approaches, such as the symbolic one, based on grammaratical complex hand-written rules with the aim of parsing the language. For further details on this approach, refer to the work of Charniak et al. (1983); Grosz et al. (1987); Hirst (1984, 1987); Riesbeck (1986). An important development of symbolic approaches in NLP has been the employment of statistical models. In fact, statistical models have made it possible to overcome the limitations of the complexity of hand-written rules by allowing them to be generated through machine learning. For more information about the statistical approach to Natural Language Processing, please refer to Bahl et al. (1989); Brill et al. (1990); Brown et al. (1990, 1991); Chitrao and Grishman (1990); Lesk (1986); Liberman (1991).

The current popular method for representing words in Deep Learning, which has produced impressive results, consists of using *word embedding*, namely word representation in a continuous multidimensional space. In fact, most of the DL approaches require an input encoded as a vector of features of fixed dimensions. The idea of using neural networks that represent words in a continuous space to reduce the number of parameters to be estimated comes from Bengio et al. (2003). Exploiting the fact that words have multiple degrees of similarities, this distributed semantic representation is used to map words in a vector's continuous spaces where words take place according to their features (Mikolov et al., 2013). In order to use word embedding with Recurrent Neural Networks (RNNs), it is usually necessary to create an embedding layer in the RNN as in Mikolov (2007) or Mikolov et al. (2009). Word embedding comprises a dense representation of words as multidimensional vectors, and the vector dimensions vary according to the amount of vocabulary.

One way to create word embedding is to start with dense vectors for each token containing random numbers, and then use these to train a model such as a document classifier or sentiment classifier. The vectors, which represent the tokens, will be distributed in such a way that semantically closer words will have similar representations. However, in order to use word embedding, it is also possible to employ the most common models such as Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014). This approach has been widely adopted, especially in light of improvements with respect to vector quality and model training speed, and advances in hardware have also been made. In general, DL-based approaches allow for the production of results that highly depend on the textual input data, which can be very large. Unlike with the previously mentioned approaches, the more observations the DL algorithm encounters, the more it will improve and gain in precision. Indeed, in the last decade, neural networks have established themselves as the most state-of-the-art models in most NLP tasks. This can be also explained by the fact that textual content is becoming more available on social media and sizable corpora can be collected and used more easily. Data is, in fact, the lymph of NLP, and this is true for all rule-based approaches that require lexicons and carefully created rules, as well as for DL supervised approaches that require corpora, especially annotated corpora. The next section (namely Section 4) briefly mentions sequence-to-sequence models (Sutskever et al., 2014; Vaswani et al., 2017), which are built on top of Recurrent Neural Networks (RNN) (Cho et al., 2014a; Hochreiter and Schmidhuber, 1997b), and attention mechanism (Bahdanau et al., 2014; Vaswani et al., 2017) structured in encoder-decoder architectures, these are currently among the most effective solutions for NLP problems.

### *Sequence Modelling*

Sequential data prediction is considered by many as a key problem in machine learning and artificial intelligence (Mikolov et al., 2010). There are several Deep Learning (DL) models used for data processing, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and in particular LSTM, etc. These models are different from each other, and each of them is known to be particularly proficient at performing certain specific tasks, although, basically, they all share a similar architecture: an input layer, an output layer, and one, two (*shallow NN*) or more hidden layers (*deep NN*) in between same. Information propagates through layers,

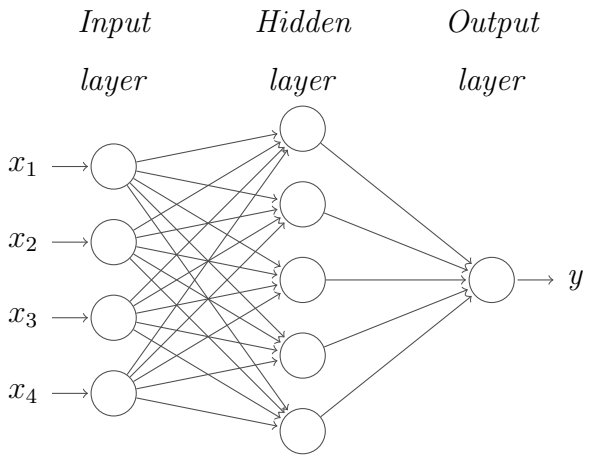



*FIGURE 1 FEEDFORWARD NETWORK.*

thanks to the connections between them. The method of connection varies for different networks. As already mentioned at the beginning of this section (4), the Feed Forward Network (FFN) 'perceptron' was the first and simplest type of artificial neural network devised. A basic idea of the FFN structure is given in Figure 1, where $x$ represent the input values and $y$ is the output. The arrows are the weight matrices that connect the layers of the represented shallow net. The name 'Feed Forward' comes from the direction of the information through this kind of net, which is always forward.

Otherwise, in the Recurent Networks (RNNs), a recurrent matrix connects the hidden layer to itself, using time-delayed connections, as shown in Figure 2. In the RNN unfolding graph in Figure 2, each node is associated with one specific time step. The RNN chunk represented in the figure as ($R$) receives one input $x$, and outputs only a single hidden state $h_t$.[15]

With respect to sequence modelling, some of today's applications include speech recognition, machine translation, topic extraction (i.e. keyphrase lists extraction), speech generation, or text summarising of customer feedback. Indeed, sequences are a data structure in which each example can be viewed as a series of data points. In order to model sequences, a specific framework that can

[15] The image is inspired by the unrolling explanation of Goodfellow et al. (2016) and the Christopher Olah's blog: `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`.

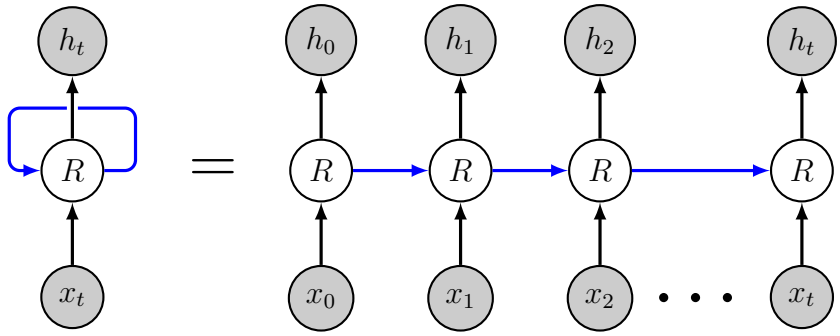



*FIGURE 2 RECURRENT NETWORK UNFOLDING GRAPH.*

handle sequences of varying lengths is required, while also maintaining their order, continuing to track long-term dependencies, and sharing parameters across the sequence. Therefore, language models based on RNNs have been proposed to overcome some limitations of FFN networks, such as their inability to detect dependencies between input patterns (Bengio et al., 2007; Mikolov et al., 2010). In fact, RNNs can efficiently represent complex patterns thanks to their (already mentioned) time-delayed connections (unlike FFN that operate at a single time). The time-delayed connection, which is represented by the blue arrows in 2, allow the recurrent model to form a kind of short-term memory, since past information can be represented by the state of the hidden layer that is updated based on the current input and the state of the hidden layer in the previous time step (Mikolov et al., 2013). Finally, at each time step ($t$), a RNN produces an output based on the current input ($x_t$) and the previous state ($h_{t-1}$). An output, to which an error corresponds, is produced at each time step. For this reason, the error back-propagation in RNNs must involve all the time steps of the net. The back-propagation method, in order to automate the updating of weights in the network hidden state, is based on the gradient descent technique, and is used to correct each parameter of the model by comparing its contribution to the error on the given output (already known).

However, as pointed out by Sutskever et al. (2014), even if the RNN was provided with all the relevant information, it would still be considered difficult to train it, due to the long-term dependencies involved therein which may cause *gradient explosion* or *vanishing* problems. Therefore, in order to solve this type of problem, it has been proposed to use a particular recurrent network, i.e. the Long Short-

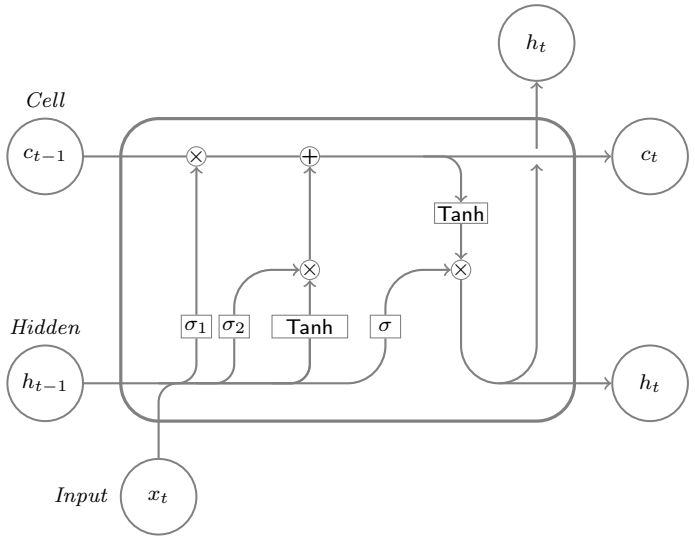



*FIGURE 3 LONG SHORT-TERM MEMORY (LSTM) CELL.*

Term Memory (LSTM), which was introduced by Hochreiter and Schmidhuber (1997a). As discussed above, a RNN has a hidden unit that computes a function of an input and its own previous output, and this is also known as the cell state or hidden state. The difference between a traditional RNN and LSTM resides in the LSTM gating system. The latter allows for control of the flow of information when entering or exiting from the cell module. The gate control has been described as enabling better storage of 'memory'. Furthermore, the cell module can maintain its hidden state through time and process data sequentially by adding or removing information thanks to four additional layers inside it. These additional layers for a traditional LSTM include the 'forget gate layer' ($\sigma_1$ in Figure 3), which is a layer tasked with keeping or removing information. The second one is a layer known as the 'input gate layer' ($\sigma_2$ in Figure 3) for selecting the values to be updated and transferring the information to the following *Tanh layer*, which produces a set of new possible values for the update.[16]

[16] A hyperbolic tangent (Tanh) activation layer applies the Tanh function on the layer inputs, which is also a sigmoid function, that takes any real value as an input and outputs values in the range of -1 to 1.

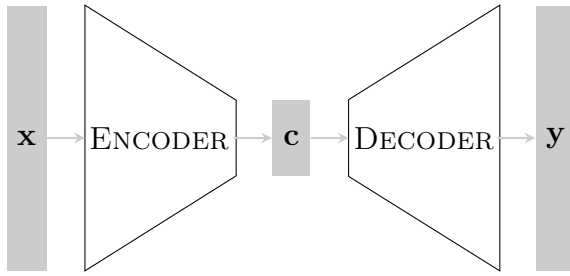



*FIGURE 4 ENCODER-DECODER MODEL.*

The LSTM cell is composed of another gate layer, i.e. the 'output' layer. This layer has the function of filtering the information that will be comprised in the new hidden state to be output. First of all, the current cell state has to be squashed through a new Tahn function. Then, the filter vector is created by the passage of the previous hidden state and the current input data through a sigmoid function. Finally, the squashed cell state is multiplied by the filter vector, giving the final new hidden state to be output. A hidden unit similar to the LSTM hidden cell, but at its simplest, has been proposed by Cho et al. (2014b), and is known as the Gate Recurrent Unit (GRU). In comparison to LSTM, it is considered to be much simpler, because it has less parameters than the LSTM cell, even if the idea is relatively similar. In order to handle previous information together with the next information, Schuster and Paliwal (1997) proposed bidirectional RNNs sequence models.

Considering the improvements given with RNN-based architectures in mapping sequences (Dinarelli and Dupont, 2017; Dinarelli and Grobol, 2019c; Dupont et al., 2017), the first sequence-to-sequence architecture, proposed by Cho et al. (2014b), was a strategy based on input-encoding into a fixed-length vector using RNNs, and output-decoding into the target sequence with another RNN. This architecture, referred to by Cho et al. (2014b) as the *RNN's encoder-decoder* architecture, is also referred to as sequence-to-sequence in the similar work proposed by Sutskever et al. (2014).

An encoder-decoder or sequence-to-sequence RNN architecture is thus employed to generate an output sequence ($y^1$...,$y^{(n_y)}$), given an input sequence ($x^1$, $x^2$ ...,$x^{(n_x)}$). The last hidden state of the RNN-encoder is employed to compute a generally fixed-size context variable ($C$). This consists of an input sequence semantic summary and

is passed, as an input, to the RNN-decoder (Goodfellow et al., 2016, 396-397).

As pointed out by Bahdanau et al. (2014), the limitation of sequence-to-sequence architecture becomes clear when the context ($C$) output by the encoder RNN has a dimension that is too small to properly summarise a long sequence (Goodfellow et al., 2016, 397). Among the various proposals suggested to overcome this limitation, one consisted of an attention mechanism (Bahdanau et al., 2014). This is a mechanism used to focus on specific parts of the input sequence $C$ vector, which are more relevant for output prediction (Goodfellow et al., 2016, 475-476). In fact, in sequence-to-sequence models, the $C$ vector of the input sequence, which is used as a fixed-length vector, that, as previously stated, has strong limitations when handling long sequences, always remains the same, regardless of the decoder's hidden state. An image of the attention mechanism is given in Figure 5. The central idea behind the attention mechanism is to overcome this problem by keeping the active focus on intermediate encoder hidden states, instead of using only the final states to initialise the decoder.

After encoding the input sequence into a set of internal states ($h_1 ... h_n$), a score for each one ($s_1 ... s_n$) is computed by a FFN, in order to focus solely on the most relevant one. The FFN is trained to identify the relevant states and assign a high score to them, and low scoring internal states will be ignored. Using the generated scores, attention weights ($a_1 ... a_n$) are then computed by a softmax layer (Bridle, 1990). As a logistic function, this squashes values between 0 and 1 and makes sure that weights add up to 1. Finally, the context vector ($C$) is computed, and this will be used by the decoder for prediction of the next word in the sequence. The computation simply consists of multiplying each attention weight by the corresponding internal state (i.e.: $a_1 * h_1$) and summing all the results. The decoder will thus be provided with the concatenation of the $C$ vector with the output of the previous step. Therefore, while being of a fixed size, $C$ is computed dynamically for every prediction. In this way, the decoder will not only receive information about all the sequences ($C$ vector), but also about the internal states which are considered to be more relevant at the specific step by the FFN. The attention mechanism is so powerful that it allows models, not only to perform better, but also to train faster. The only unsolved issue in RNN, regardless of whether an attention mechanism is used or not, is its difficulty in parallelisation, with the next hidden states depending on the previous ones. In order to solve this problem, *Google Brain* team members came up

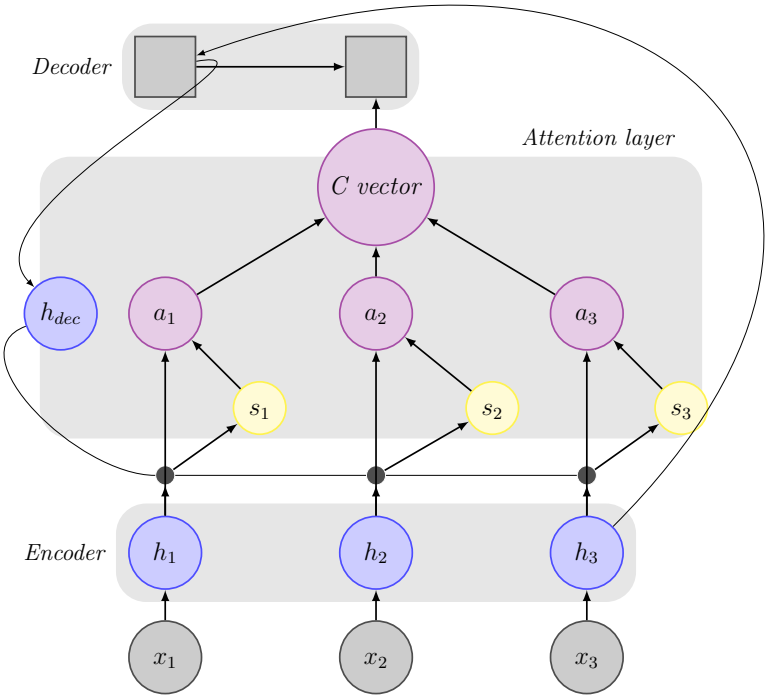



*FIGURE 5 ATTENTION MECHANISM SYSTEM.*

with the *Transformer model* which uses only attention and gets rid of all the other layers, including CNN or RNN, thus making it highly parallelisable, inherently bidirectional, and able to perform efficient computation (Vaswani et al., 2017).

### *Multi-tasking*

A multi-tasking learning approach, as a set of learning algorithms, makes what is learned when performing a task transferable, in order to improve the performance of a similar task. This approach emphasises learning multiple tasks, in a parallel or cascaded manner, that are considered to be relevant by sharing the representation of information at intermediate layers (Caruana, 1997). This approach, that is employing neural networks since early in their development and diffusion (Collobert and Weston, 2007, 2008; Collobert et al., 2011), has been proven to be particularly beneficial for ambiguous data, considering its ability to reduce sparsity, helping to process complex patterns which involve multiple features. This is the case,

for example, of POS tagging (Alonso and Plank, 2016; Bingel and Søgaard, 2017; Hashimoto et al., 2016; Rush et al., 2012; Søgaard and Goldberg, 2016). This is particularly relevant to the morphological richness of Arabic, as addressed by Inoue et al. (2017), or dialectal Arabic, i.e. in Zalmout and Habash (2019a). As an example, the multi-task architecture proposed in Collobert and Weston (2008) is a deep neural network that is *trained jointly* for all the tasks that it is required to perform. In such instances of joint learning, the information processed by deep layers, as well as the features which are trained for one task, are shared across the net, as both can be useful for other related tasks. Therefore, when training NNs on related tasks, sharing what is learned by different layers during joint training improves the generalisation performance, while the last layers of the network may be task-specific (Collobert and Weston, 2008). Another example of multi-task learning involving joint training and shared parameters is that proposed by Zalmout and Habash (2019a). Their proposal consists of a multi-task learning for joint modelling of MSA and Egyptian dialect morphological tagging with adversarial training.[17] They used a Bi-LSTM tagging model, with shared parameters of the hidden layers, to jointly learn different morphological features. In addition, the input (words and characters) representations for both MSA and Egyptian were shared. However, this joint embedding space did not perform well. Furthermore, they also used a unified feature-tag vector representation on all features, through concatenating all the feature-vectors (*v-gender* + *v-number* etc.) for each word in a unique vector, similarly to Inoue et al. (2017). Furthermore, Collobert and Weston (2008) states that:

> *'If one possesses a dataset labeled for several tasks, it is possible to train these tasks jointly in a shallow manner: one unique model can predict all task labels at the same time.'*

However, this is not the case for Tunisian, which, as will be explored in more depth in the next section (Section 5), is a low-

[17] This is a transfer-learning method used for domain-adaptation, in which the adversarial training extracts invariant features from the source domain and applies the extracted invariant features to boosting the model performance on the target domain. In Zalmout and Habash (2019a), the source domain is in both MSA and Egyptian, and the authors used this method to learn the common features between the two systems in order to transfer knowledge from the highest resource (MSA) to the lower resourced system (Egyptian), which are both modelled in the same invariant space.

resourced system, and does not have a large amount of available corpora in order to develop this kind of joint learning. Nevertheless, assuming that we need to produce mutually related levels of text annotation, such as tokenisation and POS tagging, the most intuitive way to produce this information is in a sequential order. Tokenisation and POS tagging are two exact annotation levels as in our corpus. Considering the intuitive sequential order used to organise the different tasks in a cascade, tokenisation is produced as a first level and the hidden representation of the tokenisation module is used as an input for the following one, which, in the case seen in this example, is POS tagging. In this way, each task is performed interdependently, but at the same time, each architecture layer receives the information produced by the previous layers (this topic will be more deeply reflected on in Section 4). A path involving the use of a multi-task architecture was also investigated by Meftah et al. (2020), who proposed an interesting approach based on sequential transfer learning and multi-task learning, with pre-training occurring on a resource-rich domain and fine-tuning on a low-resourced domain in a multi-task modality. In view of the importance of the dataset structure for this type of model (Collobert and Weston, 2008), the next section features records regarding the situation for Arabic processing, in particular for dialectal Arabic. The focus of this will be the already developed and available corpora for dialectal Arabic, and in particular for Tunisian. Finally, we will outline the available corpora for Tunisian encoded in both the Arabic and Latin scripts.

## 5. State of the Art

### *Arabic Natural Language Processing*

As discussed in Section 4, Natural Language Processing is a highly interdisciplinary field with links to computer science, linguistics, cognitive science, psychology, mathematics, etc. Data is the lifeblood of NLP: this is true for rule-based approaches that require carefully crafted lexicons and rules, but also for machine learning approaches that require corpora and specially annotated corpora. As a Semitic language, Arabic has a rich inflectional and derivational morphology, which makes processing Arabic a compelling challenge. This morphological complexity in early automatic Arabic processing was often handled through morphological analysers, such as BAMA (Buckwalter, 2004, 2002). Recently, there has been a substantial increase in the amount of NLP research for morphological analysis,

disambiguation, Part-of-Speech (POS) tagging, and lemmatisation for both Standard Arabic (MSA) and Dialectal Arabic (DA) (Gugliotta et al., 2020). Darwish et al. (2021) notes that automatic processing of Arabic has undergone three waves: the rule-based phase (since the early 1980's), the period of US funding for major ANLP projects (since September 11), and the Arab World ANLP (since 2010). Large projects have been funded for companies and research centers with the aim of developing NLP tools for Arabic and its dialects, and these have began to gradually shift away from rule-based approaches. As already illustrated in Section 4, Machine Learning (ML) requires much less language knowledge and is both faster and more accurate. However, ML requires a lot of data, which is not always easy to collect (i.e. for parallel texts in dialectal Arabic). The second wave has begun to yield early success in examples of blended systems (such as a combination of rule-based morphological parsers and ML disambiguation), such as the Penn Arabic Treebank (PATB) (Habash and Rambow, 2005; Maamouri et al., 2004), or MADA+TOKAN, which is a toolkit for Arabic tokenisation, diacritisation, morphological disambiguation, POS tagging, stemming and lemmatisation (Habash et al., 2009). The period that began around 2010 observed exponential growth in ANLP research originating in the Arab world. According to Darwish et al. (2021), this period also overlapped with two important independent developments: **1.** The rise of Deep Learning and neural models; and **2.** The development of social media, which led to the increase in available data for research. As an example, Zalmout and Habash (2017) presented a model for Arabic morphological disambiguation based on Recurrent Neural Networks (RNN). The authors based their work on the Penn Arabic Treebank and used LSTM by showing that this model has a significant performance. In addition, the success of word embedding (Mikolov et al., 2013; Pennington et al., 2014) being trained on non-annotated data and the resulting improved performance for NLP tasks has also contributed to the growth of interest in ANLP (Al Sallab et al., 2015; Farha and Magdy, 2019; Soliman et al., 2017).[18] Table 11 reports some of the

[18] In recent years, work with contextualised embedding trained on non-annotated data, such as BERT (Devlin et al., 2018), is increasing. Such research constitutes promising possibilities for improving many ANLP tasks. Some examples of this include Arabic BERT (Safaya et al., 2020), AraBERT (Antoun et al., 2020), GigaBert (Lan et al., 2020), and Marbert (Abdul-Mageed et al., 2020).

most recent and well-known tools in the field of automatic processing of MSA morpho-syntax.

However, as mentioned just above, in order to train Deep Learning models to perform tasks efficiently, a large amount of structured data is required. Building up resources is a long and expensive task that requires significant work over long periods of time. As seen in Section 3, corpus building is often tackled by teams of linguists or lexicographers. NLP relies heavily on the existence of corpora to develop and evaluate its models, and the performance of NLP applications depends directly on the quality of these corpora (Darwish et al., 2021).

MSA undoubtedly has an advantage over dialectal Arabic in terms of data collection, due to the large amount of data presented online, even before the emergence of social networks. In fact, MSA is used to write newspaper articles, which have been used to build MSA corpora in some cases, as in Selab and Guessoum (2015) and El-Haj and Koulali (2013). Corpora do not necessarily need to have annotations, as in the case of the ArTenTen corpus (Arts et al., 2014), the LDC's Arabic Gigaword (Parker et al., 2011) or the Qatar Arabic Language Bank (QALB), which was built with the aim of supporting automatic correction of MSA spelling (Habash et al., 2013). On the other hand, the most notable annotated resources include the LDC's Penn Arabic Treebank (PATB), which provides a relatively large MSA corpus that is morphologically analysed, segmented, lemmatised, tagged with fine-grained parts of speech, diacritised and parsed (Maamouri et al., 2004), the Prague Arabic Dependency Treebank (PADT), which was the first dependency representation for Arabic (Smrz et al., 2002), and the Columbia Arabic Treebank (CATiB) (Habash and Roth, 2009; Taji et al., 2019). The latter was actually an effort to develop a simplified dependency representation with a faster annotation scheme for MSA. As such, it consists of a small dependency treebank of travel domain sentences in MSA, created by translating 2000 selected sentences from the Basic Travelling Expression Corpus (BTEC) (Takezawa et al., 2007).[19] A different form of annotation concerns resources built for Sentiment Analysis (SA), such as the Large Scale Arabic Sentiment and Emotion Lexicon (ArSEL) (Badaro et al., 2018), the Sentiment Lexicon for Standard Arabic (SLSA) (Eskander and Rambow, 2015), the Opinion Corpus

[19] The BTEC is a multilingual spoken language corpus containing tourism-related sentences similar to those that are usually found in phrasebooks for tourists going abroad.

| Tool | Brief Description | Reference |
|---|---|---|
| ARLSTem | An Arabic Stemmer (removes adfixes). | (Abainia et al., 2017) |
| CALIMA Star | An Arabic morphological analyser and generator. | (Taji et al., 2018) |
| analyzs | CALIMA Star is part of the CAMeL Tool toolkit. | (Obeid et al., 2020) |
| MADAMIRA | Provides diacritisation, lemmatisation, morpho- | (Pasha et al., 2014) |
| | logical analysis and disambiguation, POS-tagging, | |
| | stemming, glossing, tokenisation, base-phrase | |
| | chunking and NER for MSA and Egyptian Arabic. | |
| Farasa | Uses independent models for tokenisation and | (Abdelali et al., 2016) |
| | POS-tagging of MSA. | (Darwish et al., 2017) |
| YAMAMA | A morphological analyser for MSA and Egyptian. | (Khalifa et al., 2016b) |
| CamelParser | Provides Arabic syntactic dependency analysis. | (Shahrour et al., 2016) |
| NUDAR | Universal Dependency Treebank for MSA. | (Taji et al., 2017) |
| Stanford Arabic Tools | Parser, word segmenter, and POS tagger for MSA. | (Chen and Manning, 2014) |
| | | (Monroe et al., 2014) |
| | | (Toutanova et al., 2003) |
| Stanza | Python NLP Package for many languages including MSA. | (Qi et al., 2020) |

*Table 11. MSA morpho-syntactic work.*

for Arabic (OCA) (Rushdi-Saleh et al., 2011), the Multi-Genre Corpus for MSA Subjectivity and Sentiment Analysis (AWATIF) (Abdul-Mageed and Diab, 2012), and the Large-scale Arabic Book Review data set (LABR) (Aly and Atiya, 2013). Regarding the construction of MSA corpora, further efforts have been devoted to the construction of parallel corpora, including the United Nations Parallel Corpus (Ziemski et al., 2016), where MSA documents are aligned to their corresponding version in Chinese, English, French, Russian and Span-

ish, an Arabic-Japanese corpus, which was manually aligned at sentence level (Inoue et al., 2018), and a subtitled parallel corpus covering 60 languages including MSA, which was a release of an extended version of the OpenSubtitles collection of parallel corpora (Lison and Tiedemann, 2016).

### *Processing of Arabic Dialects*

With the new availability of easily accessible dialectal Arabic data, there is growing interest in applying ANLP to dialectal Arabic processing (Al-Sabbagh and Girju, 2012; Bouamor et al., 2014, 2018; Diab et al., 2010; El-Haj, 2020; Gadalla et al., 1997; Harrat et al., 2014; Sadat et al., 2014a; Salama et al., 2014). In particular, when narrowing the field to NLP work applied to DA, two main macrostrategies, which are aimed at overcoming the lack of data for DA, can be observed: **1. The adaptation of MSA systems to DA processing**, such as (David et al., 2006), who exploited the Penn Arabic Treebank (Maamouri et al., 2004) and used explicit knowledge about the relationship between MSA and Levantine Arabic. In addition, (Duh and Kirchhoff, 2005) constructed a POS tagger for Egyptian through a minimally supervised approach by using the CallHome Egyptian Colloquial Arabic Speech (ECA) corpus (Canavan et al., 1997). Regarding dialectal Arabic parsing, David et al. (2006) addressed the problem of parsing transcribed spoken Levantine Arabic (LA) using explicit knowledge about the relationship between LA and MSA. **2. The constitution of new resources not based on MSA-DA relations**, as lexicons, corpora or dialectal NLP systems. Regarding lexicons, the Algerian lexicon of Abidi and Smaïli (2018), or the Moroccan one, built by Tachicart et al. (2014), have already been mentioned. Concerning corpora, there are different types, including parallel corpora, raw single-variety corpora, annotated corpora or treebanks, such as the LDC's Levantine and Egyptian Arabic Treebanks (Maamouri et al., 2014). There is also a distinction between written-based corpora and speech-based, such as Fisher Levantine Arabic Conversational Telephone Speech (Maamouri et al., 2007). In contrast, the Levantine Dialect Corpus (Shami), is a written-base corpus covering the dialects of Palestine, Jordan, Lebanon and Syria and containing 117,805 Twitter sentences (Kwaik et al., 2018). Shami is not annotated, while Curras is a Palestinian Arabic annotated corpus (Jarrar et al., 2017) consisting of approximately 56,000 tokens. It presents both lexical and morphological manually checked rich annotation. When building same,

the authors employed MADAMIRA (Pasha et al., 2014), as did the authors of SUAR, a semi-automatically annotated Saudi corpus, which contains 104,079 words. Moreover, in this case, the annotation had been manually checked (Al-Twairesh et al., 2018). By exploiting web data, Alsarsour et al. (2018) built the manually-annotated Dialectal Arabic Tweets data set (DART). This corpus is a collection of approximately 25,000 tweets annotated via crowd-sourcing. Diab et al. (2010) built COLABA, an Arabic corpus that was built for NLP resources covering four Arabic dialects: Egyptian, Iraqi, Levantine, and Moroccan. The authors utilised MAGEAD (Habash and Rambow, 2006) and the Buckwalter morphological analyser and generator for MSA and DA (Buckwalter, 2004). The COLABA corpus is part of the COLABA project objectives, namely:

> '*an initiative to process Arabic social media data such as blogs, discussion forums, chats, etc. Given that the language of such social media is typically DA, one of the main objective of COLABA is to illustrate the significant impact of the use of dedicated resources for the processing of DA on NLP applications*.' (Diab et al., 2010).

The objectives of dialectal corpus building include that of developing dedicated tools for a specific variety, such as the Columbia Arabic Language and Dialect Morphological Analyser (CALIMA) for the Egyptian dialect (Habash et al., 2012b). For the same dialectal variety, Darwish et al. (2018) proposed a POS tagger based on a Conditional Random Fields (CRF) sequence labeller.[20] The system works also on Levantine, Maghrebi and Gulf dialects. Concerning the Gulf Arabic, Khalifa et al. (2017) built a morphological analyser covering over 2600 verbs, while Khalifa et al. (2020) proposed a full morphological analysis and disambiguation system for Gulf Arabic based on a morphologically annotated corpus of Emirati Arabic (Khalifa et al., 2018). The authors employed 200,000 words selected from the Gumar corpus (Khalifa et al., 2016a), while for the annotation, they used the MADARi interface (Obeid et al., 2018), which produces morphological annotation and spelling corrections. MADARi was initially using MADAMIRA Egyptian, but was later extended with CALIMAGLF for better lexical matching. Alharbi et al. (2018) presented a POS tagger for the Arabic Gulf dialect, using a Bi-LSTM labeller, which showed better results in comparison with a Support Vector

[20] CRFs are a class of statistical modelling method applied to the performance of conceptual tagging at word level. These models mainly exploit features based on n-grams (Dinarelli et al., 2011; Lafferty et al., 2001).

Machine (SVM) model.[21] In addition, Zalmout et al. (2018) presented a neural morphological tagging and disambiguation model for the Egyptian dialect. The system particularly relies on LSTM and CNN models for generating character embedding, with various extensions to handle noisy and inconsistent content. For the Moroccan and Sanaani Yemeni dialects, Al-Shargi et al. (2016) built a morphologycal analyser, which was trained on a morphologically annotated corpus, that the authors manually constructed exploiting the DIWAN annotation interface (Al-Shargi and Rambow, 2015). A general dialectal morphological analyser is the Analyser for Dialectal Arabic Morphology (ADAM), which Salloum and Habash (2014) present as comparable in its performances to CALIMA for Egyptian Arabic. Finally, Samih et al. (2017) dealt with dialectal Arabic tokenisation with a neural architecture. The authors addressed the task as a sequence labelling problem at the character level, showing the importance of annotated data in developing ANLP models and that these models are good competitors to state-of-the-art methods.

Considering that the main source of dialectal Arabic data for corpora building is the world wide web, and that both MSA and the various dialects use the same encoding systems in Arabic or in Latin characters, one of the main problems addressed by ANLP is the Dialect Identification (DI), namely, the creation of models for the recognition of dialect, or even regional varieties.[22] In general, these models are trained on multi-dialectal corpora. El-Haj et al. (2018) presented a Machine Learning approach to automatically detecting dialects, which was trained through the Arabic Online Commentary (AOC) data set, comprising four Arabic dialects groups (Egyptian, Gulf, Levant and North African), in addition to MSA (Zaidan and Callison-Burch, 2011). In order to use bivalency and written code-switching as features in the classification process, they created dialect-specific frequency lists to distinguish the vocabularies spoken in each dialect compared to MSA. The authors of AOC data set have been annotating it, via crowd-sourcing, for four years, in order to use it for training a sentence-level DI system, which is an automatic classifier relying on n-gram language models

[21] SVMs are supervised learning models with associated learning algorithms that analyse data for classification and regression analysis.

[22] This problem has also been addressed for oral texts, i.e. in Ali et al. (2015); Biadsy and Hirschberg (2009); Biadsy et al. (2009); Lei and Hansen (2010), in that case where dialect identification is useful in developing robust speech systems, such as speech recognition and speaker identification.

(Zaidan and Callison-Burch, 2014). In addition, Elfardy and Diab (2013) focused on sentence-level DI of Egyptian and MSA and proposed a supervised approach combining token-level DI and other features used to train a generative classifier that predicts the given sentence class. The year before this, the same authors proposed AIDA, a system for dialect identification, classification and glossing on the token and sentence level (Elfardy and Diab, 2012a). On the other hand, Elfardy and Diab (2012b) also presented a set of guidelines for token-level identification of DA. In fact, the same authors, in (Elfardy and Diab, 2012c; Elfardy et al., 2013), further addressed the problem of token-level dialect-identification by casting it as a code-switching problem. Elfardy et al. (2014) then presented the latest version of their system for identifying linguistic code-switching in Arabic text. The system relies on probabilistic language models and a tool for morphological analysis and disambiguation for Arabic to identify the class of each word in a given sentence. In the same year, Darwish et al. (2014) also showed that targeting words that exhibit distinguishing traits is essential for proper dialect identification. Indeed, the authors identified lexical, morphological, phonological, and syntactic features in order to support Egyptian-MSA identification. Starting with Egyptian Arabic, Bouamor et al. (2014) selected a set of 2000 sentences and translated them into four other Arabic dialects (Tunisian, Jordan, Palestinian and Syrian) to present the first multi-dialectal Arabic parallel corpus for DI. The original sentences in Egyptian were extracted by the Egyptian-English corpus built by (Zbib et al., 2012), who employed non-professional translators to perform the Egyptian-English translation on Mechanical Turk. Tachicart et al. (2017) also presented a Moroccan dialect and MSA DI system. The authors relied on two different approaches: rule-based (relying on stop-words) and that based on statistics (using several machine learning classifiers). The statistical approach outperformed the rule-based approach, wherein the SVM classifier was more accurate than the other statistical classifiers. For the identification of the same varieties as in El-Haj et al. (2018), Ali (2018) proposed a character-level convolutional neural network model working with dialect embedding vectors. As regards the latter, Shon et al. (2018) extracted dialect embedding with an end-to-end dialect identification system and a Siamese neural network.[23] By exploiting the MADAR corpus (Bouamor et al., 2018), which will

[23] Networks designed on the end-to-end principle collect application-specific features at the communicating end nodes of the network, rather

be presented below, Salameh et al. (2018) presented a fine-grained dialect classification covering 25 Arab-city dialects in addition to MSA.

Regarding the construction of a dialectal Arabic corpora, first of all, the 'problem' of non-standardised encoding affects all Arabic dialects with different degrees of conventionalisation of the systems. This problem was addressed by Habash et al. (2012a) by presenting the Conventional Orthography for Dialectal Arabic. This is a set of guidelines, initially dedicated only to the Egyptian dialect, but later extended to other dialects, such as Algerian (Saadane and Habash, 2015), Tunisian (Zribi et al., 2014), Maghrebi (Turki et al., 2016) or Gulf Arabic (Khalifa et al., 2016a). Finally, Habash et al. (2018) proposed *CODA Star*, which acts as a common set of guidelines with enough specificity to help create dialect-specific conventions. A second matter for consideration, while building a dialectal Arabic corpora, is the need for guidelines for their annotations, an issue addressed, for example, by Habash et al. (2008), Zaghouani and Charfi (2018b) and Elfardy and Diab (2012b). The first of these works had the aim of producing the basis for the annotation of data collection suitable for dialect identification and ANLP, highlighting dialect switching between MSA and DA. The *dialectness* (collocating each word in a continuum ranging from 0 to 4, where 0 is *pure* MSA) of each word is annotated, and relies on lexemes and inflectional morphemes. In terms of the second of these works, the authors have presented an annotation pipeline. The guidelines that they wrote for creating a large manually annotated Arabic author profiling data set from various social media sources (Zaghouani and Charfi, 2018a) will be presented below. The third work exhibits a simplified set of guidelines for detecting MSA-DA code-switching and foreign words on the word and token level.[24] The MADAR corpus, which is a parallel of 25 Arab-city dialects in addition to the pre-existing parallel set for English, French and Modern Standard Arabic (MSA), built by (Bouamor et al., 2018) has already been briefly mentioned. This corpus was created by translating, 2000 French and English sentences, selected from the Basic Traveling Expression Corpus (BTEC) (Takezawa et al., 2007), into the different dialects. In addition to the 2000-sentence strong corpus (CORPUS-25), the authors also selected 10,000 other sentences representing five cities: Doha,

than at the intermediate nodes, such as gates, that exist to build the network.

[24] The assigned class labels are: MSA, Dialectal, Both, Foreign, Named Entity, Unknown and Typo.

Beirut, Cairo, Tunis and Rabat (CORPUS-6). The translators, from the *Ramitechs* company,[25] started from French or English, and were asked to use Arabic script, avoid any code-switching and to be internally consistent in spelling words. They were not provided with any orthographic guidelines.[26]

As noted on several previous occasions, the arrival of social media has allowed researchers to build corpora, extracting data directly from the Internet. There are many dialectal-Arabic corpora based on web data. In tables 9 and 10, we report some of the best-known Arabic-encoded corpora, excluding Tunisian corpora, that will be addressed separately. While ANLP researchers are making strong efforts in building DA corpora, Arabic dialects still lack fine-quality word embedding. As already stated in Section 4, word embedding is crucial to many ANLP tasks. However, this relies on large volumes of annotated and non-noisy data. Erdmann et al. (2018) addressed the problem of noise in multi-dialectal word embedding by collecting a multi-dialectal corpus by concatenating pieces from six different corpora: Almeman and Lee (2013); Bouamor et al. (2018); Jarrar et al. (2017); Khalifa et al. (2016a); Zaidan and Callison-Burch (2011); Zbib et al. (2012). These corpora are outlined in tables 9 and 10.

As has emerged from tables 9 and 10, dialectal Arabic corpora have multiplied in recent years, with most of them being concerned with the varieties of Egyptian Arabic, Gulf Arabic, and the two macro-areas of Levantine and North African or Maghrebi Arabic. However, for the specific varieties of the last geographical group, only three corpora exist for the most investigated variety (Algerian), which are very different from each other as they are based on very divergent fields, i.e. one is extracted from newspapers and is specific to code-switching with French (Cotterell et al., 2014), one is dedicated to YouTube comments (Abidi et al., 2017), and the other is based on an English lexicon automatically translated into Algerian and built for the purpose of Sentiment Analysis (Guellil et al., 2018).

### *Tunisian Processing*

Regarding Tunisian, Younes et al. (2018a) state that

[25] `http://www.ramitechs.com/`. Consulted on 7th April 2021.
[26] For further information about the translation process, see Bouamor et al. (2018).

| **Corpus References** | **Brief Description** |
|---|---|
| COLABA | Egyptian, Levantine, Iraqi and Moroccan Arabic web crawled |
| (Diab et al., 2010) | from blogs. Annotations include POS tagging. |
| Arabic On-line Commentary | MSA, Egyptian, Levantine and Gulf Arabic. 855,000 words. |
| (Zaidan and Callison-Burch, 2011) | Annotated for DI. |
| (Zbib et al., 2012) | Levantine-English and Egyptian-English parallel corpora, |
| | crowd-sourced and translated on *Amazon's Mechanical Turk*. |
| | The corpora respectively consist of 1.1 million and 380,000 words. |
| YADAC | Twitter, blogs and forums in Egyptian, Levantine and Gulf |
| (Al-Sabbagh and Girju, 2012) | Arabic. Tokenised and POS-tagged (tagset adapted from |
| | PATB (Maamouri et al., 2004)). 11 million words. |
| No-named | Forums, comments and blogs web crawled in Gulf (14.5 million), |
| (Almeman and Lee, 2013) | Levantine (10.4 million), Egyptian (13 million) and North African (10 million). |
| YouDACC | YouTube Dialectal Arabic Commentaries in: |
| (Salama et al., 2014) | Egyptian, Gulf, Iraqi, Maghrebi and Levantine. |
| | Comment-annotation with user provenience for DI. |
| Callhome | Egyptian Arabic-English telephone conversations translations. |
| (Kumar et al., 2014) | Each conversation between native speakers is about 5-30 min. |
| | Gender, age, education and accent of speakers were also added. |

*Table 12. Dialectal-Arabic corpora before 2016.*

| Corpus References | Brief Description |
|---|---|
| SANA | MSA, Egyptian and Levantine dialects, with English glosses. |
| (Abdul-Mageed and Diab, 2014) | Manually annotated for subjectivity and sentiment analysis. |
| | 224,564 entries. |
| no-named | Multi-Dialectal Arabic Twitter Corpus. |
| (Mubarak and Darwish, 2014) | 175 million messages annotated with user locations. |
| No-named | Algerian dialect-French code-switched corpus, extracted from |
| (Cotterell et al., 2014) | Algerian newspaper website. 339,504 comments (+ Arabizi). |
| Gumar | Gulf AD. 110 million words from 1200 forum novels, encoded in |
| (Khalifa et al., 2016a) | Gulf Arabic CODA, inspired by Habash et al. (2012a) |
| no-named | Two corpora for Moroccan (64,000 words) and Sanaani Yemeni |
| (Al-Shargi et al., 2016) | (32,500 words) Arabic, morphologically annotated with DIWAN |
| | (Al-Shargi and Rambow, 2015). Texts are in specific CODA. |

*Table 9. Dialectal-Arabic corpora before 2016. (Continued)*

'*especially with regard to annotated corpora, the Tunisian dialect still lacks large consistent corpora allowing the exploration of innovative methods for its automatic processing*.'

Likewise, Guellil et al. (2019) states that an important

'*number of works are focused on multi-dialects. For the rest of the work on dialects, they mostly focused on Saudi, Gulf and Egyptian Arabic. The work on Maghrebi dialets is less sizeable, particularly for Tunisian. The situation is worse for the Palestinian and Levatine dialect. No work deals with Arabizi identification for distinguishing between Algerian, Tunisian, Egyptian Arabizi, etc.*'

In fact, there is a good quantity of corpora including or focusing on Tunisian Neo-Arabic. However, there are not a lot of Tunisian corpora available for free. In tables 11 and 12, Tunisian corpora in Ara-

| Corpus References | Brief Description |
|---|---|
| AraSenTi-Tweet | 17,573 Saudi tweets semi-automatically annotated into four classes: |
| (Al-Twairesh et al., 2017) | positive, negative, neutral and mixed. |
| Curras | 56,000 tokens in Palestinian Arabic, annotated using MADAMIRA and |
| (Jarrar et al., 2017) | DIWAN tools (Al-Shargi and Rambow, 2015; Pasha et al., 2014). |
| no-name | Arabic-Egyptian comparable Wikipedia corpus, that contains |
| (Saad and Alijla, 2017) | 10,197 aligned documents. |
| CALYOU | 12.7 million Algerian YouTube comments, aligned using Word2Vec |
| (Abidi et al., 2017) | (Mikolov et al., 2013). |
| SentiAlg | Algerian sentiment corpus based on an Algerian sentiment lexicon |
| (Guellil et al., 2018) | automatically built by translating two English lexicons (SOCAL and |
| | SentiWordNet (Esuli and Sebastiani, 2006; Taboada et al., 2011)) |
| | using Glosbe API (Guellil and Faical, 2017). It contains Arabizi. |
| No-name | 1/4 billion tweets from 29 major Arab cities. |
| (Abdul-Mageed et al., 2018) | Data is tagged at city level. |
| QADI | Automatically collected dataset of 540,000 tweets from 2,525 users |
| (Abdelali et al., 2020) | across 18 Arab countries. |

*Table 10. Most recent Dialectal-Arabic corpora.*

bic script are outlined (those in Arabizi will be outlined separately, namely in Table 13), highlighting whenever these are freely available.

As in the results shown in Table 11 and 12, among the Tunisian-based freely available corpora there are the Tunisian Dialect Corpus Interlocutor (TuDiCoI) (Graja et al., 2010) and the Spoken Tunisian Arabic Corpus (STAC) (Zribi et al., 2015), which is also morpho-syntactically annotated with a tag set based on Tunisian-*Al-Khalil* conventions (Zribi et al., 2013b), adopted for the purpose of the

STAC. The STAC (Zribi et al., 2015) is composed of 42,388 words collected from audio files downloaded from the web (as files from TV channels and radio stations) and then transcribed using the OTTA convention (Zribi et al., 2013a). OTTA is a Tunisian dedicated convention for orthographic transcription, which is oriented to Tunisian phonetics, unlike *CODA Star* (Habash et al., 2018), which is MSA-orthography-oriented, and was made to represent 28 Arab city dialects and to facilitate their processing. The STAC also contains 30 minutes taken from the TuDiCoI corpus (Graja et al., 2010), which is a domain-specific spoken-dialogue corpus that gathers a set of conversations between staff and clients recorded in railway stations. TuDiCoI consists of 21,682 words in client turns (Graja et al., 2013). Another domain-specific corpus is the Tunisian Arabic Railway Interaction Corpus (TARIC) (Masmoudi et al., 2014b), which is also based on spoken conversations between staff and clients at Tunisian train stations. It contains 20 hours of TUN speech, manually transcribed into Arabic characters by three university students (Masmoudi et al., 2014b). TARIC is a task-oriented resource built for Automatic Speech Recognition (ASR). There are also two parallel corpora, one of which is the Parallel Arabic DIalect Corpus (PADIC) (Meftouh et al., 2015, 2018), which is composed of 6,400 sentences in six Arabic dialects aligned at sentence level and encoded in Arabic script. The dialects represented in PADIC are: Algiers' and Annaba's Algerian dialects, Sfax's Tunisian dialect, Damascus' Syrian dialect, Gaza's Palestinian dialect and the Moroccan dialect. The approach to the building of PADIC is quite similar to Bouamor et al. (2014), which started by collecting and manually translating 2000 words into various dialects, starting from the Egyptian part of the Egyptian-English corpus built by Zbib et al. (2012). The Tunisian sentences included in PADIC, as well as the Palestinian and Syrian ones, have been translated by native speakers starting from MSA translations of Algerian texts. The latter have been collected by recording conversations and TV broadcasts such as shows and movies (Meftouh et al., 2015). Additionally, the MADAR parallel corpus (Bouamor et al., 2018) was initially a collection of translated sentences from English and French into Arabic dialects. In particular, as mentioned earlier, the source texts were selected from the Basic Traveling Expression Corpus (BTEC) (Takezawa et al., 2007). With regard to the latter two instruments, these are therefore parallel corpora, acting as the raw material for several natural language processing tasks, such as cross-language applications including machine translation, bilingual lexicon extraction and multilin-

gual information retrieval (Meftouh et al., 2015). However, it is also necessary to take into account that the texts collected in these resources are not examples of spontaneous conversations. A version of MADAR encoded in *CODA Star* has recently been released. The MADAR CODA Corpus contains 10,000 sentences written in CODA, along with their original raw form. The sentences are part of the MADAR CORPUS-6 that includes Beirut, Cairo, Doha, Rabat, and Tunis dialects. The corpus was created by a human annotator with the aid of a bootstrapping technique, and was manually validated (Eryani et al., 2020).

Besides the corpora, there have been a number of pieces of NLP work focused on building lexicons, such as Boujelbane (2013); Boujelbane et al. (2013a,b); Sghaier and Zrigui (2017). In particular, the latter two describe the methods used for building a bilingual MSA-Tunisian lexicon by transforming PATB Maamouri et al. (2004) into Tunisian through a rules-based system based on MSA-Tunisian relations. Boujelbane et al. (2015) projected an MSA corpus, and in particular its lemmas, to a Tunisian corpus by relying essentially on the bilingual MSA-Tunisian lexicon built by Boujelbane (2013); Boujelbane et al. (2013b). Hamdi et al. (2014) built *lexica* (lexicon of verbs, of de-verbal nouns and of particles) with the aim of building a Part-of-Speech tagger for Tunisian. By contrast, Sadat et al. (2014b) instead built a Tunisian-MSA lexicon of 1,600 Tunisian words manually translated into MSA, in order to develop a rule-based translation system. Specific ontologies were then created to support dialogue systems for certain contexts, such as those in Graja et al. (2011a,b, 2015), Karoui et al. (2013a,b), based on the TuDiCoI corpus (Graja et al., 2010) again referred to in Table 11. Among the Tunisian ontologies already mentioned, AebWordNet (Karmani and Alimi, 2015; Karmani et al., 2014; Moussa et al., 2015) also developed starting from the English-Tunisian Peace Corps Dictionary (Abdelkader, 1977).

Section 5 provided an insight into how many studies have been carried out for the Dialect Identification task. The work that also includes Tunisian can be found in Belgacem et al. (2010), Lachachi and Adla (2015, 2016a,b), Neifar et al. (2014), Sadat et al. (2014a), Hassine et al. (2016) and Masmoudi et al. (2016, 2018). Concerning the morpho-syntactic analysis, a rule-based morphological analyser has been presented by McNeil (2012) and Karmani and Alimi (2015). On the other hand, Zribi et al. (2016) focused on sentence boundary detection of Tunisian transcriptions by employing three different methodologies, i.e. one based on linguistic rules, another with a

| Corpus References | Brief Description |
|---|---|
| OrienTel | Transcription of real conversation into raw database for |
| (Iskra et al., 2004) | Gulf, Levantine, Egyptian and Maghrebi Arabic |
| | (Morocco, Algeria, Tunisia and part of Libya). |
| Part of the Orédule project | Transcription of 10 h of TV and radio broadcasts in Tunisian, |
| (Belgacem, 2009) | Algerian and other ADs, for system of speech-recognition. |
| TuDiCoI | Tunisian Dialect Corpus Interlocutor. Spoken corpus of 1,825 |
| (Graja et al., 2010, 2013) | dialogues recorded in Tunisian railway stations, transcribed in |
| *Freely available* | Buckwalter. 7,814 words annotated at multiples levels. |
| Tunisian Arabic Corpus | 881,964 words from the web and other resources McNeil (2015). |
| (McNeil and Faiza, 2011) | The corpus can be queried online via RegEx, Stem, or the exact word. |
| STAC | Spoken Tunisian Arabic Corpus. |
| (Zribi et al., 2013b, 2015) | 3 h 20 min of TV and radio broadcasts transcribed in OTTA conventions |
| *Freely available* | (Zribi et al., 2013a) and annotated at multiples levels, including POS. |
| Tunisian version of ATB | Built in order to train a Stanford POS Tagger adaptation to Tunisian. |
| (Boujelbane et al., 2014) | 2 h 15 min of transcribed Tunisian speech, manually annotated. |
| Multidialectal Corpus | Multidialectal Arabic Parallel Corpus: 2000 sentences in MSA and |
| (Bouamor et al., 2014) | DAs (Tunisian) + English, manually translated from Egyptian. |
| TARIC | Raw corpus of 20 h of real conversations manually transcribed into Arabic |
| (Masmoudi et al., 2014a,b) | script. The corpus was built for Tunisian Automatic Speech |
| *Freely available* | Recognition. |

*Table 11. Tunisian Neo-Arabic corpora, excluding Tunisian Arabizi corpora*

| Corpus References | Brief Description |
|---|---|
| Oral Tunisian of Media (Boujelbane et al., 2015) | Raw corpus of 5 h 20 min of radio and TV shows, transcribed in CODA-TUN (Zribi et al., 2014). The corpus has been built for language modelling aims. |
| Test corpus (Hamdi et al., 2015) | For Tunisian POS-tagger building. 805 sentences containing 10,746 tokens, annotated using Lemma and POS tags. |
| DI Corpus (Lachachi and Adla, 2015) | Raw corpus of Moroccan, Tunisian and three Algerian dialects collected from radio stations and TV show speeches, for automatic dialect identification. |

*Table 11. Tunisian Neo-Arabic corpora, excluding Tunisian Arabizi corpora (Continued)*

statistical approach, and the latter being a hybrid approach based on the previous two. Zribi et al. (2013b) adapted the MSA morphological analyser *Al-khalil* (Boudlal et al., 2010) for Tunisian, using a set of MSA verbs and nominal patterns transformed into Tunisian patterns, as well as the definition of a set of Tunisian affixes. Zribi et al. (2017) proposed a disambiguation method for Tunisian morphological analysis. As has already been made clear, Hamdi et al. (2015) exploited the relationship between MSA and Tunisian in order to build a POS-tagger for the latter, while Mekki et al. (2017) focused on Tunisian Treebank building.

### *Tunisian Arabizi Processing*

As briefly outlined above, a number of studies on dialectal Arabic have been set up with the aim of solving the lack of a standard orthography through several different methods, such as rule-based, conventional orthography, language modelling and machine learning approaches. However, transliterations involving dialectal Arabic were mainly treated by adopting language models and orthographic conventions (Younes et al., 2020). Nowadays, the most extensive Conventional Orthography for Dialectal Arabic is *CODA Star*, a unified set of guidelines for 28 Arab city dialects (Habash et al., 2018).

| Corpus References | Brief Description |
|---|---|
| (Adouane et al., 2016) | 579,285 words, collected for language identification with ML models. 7 Arabic dialects (with Tunisian) + MSA and Berber. |
| (Hassine et al., 2016) | Real conversation oral recording raw corpus for Moroccan-Tunisian recognition system based on SVM and NN classifiers. |
| (Sayadi et al., 2016) | 5,514 tweets (49,940 words). Annotated for sentiment analysis. |
| (Lachachi and Adla, 2016a) | Raw corpus of radio stations and TV show speeches. |
| PADIC, *freely available* (Harrat et al., 2017) (Meftouh et al., 2015, 2018) | Parallel corpus built translating 5 dialects, including Tunisian, into MSA. 6,400 sentences for each dialect, annotated for LI at sentence-level. |
| (Saadane et al., 2017) | 4 dialects + MSA (19,100 Tunisian messages). Annotated for LI at word-level. |
| (Aridhi et al., 2017) | 86,940 words annotated for LI at word-level. |
| (Mekki et al., 2017) | 492 sentences, annotated with sentence boundaries, tokenisation and syntactic analysis as a Treebank of 928 syntactic trees. |
| (Hassine et al., 2018) | Real oral conversation recordings collected into raw corpus. Built for dialect recognition based on hybrid techniques. |
| (Masmoudi et al., 2018) | Corpus collected from mixture of contents (web and others). |
| MADAR, *freely available* (Bouamor et al., 2018) | Parallel translated corpus of 25 dialects + MSA, English and French. 47,466 words obtained by translating BTEC corpus (Takezawa et al., 2007). |
| (Saadane et al., 2018) | 102,000 Tunisian comments together with 4 dialects + MSA. Annotated for LI at word-level. |
| Arap-tweet (Zaghouani and Charfi, 2018a) | 2.4 million tweets covering 16 Arabic countries. Built for LI. Annotated at sentence-level with user metadata (age and provenience). |

*Table 12. Most recent Tunisian Neo-Arabic corpora, excluding Tunisian Arabizi corpora.*

> '*As a general rule, CODA* uses MSA-like orthographic decisions (rules, exceptions and ad hoc choices), e.g., cliticizing single letter particles, using Shadda for phonological gemination, using Ta-Marbuta, Alif Maqsura, silent Alif in Waw-Alif of plurality, and spelling the definite article Al morphemically.*'[27]

The main CODA features consist of:

1. Having a single orthography for each word;
2. Supporting Natural Language Processing for Dialectal Arabic;
3. Employing the Arabic script;
4. Unifying all Arabic dialects under a unique encoding framework;
5. Balancing morpho-phonological individual dialect representation and the exploitation of MSA-Dialectal Arabic similarities for unification purposes.

The first widespread convention for dialectal Arabic, named CODA, was proposed by (Habash et al., 2012a). Regarding Tunisian, the first CODA was adapted to Tunisian from the Egyptian CODA, resulting in the CODA-TUN (Habash et al., 2012a), by Zribi et al. (2014), and has been used, for example, by Boujelbane et al. (2015), while the Egyptian CODA was instead used by Al-Badrashiny et al. (2014) to build 3ARRIB, i.e. a transducer trained at the character level to generate all possible transliterations for input words in Arabizi. Additionally, Bies et al. (2014) used the Egyptian CODA to build a parallel corpus of Egyptian Arabizi-Egyptian Arabic of informal written conversation, such as for SMS and chats. Another work on social media Arabizi transliteration into Arabic script is that of Eskander et al. (2014). Regarding the CODA-TUN applied to Tunisian Arabizi (TA), the study carried out by Masmoudi et al. (2015), who realized a TA-Arabic script conversion tool, which was implemented with a rule-based approach, should also be mentioned. With these aims in mind, the scholars collected a Tunisian corpus of texts written in Arabizi, automatically generating a set of its possible transliterations into CODA-TUN (Zribi et al., 2014) by applying transliteration rules. The best candidate was then manually selected by the annotator.

In recent years, several researchers, when dealing with the task of Arabic language transliteration, have adopted language models based on different approaches. These researchers included the work of Chalabi and Gerges (2012), mentioned above, who implemented a hybrid rule-based and language models approach. Another ex-

[27] `https://camel-guidelines.readthedocs.io/en/latest/orthography/`. Consulted on 7th April 2021.

| **Corpus References** | **Brief Description** |
|---|---|
| MultiTD corpus | Gathers texts (32,848 words) from many sources: |
| (Bouchlaghem et al., 2014) | social networks, written pieces of theatre, dictionaries, |
| | transcriptions of spontaneous speech, etc. |
| LETD *contact* | Electronic Tunisian Dialect: 43,222 messages in Latin |
| (Younes and Souissi, 2014) | script exclusively constructed from the web. |
| (Masmoudi et al., 2015) | Social media and SMS corpus (870,904 words) |
| | for CODA transliteration. |
| TLD & TAD *contact* | 420,897 & 160418 words in a Latin & Arabic script raw |
| (Younes et al., 2015) | corpora, automatically constructed from the web. |
| Sentiment Analysis corpus | Comments from the Facebook pages encoded in Arabic |
| (Ameur et al., 2016) | and Latin scripts. Annotated for SA with 9 different tags. |
| TSAC, *downloadable* | Tunisian Sentiment Analysis Corpus. 17,060 Tunisian |
| (Mdhaffar et al., 2017) | Facebook comments, in Arabic and Latin script. |
| | Manually annotated with polarity. |

*Table 13. Corpora including Tunisian Arabizi.*

ample is the work of van der Wees et al. (2016), who implemented an Arabizi-Arabic transliteration pipeline based on Arabizi character mapping to Arabic sequences with the aim of improving Arabizi-to-English Statistical Machine Translation (SMT). With regards to the relationship between translation and transliteration, Guellil et al. (2017a,b) presented a system for translating between Algerian Arabizi and MSA. The authors proposed a comparison between statistical and neural translations after the translation step, showing that the quality of transliteration directly affects the translation. In another work, namely Guellil et al. (2018), in order to process Arabizi, the transliteration task has been shown to be a necessary previous step in order to reduce Arabizi ambiguity, which is a consequence of its lack of spelling conventions. In that particular study, the main task was sentiment classification of an Algerian Arabizi corpus. It ap-

pears that most of the work has been focused on automatic transliteration from Arabizi into the Arabic script, such as in Chalabi and Gerges (2012), Darwish (2014) and Al-Badrashiny et al. (2014). These three works are based on a character-to-character mapping model that aims to generate a range of alternative words that must then be selected through a linguistic model. In particular, the aim of Darwish (2014) was to classify words as either Arabizi or English by using sequence labeling based on Conditional Random Fields (CRF). Eskander et al. (2014) focused on foreign words and the automatic processing of Arabic social media text written in Roman script. Guellil and Azouaou (2016) presented an approach for social media dialectal Arabic identification based on supervised methods, using a prebuilt bilingual lexicon, of 25,086 words, which was previously proposed in (Azouaou and Guellil, 2017; Guellil and Faical, 2017). The primary goal of this approach has been to identify words encoded in the Algerian Arabizi dialect. Continuing the focus on Algerian Arabizi, Guellil and Azouaou (2017) proposed a Syntactic Analyser for the Algerian dialect. The authors focused on Algerian Arabizi and in order to build their parser, enriched a basic dictionary that contains translations between the Algerian dialect and French, with different phonological extensions. 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). As regards Tunisian dialect transliteration in general, Younes et al. (2020) noticed that very few studies have been developed, namely only Masmoudi et al. (2015); Younes et al. (2016, 2018b). Only a few studies have explored Deep Learning-based approaches for TD language transliteration, with one of those (Younes et al., 2018b) being a sequence-to-sequence approach and obtaining a word-level accuracy of 95.59%. Younes et al. (2020) therefore proposed an exploration into a word-level approach based on modelling transliteration as a sequential labelling task implemented using different techniques (CRF, BLSTM and BLSTM-CRF). Finally, when evaluating the transliteration approaches at the sentence-level of their Latin-encoded Tunisian Corpus content, this showed that the BLSTM-CRF model outperformed the other models. As in the case of Younes et al. (2020), most research on the automatic processing of Arabizi involves a preliminary phase of collecting a corpus, so as to train the models or test the various experiments. As we have seen in Section 3, creating corpora from scratch is a very time and energy consuming practice; moreover, it is evident that experiments can only really be reproducible if carried

out on the same data used by scholars, if same is made available. However, this is not always the case. In the following table (Table 13), the Arabizi Tunisian corpora are therefore listed, including a note on whether they are available for free download (*downloadable*) or by contacting the author (*contact*).

# 3. TUNISIAN ARABISH CORPUS - TARC

After defining, in the previous chapters, the motivations that led to the building of a Tunisian Arabizi corpus from scratch and the methodology adopted by referring to the current state of the art, this chapter will move on to a description of the specific operations that have led to the realisation of this goal. The phases that will be described trace the path from its beginnings, that is to say, from data collection up to decisions relating to the selection and collection of metadata (see Section 1). We will also describe the semi-automatic annotation phases of the corpus in the levels that it is composed of: transliteration in Arabic characters, tokenisation, part-of-speech tagging and lemmatisation. The phases and experiments that led to the identification of the best strategies for achieving the goal will be reviewed (see Section 2 and 3). Finally, the Multi-Task Sequence Prediction Architecture that was built in order to produce the different annotation levels that make up TArC from the Arabizi text will be described (see Section 4). This architecture was used to precisely achieve the main goal, but at the same time also constitutes a useful tool which makes it possible to extend the same work in the future. Through the same architecture it will be possible to extend TArC with further data that will be annotated automatically and which requires minimum participation in manual correction, and to adapt the same tool to the treatment of other Arabic dialects.

## 1. Data collection

### *Data & metadata selection*

Considering what was discussed in the previous chapter (see Section 2) about the nature of online conversation applied to the case of Arabizi, three main sources of DNW were identified in Tunisian Arabizi to collect the data we wanted to include in our corpus. The three sources of these written texts were conversations on social networks, on forums, and texts extracted from blogs. Regarding the latter genre, we have selected two Tunisian Blogs, *Hatem Jouher*'s blog and the *TounsiaDigordia*'s blog, from whom we have requested written permission to use their texts.[1] The extraction of this data was possible through a web scraping tool, *ParseHub*, which permits precise extraction of selected textual data and metadata.[2] Web scraping is a method of extracting data from several websites into a spreadsheet or database. The selection of data from blogs was actually limited by the identification of Tunisian blogs encoded in Arabizi. The only interference exercised on this data is that of having chosen one blog produced by a female author and one by a male author, in order to move as close as possible towards an equal representation of the two sexual genders. However, it was necessary to devise a more complex system for data extracted from social networks and forums, in order to automatically identify Tunisian texts encoded in Arabizi which were available on the Internet, in order to crawl (select and extract) the texts and metadata of the users.

In order to quickly identify Tunisian texts, or at least those produced in Tunisian territories, one option would be to use Twitter data, taking advantage of the possibility to extract text based on the location of users. However, Twitter's textual data was not suitable for our purposes. In fact, our goal was to collect texts of variable lengths, but possibly medium-to-long, so that they would be as contextualised as possible and rich in linguistic information. As such, the crawling process had to be organised differently.

The choices made resulted in a selection made of keywords that were unequivocally Tunisian, but which were generic enough to not influence the data. In order to ensure that the work was as unbiased as possible, a corpus collection procedure was adopted, and was composed of the following: steps:

---

1 The bloggers' profiles can be visited at the following links: `https://draft.blogger.com/profile/08907253046329916251` and `https://tounseyyadigordeya.wordpress.com/a-propos/`.

2 The ParseHub website is at `https://www.parsehub.com/`.

1. Thematic category identification.
2. Matching of categories with sets of semantically related TA keywords.
3. Texts detection based on keyword search.
4. Texts and metadata extraction.

*The first step.* In order to build a corpus which is as representative as possible of the Tunisian system, identifying wide thematic categories that could represent the most common topics of daily conversations on DNW was considered to be useful. In this regard, two instruments with a similar thematic organisation were employed:

- 'A Frequency Dictionary of Arabic', and in particular its 'Thematic Vocabulary List' (TVL) (Buckwalter and Parkinson, 2014).
- 'Loanword Typology Meaning List', which is a list of 1460 meanings (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 headings. Considering that the boundaries between some categories are very blurred, some categories have been merged, such as 'Body' and 'Health', (see Table 14).[3] Some others have been eliminated, due to being irrelevant for the purposes of this research, including 'Colors', 'Opposites', 'Male names'. In the end, the fifteen macro-categories listed in Table 14 were obtained.

*The second step.* By aiming to easily detect texts and their respective Source URLs, without introducing relevant query biases, it was decided to avoid the use of category names in the query, instead generating a range of keywords (Schäfer and Bildhauer, 2013). Therefore, each category was associated with a set of keywords in Arabizi belonging to the basic Tunisian vocabulary. A semantic category with three meanings was found to be enough to obtain a sufficient number of keywords and URLs for each category. For example, for the category 'family', the meanings: 'child', 'marriage', 'divorce' were associated with all their TA variants, resulting in an average of 10 keywords for each macro-category (see Table 14). Concerning the keywords associated with the fifteen macro-categories in Table 14, in order to avoid possible bias resulting from a specific spelling of an

[3] The Tawḥīd, reported in Table 14, is the principle of the uniqueness of God, the first affirmation of the Muslim faith expressed in the šahāda, and a testimony of faith in Islām.

employed keyword and the associated words, native speakers were involved in the process. The latter were asked to write the keywords (in Arabizi) in all possible ways, which was presented to them written in Arabic script. A total of four native speakers of different sexual genders and age ranges (specifically, two males from the 25-35 and 50+ ranges and two females from the same ranges) were involved.[4] Considering that these keywords did not serve to identify the exact texts to be extracted, but only served to lead us to Tunisian language forums, where we could then find people of different age groups, sexual genders, and backgrounds, these expedients were considered to be sufficient to avoid introducing bias during the text search phase.

*The third step.* The adopted Parsehub collection process consists of:

1. Manually insert the selected URL into the Parsehub client new-project section. The system loads the page and shows, to the user, the possibility of selecting the data he/she wants to collect by clicking on an interactive area of the client, where the web page is shown.
2. Click on the first message. The system automatically identifies the structure of the page and highlights all the other elements which are similar to the one clicked on.
3. Click on the first highlighted message. The Parsehub system will select all the other similar elements until the end of the page and will contextually structure the selected data into a CSV/Excel or JSON format. The second possibility was chosen for this study.
4. Add relative select comments. These comprise information belonging to the selected data (i.e. the text's message), such as, for example, the message metadata. After the first relative selection, the system automatically repeats the same process on the other selected elements.
5. Get data, once selection has concluded.

Considering that Parsehub permits the saving of the data selection project registering the different steps described above, it is not necessary to repeat these for each page in case where the same forum is used. Therefore, only the URL had to be changed to crawl a page from the same forum or blog.

*The fourth step.* After a manual check of the collected data to ensure that they were indeed Tunisian texts and relevant to the purposes of

[4] Among the two males, the youngest was from the Djerba area, and the other from Tunis; in terms of the female representation, the youngest woman was from the Zarzis area and the other was from Bizerte.

| **Macro-Categories** | **Words Associated** |
|---|---|
| 1. Family<br>*son, wedding, divorce* | weld, wild, 3ars, 3ers, tla9, 6la9, tlaq, 6laq, tle9, tleq, 6leq |
| 2. Clothing<br>*dress, shoes, T-shirt* | robe, lebsa, rouba, sabat, spedri, spadri, marioul, maryoul, meryoul, merioul |
| 3. Automobiles<br>*gasoil, engine, occasion* | mazout, motor, moteur, motour, forsa |
| 4. Animals<br>*cock, dog, cat* | sardouk, kelb, kalb, 9attous, gattous |
| 5. Body and Health<br>*sick, doctor, health* | maridh, marith, mridh, ettbib, tbib, sa77a, sa7a, sahha, saha |
| 6. Food and Drinks<br>*recipe, kitchen, enjoy* | recette, r7, koujina, coujina, bchfé, bchfè, bchfe, bechfé |
| 7. Public Transport<br>*bus, taxi, train* | lkar, kar, lbus, bus, taxi, trino, trinou, train |
| 8. Nature and Travel<br>*nature, trekking, trip* | tabi3a, ettabi3a, randonné, ri7la, rihla |
| 9. Services and Technology<br>*telephone, TV, services* | telifoun, talifoun, bortable talefza, talvza, telé, tv, khidmet, 5idmet |
| 10. Sport<br>*sport, football, work out* | spor, sport, foutbol, lkoura, koura, lentrainement |
| 11. Careers<br>*payment, salary, job* | 5las, khlas, l5las, chahriya chahria, l5edma, l5idma, 5edma, khedma, khidma |
| 12. Climate and Weather<br>*summer, spring, hot* | sif, essif, rbi3, errabi3, s5oun, skhoun |
| 13. Emotions and Values<br>*happy, sad, fear* | farhan, far7an, far7ana, hzin, 7zin, 7azin, l5ouf, khouf, khaouf, 5ouf |
| 14. Political and Social Relations<br>*party, democracy, freedom* | 7izb, hizb, 7ezb, hezb, dimo9ratia, dimo9ratiya, dimoqratia, dimokratia, hourriya, 7ourriya, horriya, 7orriya |
| 15. Faith<br>*religion, god, Tawḥīd* | din, eddin, dine, eddine, rabbi, rabb, tawhid, taou7id, taouhid, taw7id |

*Table 14. Example of the fifteen thematic categories*

the research, approximately 25,000 words and their metadata were collected as the first part of the corpus. As can be seen in the following sections, the TArC data was augmented at a later time using the same techniques as those described here.

Regarding the collection of metadata, while this could be performed automatically using the tool for blogs (and some forums), for social networks it was necessary to proceed manually. The only metadata that was automatically collected by the web crawling system, and developed for purpose, was the date of publication of the text and the link to the profile of the user who was the author. In a second phase using that link, user data was manually collected where publicly available.

Concerning the date of publication, this was useful for assisting with the selection of data extracted from the web, in an attempt to make the corpus as balanced as possible in the diachronic representation. In fact, texts were chosen and collected from a range of approximately 10 years in order to observe their diachronic variation, where it was present. In particular, this selection was done with the purpose of observing to what extent the TA orthographic system has evolved toward a writing convention.

With regard to user metadata, the information posted by users was extracted, and the focus was on the three types of information generally used in ethnographic studies:

1. Gender: Male (M) and Female (F).
2. Age range: [-25], [25-35], [35-50], [50+].
3. City of origin.

In a second stage, cities were traced back to the governorate to which they belonged, in order to simplify analyses of diatopic variation on TArC (as presented in Chapter 4). In fact, this user metadata was collected with the general intention of making diastratic variation analyses possible on the TArC data.

## 2. Corpus structure - annotation levels - first phase

Since Arabizi is a spontaneous orthography of Tunisian, it was considered to be important to adopt the CODA* guidelines (Habash et al., 2018) as a model for producing a unified encoding in Arabic script for each Arabizi token.[5]

> '*CODA* (pronounced CODA Star, as in, for any dialect) is a conventional orthography for dialectal Arabic. It is designed primarily for the purpose of developing computational models of Arabic dialects.*'

[5] CODA* guidelines are available at `https://camel-guidelines.readthedocs.io/en/latest/orthography/`.

This choice was primarily made to facilitate correspondence with existing tools and studies for Arabic processing. Since CODA* is a unified convention, the specific guidelines for Tunisian Arabic (CODA TUN) were also taken into account (Zribi et al., 2014). These guidelines adhere to those of the CODA convention project. In fact, the negation mark that the CODA TUN proposes for maintaining the MSA rule, that is, the insertion of a space between the first mark of negation and the verb, was executed to make CODA TUN conform to the first CODA (Habash et al., 2012a). However, as explained by Zribi et al. (2014), in Tunisian Arabic, this rule does not have a correspondence on the phonetic level, but should still be applied in order to preserve consistency across CODA guidelines. Indeed, in our transliteration we report what was produced in Arabizi following the CODA* rules, while in the lemmatisation we report the lemma of the verb. At the same time, on the tokenisation level, we segment the negative verb into its constituent parts: the negation prefix, the conjugated verb, and the second negation mark (if present).

In order to guarantee accurate transliteration, the first 6,000 tokens were manually annotated across all the annotation levels. Some annotation decisions were taken before this step, with regard to specifically Tunisian Arabic features:

1. Concerning foreign words, we transliterated Arabizi words into Arabic script, except for code-switching terms. Regarding the latter, at the Arabic encoding level (and the other levels, as can be sees later in this chapter), the token with the tag later used for its classification was replaced, i.e. '*foreign*'.
2. With regard to typographical errors and typical problems related to informal writing habits on the Internet, such as repeated characters to simulate the prosodic features of the language, it was not possible to maintain all of these characteristics in the Arabic script level, according to CODA* conventions.
3. Regarding the phono-Lexical exceptions, the grapheme < ڤ >, [g], was only used in acclimatised loanword transliterations, such as the loan verb /rīgəl - yrīgəl/ 'to arrange', from French *régler*. Instead, the Hilalc phone [g], which can be found, for example, in the Arabizi word *gamra*, /gamṛa/, has been transliterated and lemmatised with the grapheme < ق >, [q].
4. Concerning the glottal stop, as explained in CODA TUN, the real initial and final glottal stops have almost completely disappeared from Tunisian Arabic or have instead become a long vowel (Zribi et al., 2014, 2357, 2359). Glottal stops still remain in some words

that are treated as exceptions, such as the verb 'to ask a question' <سأل>, /s?al/, also pronounced /shal/. Indeed, glottal stops are only transcribed when they are usually pronounced, and if not, glottal stops are not written at the beginning of the word or at the end, neither in the transliteration, nor in the lemmas.

### *Transliteration Model*

In order to facilitate easier and faster corpus collection, a semi-automatic procedure based on sequential neural models was adopted (Dinarelli and Grobol, 2019a,b). To begin, we first used the first set of approximately 6,000 manually transliterated tokens as the training and testing data sets in a 10-fold cross-validation setting. Cross-validation is a statistical technique that can be used in the presence of a good numerosity in the observed sample. In particular, the so-called k-fold cross-validation consists of the subdivision of the total data set in k parts of equal numerosity, and, for every step, the k part of the data set comes to be the validation part, while the remaining part always constitutes the training set. As such, the model is trained for each of the k parts, therefore avoiding problems concerning the improper representation of the data which is typical of the data set subdivision in only two parts (that is training/validation sets) (Russel et al., 2009, p. 708). In other words, the observed sample is divided into groups of equal numerosity, one group is iteratively excluded at a time, and one tries to predict it with the non-excluded groups, in order to verify how good the prediction model used is.

As mentioned previously, the Arabizi tokens that were the result of code-switching were classified as *foreign*. However, the decision to include a level of classification of Arabizi tokens, in *arabizi*, *foreign* and *emotag*, came at a later date, with the aim of solving the following question. That is to say that, considering that the goal was to transliterate Tunisian Arabizi (not the result of code-switching) into Arabic characters, if the tokens resulting from code-switching were also transliterated, this would have generated noise for an automatic and probabilistic model. In fact, the correspondence between the phonological and orthographic planes of the two systems necessarily presents asymmetries. In this first phase of testing, however, not having yet identified an ideal treatment solution for these tokens, it was simply decided to remove them from the data. After removing the French tokens, the data was reduced to approximately 5,000 tokens. By combining the index of sentences, paragraphs, and tokens

in the corpus, whole sentences can be reconstructed. Specifically, these 5,000 tokens corresponded to approximately 300 sentences, which is far too few to be used for neural model learning.[6]

In order to reduce the variety of data to be predicted, the Arabizi tokens were divided into characters and the Arabic tokens were divided into morphemes, and treated each token itself as a sequence. In doing so, the model was required to learn how to map Arabic characters into Arabic morphemes. The 10-fold cross validation with this setting gave a token-level accuracy of approximately 65%. This result was not satisfactory on an absolute scale, however it was encouraging when taking into account the small size of the sample. Using this model, approximately 700 additional tokens were automatically transliterated into Arabic morphemes. The manually corrected additional tokens were added to the training data of the neural model, and a new 10-fold cross-validation was performed. After this second stage, the results were approximately 70% on average. This procedure was repeated three times in total, so as to transliterate the first four blocks of TArC. The average accuracy on the fourth block was approximately 76%.

Once this level of accuracy was reached however, it became apparent that it would be very difficult to further improve the performance of this model, and a new strategy was evaluated that would allow us to:

1. Automate some of the operations performed manually, such as the classification of Arabic tokens, in order to solve the problem of noise caused by the presence of code-switching tokens.
2. Improve the performance of the transliteration task through the information contained in the other levels of analysis.
3. Share information across levels in order for same to improve one another.

From this insight came the idea of building a multi-tasking architecture, which will explored in the next section.

## 3. Corpus structure - annotation levels - second phase

### *String classification*

As introduced in the previous paragraph, after the first experiments on the task of transliterating data into Arabic characters, the need

[6] As shown by preliminary experiments that yielded rather poor results, with below 50% accuracy on average.

to treat this data differently depending on its graphematic nature was imperative if also summarising it that way. It is true in fact that, as mentioned several times in previous chapters, Arabizi is encoded in Latin characters, and this facilitates switching to different systems that use the same encoding in Latin characters. However it is clear that the rules underlying the encoding of the system 'Tunisian Arabic' and that of the system 'French' for example, are divergent. The two in fact, despite sharing coding conventions, probably due to a strong linguistic contact continued over time, are governed by divergent historical-linguistic traditions. For example, looking at the TArC data, the first code-mixing element to be found is the French word 'mais', pronounced [mɛ]. However, in Arabizi, the correct French spelling is maintained. As for Arabizi encoding words of Arabic origin, the Tunisian DNW system tends to reflect the phonetic realisation of Tunisian, rather than orthographic. If we had wanted to transliterate French code-mixing into Arabic characters as well, we would have had to choose between two options:

1. encoding French code-mixing being based on its pronunciation.
2. encoding French code-mixing being based on its spelling.

Both of these options were discarded for the following reasons: the first case would have increased the difficulties faced by the system at the level of disambiguation. For example, 'mais' would have to be encoded as <ما>, which is the encoding for both the first element of the circumfixed negations and the word designating 'water'. In the second case, it would be necessary to encode 'mais' as <مايس> by generating a token that has no meaning in Arabic characters, which can be referred to as an artificial token. These artificial tokens would also contain generated noise for the neural system, for example, as in the phonetic realisation of French vowel clusters or silent consonants. The system would have definitely had more difficulties in generalising the transliteration rules of Arabizi.

The initial solution was to remove the code-mixed tokens. However, this could not be considered to be a permanent solution. It was necessary to treat the two systems differently, and therefore first of all, it was also necessary to prefix all levels of treatment we wanted to provide for TArC, i.e. a level of classification of the text into code-mixing elements or elements with an Arabic-Tunisian etymological base. In short, we made a functional classification for the automatic treatment tasks, separating what should be transliterated from what should not be. In this way, the three macro-classes of *foreign*, *arabizi*

and *emotag* were born. In addition, a class was needed also for all the paratextual elements which were typical of DNW. It was desirable to keep these for their potential usefulness in analyses, but they also created a good amount of noise for the neural system, due to largely being constituted of punctuation marks. Initial classification experiments were conducted using 'more controlled' data than that of TArC. 'More controlled' here refer to the use of data that contained only elements that belonged to one of the classes or a mixture of them. In fact, the texts of TArC contain texts of social networks, which are sometimes very informal, and at other times artistic or contain word-play, and therefore my be more difficult to handle from a computational point of view.

The token-level classification has been carried on through a RNN character-level model pre-trained on:

1. Hussem Ben Belgacem's French dictionary, consisting of 336,351 tokens.[7]
2. A Tunisian Arabish dictionary of 100.936 tokens, resulting from the merging of the following datasets:
   a. The Tunizi Sentiment Analysis Tunisian Arabic Dataset (Fourati et al., 2020).[8]
   b. The TLD dataset of Arabish (Younes et al., 2015).

Therefore, the procedure employed to bootstrap the classification of the TArC data included performing model pre-training on an external corpus, by exploiting the above resources. This decision was made to speed up the process. In fact, as will be outlined below, an Arabizi corpus annotated with a classification level was initially built, instead of manually annotating a first block of TArC, as it was, for example, for the transliteration of the bootstrapping process. Moreover, in this way, it is possible to obtain a good quantity of pre-training samples without employing TArC data, such that classification was only performed once the model was already pre-trained.

The building of this pre-training corpus included the following steps:

1. Easily classifying all the tokens of Belgacem's French dictionary using the *foreign* tag.

[7] The dictionary is available at the following link: `https://github.com/hbenbel/French-Dictionary`.

[8] The dataset is available at the following link: `https://github.com/chaymafourati/TUNIZI-Sentiment-Analysis-Tunisian-Arabizi-Dataset`.

2. Merging the Tunisian corpora into a unique corpus (TUN).
3. Using Belgacem's French dictionary to remove French words from TUN.
4. Building a small dictionary of emoticons (EMO) by using the *emoji* python-library.[9] This was classified with the tag *emotag*.
5. Employing the EMO corpus to extract, from TUN, the smileys and emoticons contained in same, and obtaining a clean Tunisian Arabizi corpus that has been classified, at the token level, with the *arabizi* tag.
6. Merging and shuffling all the three classified corpora together, and obtaining the pre-training corpus.

The model pre-trained on the above data reached 97% accuracy. At this point, it became possible to start an iterative procedure for the TArC text classification. The custom of dividing TArC into blocks of approximately 6,000 tokens was maintained. After classifying each block, it was manually checked and added to the training data in order to improve the performances thereof. However, the accuracy on the last TArC block (the seventh) remained at 97%. As will be seen in Section 4, when it was decided to perform all the TArC annotation levels through an architecture, the classification task was repeated, together with the other tasks. As shown in Table 17, the architecture accuracy of the classification task slightly improved in the last step (97.63%) in comparison to the mono-task model described here. The same table also summarizes the results obtained at each step during the TArC classification.

### *Tokenisation*

Considering the high index of synthesis that characterises Arabic dialects, which is albeit less than that of Standard Arabic, it was considered fundamental to have, as a third level of TArC annotation, a tokenisation level.[10] In Tunisian Arabic, each string consists of one root that has a basic meaning. Affixes are added to make the stem word incorporate the subject, direct and indirect objects, numbers, gender, definiteness, etc., as shown in the following example (20). The tokenisation at string level deals with reducing each string to

[9] Available at `https://pypi.org/project/emoji/`.

[10] The syntheticity index of a language is a coefficient that describes, for the language in question, the degree of concentration of morphological functions within a word.

its morphemic clitic components, concatenated by the symbol +. This also reduces data sparsity and decreases the number of out-of-vocabulary (OOV) words, supporting the other annotation levels. Only those tokens classified as *arabizi*, and thus transliterated, have been tokenised, following the D3_BWFORM configuration scheme of MADAMIRA tools, where basically all clitics are tokenised, including the article (Pasha et al., 2014).[11]

(20) Arabizi: *dbrthelik*

*/dəbbart-hā-lə-k/*

<دبّرتها لك> (transliteration in CODA*)

<دبّرت+ها+ل+ك> (tokenisation)

'I turned things around for you' (lit.:I found her (the solution) for you).

This scheme was chosen because the BWFORM method is the only one, among those provided by MADAMIRA, to support dialectal Arabic input text (which is specifically trained on Egyptian Arabic). Regarding sentence boundaries, a boundary token, i.e. '<eos>' (meaning 'end of sentence'), was automatically added after each final punctuation mark, and same were then checked manually. Manual correction was quite important in this case because, especially in the case of texts coming from social networks, it is possible to find very long texts without any punctuation marks. Consequently, in addition to the automatically inserted end-of-sentence marks, at least another 30% of marks were added manually.

### *Part-of-Speech*

Part-of-Speech tagging is the morphosyntatic annotation of strings. It has been carried out mainly at the morphological level, and in the case of adverbial phrases or interjections being composed of multiple words, also at the functional level (Bender and Lascarides, 2019, 57-58).[12] As can be seen below, it was necessary to incorporate a variation to the chosen POS-tagging system in order to be able to include this additional layer of information. Regarding the morphological POS-tagging level, this describes the morphological nature of each

[11] Version used: MADAMIRA 2.0. D3 BW* schemes (Habash, 2010).

[12] To be precise, we refer here to the notion of transfer as postulated by Tesnière (*translation* in French) (Tesnière, 2015, 46-58).

element of the string, while the other one describes the grammatical function of the whole string. The POS annotation style follows the guidelines of the Penn Arabic Treebank (PATB) (Maamouri et al., 2004). The latter uses the Buckwalter tag set that presents a high degree of granularity. It is precisely because of this granularity and the possibility of tagging each morpheme that Buckwalter's annotation scheme was chosen. Buckwalter's annotation scheme includes the following tags (15). At the beginning of the project, the intention was to use Universal Dependencies (UD),[13] specifically the *New York University Abu Dhabi Universal Dependency Arabic Treebank* (NUDAR) developed by Taji et al. (2017). However, NUDAR is a tagset for Standard Arabic and, in particular, is the conversion of the Penn Arabic Treebank into the syntactic representation system of the UDs through an intermediate dependency representation. For these reasons, a more adequate solution for Tunisian Arabic tagging was evaluated, i.e. the direct use of the Penn Arabic Treebank tagset. In fact, the operation of POS-tagging of dialectal Arabic, and moreover of its informal social networking variety, required a meticulous analysis of each morpheme. Furthermore, it is easier to automatically convert a fine-grained tagset into a more simple one. It seemed to be better to ensure a fine-grained tagging in a first level of analysis of the TArC morpho-syntax, without at all excluding the possibility of generating a conversion of the adopted tag set to the UDs at a later stage. In fact, since the purpose of UDs is to facilitate the creation of treebanks in different languages that are consistent in their syntactic representation, while allowing for the extension of relationships to accommodate language-specific constructs, this would be the ideal solution in both NLP and in linguistics for comparative purposes with other languages.

Regarding the functional POS tagging level, it was deemed useful to add information such as the adverbial function of phrases, such as the one shown in the following example (21):

(21) Arabizi: *bsara7a*

*/b-ṣaṛāħa/*

<بـ+صراحة> (tokenisation)

[PREP+NOUN-NSUFF_FEM_SG]ADV (POS)

'Frankly' (lit.: With frankness).

[13] For more information about Universal Dependencies, see `https://universaldependencies.org/introduction.html`.

| Verbs | Nominals | Particles | Other |
|---|---|---|---|
| VERB | NOUN | PREP | PUNC |
| PSEUDO_VERB | NOUN_NUM | CONJ | ABBREV |
| PV | NOUN_QUANT | SUB_CONJ | INTERJ |
| PV_PASS | NOUN.VN | PART | LATIN |
| PVSUFF_DO:<PGN> | NOUN_PROP | CONNEC_PART | FOREIGN |
| PVSUFF_SUBJ:<PGN> | ADJ | EMPHATIC_PART | TYPO |
| IV | ADJ_COMP | FOCUS_PART | PARTIAL |
| IV_PASS | ADJ_NUM | FUT_PART | DIALECT |
| IVSUFF_DO:<PGN> | ADJ.VN | INTERROG_PART | |
| IV<PGN> | ADJ_PROP | JUS_PART | |
| _MOOD:<mood> | ADV | NEG_PART | |
| CV | REL_ADV | RC_PART | |
| CVSUFF_DO:<PGN> | INTERROG_ADV | RESTRIC_PART | |
| CVSUFF_SUBJ:<PGN> | PRON | VERB_PART | |
| | PRON_<PGN> | VOC_PART | |
| | POSS_PRON_<PGN> | | |
| | DEM_PRON_<GN> | | |
| | REL_PRON | | |
| | INTERROG_PRON | | |
| | NSUFF<Gen><Num><Cas> | | |
| | CASE<Def><Cas> | | |
| | DET | | |

*TABLE 15. THE BUCKWALTER TAG SET*

As can be seen in Example 21, morpho-syntactic annotations have been enclosed within square brackets, outside of which we have provided the functional annotation, since the prepositional phrase performs the syntactic function of an adverb. This variation in the annotation system has been evaluated as enriching the corpus in terms of information, which does not affect the level of compatibility with other corpora annotated according to the same scheme. In fact, in the case where the users of TArC do not want to take into account the functional annotation, they can simply choose to exclude (also at the level of automatic processing) the elements from the square brackets and limit their query to the morphological annotations.

Furthermore, since Buckwalter's tag set was created for the annotation of Standard Arabic, certain tags such as the <CASE>, the <CONNEC_PART> or the <LATIN>, <FOREIGN>, <PARTIAL> and <DIALECT> tags (in Table 15) are never used in the annotation of our corpus, as they are not tags which are compatible with the morphological structure of Tunisian Arabic.

In general, our approach to POS-tagging has been to follow Buckwalter's tag set, trying to remain as faithful to it as possible, by using the Penn Arabic Treebank's annotation guidelines.[14] At the same time, we sought a compromise appropriate to the representation of the morpho-syntax of Tunisian Arabic and a strong consistency in the use of tags in order to ensure a good result of the automatic processes. In order to simplify the work, in the initial phase, the Tunisian MADAR corpus data (Bouamor et al., 2014) was also automatically POS-tagged using the MADAMIRA software (Pasha et al., 2014). The results, as MADAMIRA is trained only for Egyptian among the Arabic dialectal varieties, were not satisfactory for Tunisian Arabic. However, despite this, this step was very useful for starting to familiarise ourselves with a POS-tagged text and to observe the various tag combinations on a Tunisian Arabic text. Finally, the scheme we used for out tag set was the following (scheme 16), in which the ∃ symbol means that the tag can receive additional morphological features, while the symbol Ø means the opposite.

### *Lemmatisation*

As will be discussed later in this chapter, lemmatisation is a substantially important annotation layer. Present in our projects since the beginning, this was the last level we decided to produce for practical reasons. In fact, despite the usefulness of lemmatisation, transliteration in CODA*, tokenisation and POS tagging are indispensable layers for one another and are necessary for TArC completeness and for the analysis we present in Chapter 4. However, even lemmatisation represents a tool of fundamental importance both at the level of linguistic analysis and data processing (Zalmout and Habash, 2019b). The procedure used to generate this layer was the same as that implemented for all the others, i.e., the incremental iterative procedure based on automatic lemma annotation generation (via the Multi-Task Sequence Prediction Architecture 4) and manual correction of data, which was then used to augment the system training data.

[14] In particular, version 3.8 of 2009 was of great assistance.

| Definition | | Tag (stem) | Additional morphological features | | |
|---|---|---|---|---|---|
| **Determiners (Articles)** | | DET | Ø | | |
| **Nouns** | | NOUN | Ø | | |
| | | | -NSUFF | _FEM | _SG |
| | | | | | _PL |
| | | | | | _DU |
| | | | | _MASC | _PL |
| | | | | | _DU |
| | | | _QUANT | -NSUFF | _MASC(_PL/_DU) |
| | | | | -NSUFF | _FEM (_SG/_PL/_DU) |
| | | | | Ø | |
| | | | _NUM | Ø | |
| | | | _PROP | -NSUFF | _FEM_SG |
| | | | | Ø | |
| **Adjectives** | | ADJ | -NSUFF | _FEM/_MASC | _SG/_PL/_DU |
| | | | Ø | | |
| | | DEM_ADJ | _3 | _F/_M | _S/_P |
| | | ADJ_COMP | Ø | | |
| | | ADJ_NUM | -NSUFF | _FEM/_MASC | _SG/_PL |
| | | | Ø | | |
| **Adverbs** | | ADV | Ø | | |
| | | INTERROG_ADV | Ø | | |
| **Prepositions** | | PREP | Ø | | |
| **Pronouns** | | PRON | _1S/1P.../INDEF | | |
| | | INTERROG_PRON | Ø | | |
| | | POSS_PRON | _1S/1P... | | |
| | | DEM_PRON | _1S/1P... /Ø | | |
| | | NUM_PRON | -NSUFF | _FEM/_MASC | _SG/_PL |
| **Verbs** | **Perfect V.** | PV | -PVSUFF_SUBJ: (1S/2S/2P/3MS/3FS/3P) | Ø | |
| | | | | ∃/ Ø | +PVSUFF_IO: (1S/2S/2P/3MS/3FS/3P) |
| | | | | ∃/ Ø | +PVSUFF_DO: (1S/2S/2P/3MS/3FS/3P) |
| | **Passive Perfect V.** | PV_PASS | -PVSUFF_SUBJ (1S/2S/2P/3MS/3FS/3P) | Ø | |
| | **Imperfect V.** | IV | IV(1S/1P/2S/2P/3MS/3FS/3P) [this is a prefix] | Ø | |
| | | | | ∃/ Ø | +IVSUFF_DO: (1S/1P/2S/2P/3MS/3FS/3P) |
| | | | | ∃/ Ø | +IVSUFF_IO: (1S/2S/3MS/3FS/3P) |
| | | | IV(1P/2P/3P) [this is a prefix] | ∃/ Ø | -IVSUFF_SUBJ: (P/2FS/2S) |
| | **Command V.** | CV | -CVSUFF_SUBJ: (2S/2FS/2MS/2P) | Ø | |
| | | | -CVSUFF_SUBJ: (2S/2FS/2MS/2P) | Ø | +CVSUFF_DO: (1S/1P/...) |
| | | | -CVSUFF_SUBJ: (2S/2FS/2MS/2P) | ∃/ Ø | +CVSUFF_IO: (1S/1P/...) |
| | **Pseudo-verbs** | PSEUDO_VERB | Ø | | |
| **Particles** | | PART | Ø | | |
| | | NEG_PART | ∃ | +NEG_PART | |
| | | | Ø | | |
| | | VOC_PART | Ø | | |
| | | FUT_PART | Ø | | |
| | | FOCUS_PART | Ø | +PRON_(1S/1P...) | |
| | | INTERROG_PART | Ø | | |
| **Interjection** | | INTERJ | Ø | | |
| **Conjunction** | | CONJ | Ø | | |
| | | SUB_CONJ | Ø | | |
| **Punctuation** | | PUNC | Ø | | |

*Table 16. The adopted tag set scheme.*

Once again, data from the MADAR corpus was exploited, for which we semi-automatically generated the lemmatisation level, using a first block of the manually lemmatised TArC as training data. The TArC lemmas are encoded in Arabic characters in the orthography also used for the Arabizi transliteration level, i.e. CODA*. This decision was made on the basis of factors of utility in linguistic research on TArC data. In fact, the CODA convention allows for normalisation of the text and a unique coding of the lemma, unlike the Arabizi encoding.[15]

In order to ensure a good consistency and rigor, both linguistically and computationally, the following methodological choices were adopted when annotating the data of this level.

1. Considering that the inclusion of a root level is foreseen in order to simplify the matching of same with other Tunisian corpora or glossaries, it was not considered necessary to refer adjectives or deverbal nouns to the etymological verbal form.
   In this way, we did not have to make a selection between those adjectives (or present and past participles) that are now so crystallised in the language in their nominal or adjectival form that they are perceived as nouns or adjectives rather than deverbal elements.
2. Adjectives with nisba are present at the lemma level in their adjectival, masculine singular version.[16]
3. Collectives nouns that quantify one unit such as /žṛāna/, 'one frog' or /burdgāna/ 'one orange', are always brought back to the masculine singular i.e. /žṛān/ and /burdgān/ .
4. In the case of elements of French origin, but which have also entered the Tunisian lexicon through transliteration, as with the adapted loanword /rəstūrānāt/, 'restaurants', as mentioned in Chapter 1, (Section 2), these have been reported in Arabic characters (رستورانات) also in TArC. However, for their lemma, the

[15] We even considered an encoding in Latin characters, according to the dialectological tradition. However, we assumed that this would create difficulties in writing queries for linguistic analysis, based on the choice of specific graphemes, which are explained in an additional user guide. Additionally, we did not want to rule out the possibility of adding this encoding in the future for the level of annotation of the roots of each lemma.

[16] The nisba adjective is the 'relative' adjective in Arabic, i.e. /tūnsiyy/ means '(smth. or sb.) of Tunis'.

French etymological lemma is reported, i.e. '*restaurant*'.[17]

5. Elements classified as *foreign* or *emotag*, or POS-tagged as *PUNC*, *NOUN_NUM* or *SYM*, show exactly these entries at the lemmatisation level.
6. Interjections, such as /nšāḷḷa/, 'God willing', are reported exactly as in the transliteration level in CODA* (ان شاء الله) and have not been etymologically reduced to a single element, having a unitary semantic value.
7. In the case of articulated nouns (*DET+NOUN*) or items preceded by prepositions, the noun lemma has been reported. In the case of articulated prepositions, the preposition has been reported as lemma. The article constitutes a lemma only when it appears as an isolated token. In the case of morpho-syntactically complex relative items (i.e. *PREP+REL_PRON*), the relative has been reported as a lemma.

As will be explored in even greater depth in Section 4, computationally speaking, the belief is that this level of annotation could make improvements to the CODA* transliteration task. For this reason, following the order of the tasks performed by the multi-task system, it was decided to place the module dedicated to the lemma between the classification module and the CODA* transliteration module.

## 4. Multi-Task Sequence Prediction Architecture

As mentioned above, considering the correlation between the different annotation levels expected for the TArC corpus, it is logical to produce these levels using a multitasking architecture rather than through models dedicated to single tasks. Indeed, a neural architecture factors in some parameters for information that can be shared between tasks, and then uses different modules for each task, which are learned interdependently. One idea was for these shared parameters to aid in disambiguation when annotating other layers. According to our insights, the first level of classification would have operated the selection of tokens to be encoded in CODA* and to then be tokenised, POS-tagged and lemmatised as learning tasks organised

[17] On the other hand, for tokens currently classified as *foreign*, the same treatment is planned to be addressed in the future, i.e. adding a POS-tagging and lemmatisation layer for these as well.

in a chain in a neural network, benefiting from each other at the disambiguation level, thanks to the sharing of parameters along the network. This insight led to the idea of a multi-task neural network, where different learning tasks are organised in cascades (Gugliotta and Dinarelli, 2020a; Gugliotta et al., 2020). An input consisting of text in Arabizi undergoes the following five tasks:

1. *Classification*
2. *CODAfication*
3. *Tokenisation*
4. *POS-tagging*
5. *Lemmatisation*

Regarding the order presented in the above list, this does not represent the final multi-task system order for the tasks organised in a cascade, but instead indicates the order in which the various layers were produced.[18]

As shown in Figure 1, the Arabizi input is converted into context-aware hidden representations by the encoder based on recurrent layers (see Section 4). Steps 2, 3, 4 and 5 consist of interdependent decoders, which not only receive, as an input, the output of each of the previous steps (in the form of a hidden state), but also receive the Arabizi input processed by the encoder (i.e. the encoder output). Regarding the figure, $x$ represents the Arabizi input, which passes through the *Encoder* and is transformed into the encoder hidden state ($h_E$) and sent to different modules dedicated to the individual tasks:

1. $Decoder_{cl}$ for *Classification*
2. $Decoder_{ar}$ for *CODAfication*,
3. $Decoder_{tk}$ for *tokenisation*,
4. $Decoder_{pos}$ for *POS-tagging*.
5. $Decoder_{lm}$ for *Lemmatisation*.

Here, once again, the order does not represent the final order of the annotation levels produced in a cascade by the multi-task system. In fact, if, from the above list, the last module ($Decoder_{lm}$) is included, it is possible to obtain the structure thanks to which we obtained the annotation of the lemmas. However, it has been found that the lemmatisation task may be more helpful for the CODAfication task

[18] The multi-task system is available at the following link: `https://gricad-gitlab.univ-grenoble-alpes.fr/dinarelm/tarc-multi-task-system`.

if placed before it, so in the end, the modules of the multi-tasks architecture are presented in the following order:[19]

1. $Decoder_{cl}$ for *Classification*
2. $Decoder_{lm}$ for *Lemmatisation*.
3. $Decoder_{ar}$ for *CODAfication*,
4. $Decoder_{tk}$ for *Tokenisation*,
5. $Decoder_{pos}$ for *POS-tagging*.

Each decoder gives its predicted output ($\hat{o}_1$, $\hat{o}_2$, $\hat{o}_3$, $\hat{o}_4$ and $\hat{o}_5$) and a hidden state ($h_1$, $h_2$, $h_3$, $h_4$ and $h_5$), while $h_E$ is the encoder's hidden state. Moreover, each decoder has a number of different attention mechanisms (see Section 4), including one for attending to encoder's information, and one for each previous decoder's hidden state. The predicted output is used to learn the single task by computing a single loss (i.e. $\mathcal{L}_1(o_1, \hat{o}_1)$, which refers to the comparison between the predicted output of the *Classification*, in this case, $\hat{o}_1$ with the expected output $o_1$). Each loss ($\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$, $\mathcal{L}_4$ and $\mathcal{L}_5$) is than summed to obtain the global loss of the model ($\mathcal{L}$) and jointly learn all the tasks ($\mathcal{L} = \sum_{i=1}^{5} \mathcal{L}_i(o_i, \hat{o}_i)$). The sum is represented with the circled + symbol at the upper part of the Figure 1

As described in the previous section, in order to obtain the *CODAfication* annotation level, a dedicated mono-task sequence-to-sequence model, the one of Dinarelli and Grobol (2019b), was used. Using this model, annotation was performed up to the fourth block of TArC, obtaining an accuracy of approximately 76%. We identify this procedure as the first phase (described in Section 2). The second phase concerns TArC annotation in all its levels. In addition, in the second phase, which was developed with the multi-task system, we repeated the iterative semi-automatic annotation procedure of the first phase for several levels (classification, tokenisation, POS-tagging and lemmatisation). A summary table of the steps is Table 17.

In order to train the model to annotate the first block of TArC, Tunisian Arabic-encoded texts belonging to the MADAR corpus were employed (Bouamor et al., 2014), as previously introduced. The MADAR texts used amounted to approximately 12,000 tokens, i.e. 2,000 sentences, and were manually checked in order to be consistent, at the orthography level, with the CODA* conventions,

[19] Experiments on the lemmatisation level are not yet complete, which is why we will not report percentages for this annotation level.

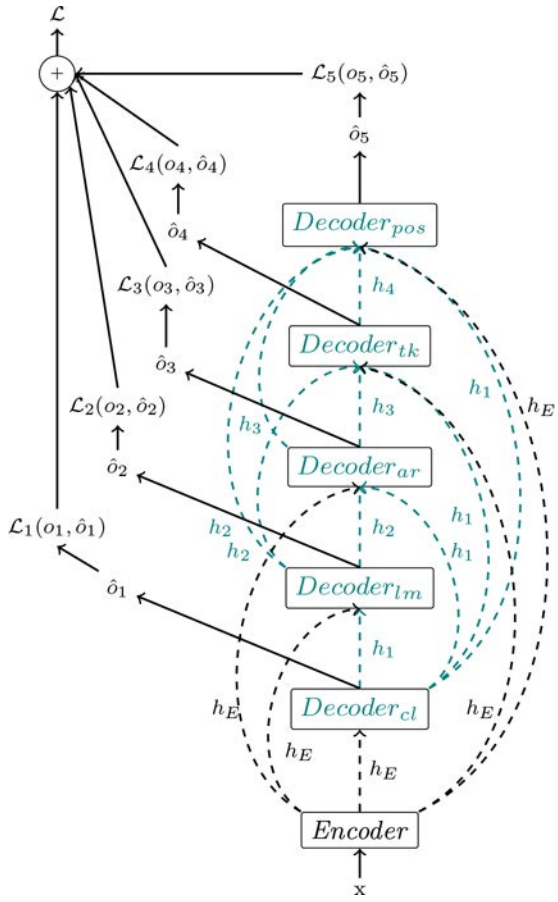



*Figure 1 A high-level schema of our architecture*

not being encoded in CODA*.[20] The same data, as only rough texts, were also annotated on the tokenisation and POS-tagging levels, by exploiting the MADAMIRA tools, and manually performing checking. Regarding the classification level, each token of this data is classified as *arabizi*, in consideration of the efforts done by MADAMIRA builders in order to avoid any code-switching (Bouamor et al., 2014). The MADAR data has been used to train a first model which is useful for starting the iterative procedure of TArC annotation with all its levels

[20] A CODA version of MADAR was recently published by Eryani et al. (2020). We are already planning to use this data for future projects.

(step 0 in Table 17). However, it was also helpful to merge MADAR training data with TArC training data (after manually performing correction) for annotation of the next block. Indeed, this merging was possible until TArC data was already encoded in Arabic-script (until the fourth block). Different solutions were tried for the annotation of the last three blocks:

1. Using only TArC data as training data, since MADAR is not provided with an Arabizi encoding level (Step 4 and Step 5 in Table 17);
2. Using only TArC data as training data, together with *ad hoc* parameter initialisation, in order to exploit the model trained at step 0 (on the MADAR data) (*smart init* steps in Table 17);
3. Provide MADAR with an Arabizi encoding level and merge same with the TArC data (steps found under the header $MADAR_{Arabizi}$+*TArC* in Table 17).

In the following Table (17), it is possible to observe the results of the different strategies tried. The tasks are indicated as **Class** for classification, **CODA*** for the CODA* encoding, **Token** for tokenisation, and **PoS** for POS-tagging. In the same table, 'Train. Tokens' represents the number of tokens used to train the model at each step (the number of tokens from TArC are reported in parentheses, while the rest are tokens from the MADAR corpus).

Regarding Table 17, *Step0* represents the bootstrapping phase of the multi-task system, using only MADAR data. Moreover, $Step0_{complete}$ uses only the MADAR data, but also includes the semi-automatically produced MADAR Arabizi annotation. The next steps (1, 2, 3) also employ the TArC blocks, respectively one block (the first of TArC), two blocks (the first and the second) and three blocks (from the first to the third). As mentioned above, these TArC blocks were already provided with a CODA* encoding level thanks to the preceding models, and were manually checked. For this reason, it is possible to concatenate MADAR and TArC until the third step. In the next steps (*Step4* and *Step5* in the table), the first solution mentioned above was adopted and employed only when the TArC data observed a substantial drop in the tokenisation and POS-tagging accuracies. This is due to the following reasons:

1. Arabizi (less consistent than CODA*) has been used as the input;
2. The presence of less training data;
3. An additional annotation level to be predicted by the system (CODA*/$Decoder_1$ in Figure 1).

| Task | Train. tokens | LSTM | | | |
|---|---|---|---|---|---|
| | | Class | CODA* | Token | PoS |
| Corpus: MADAR | | | | | |
| Step0 | 12,391 | 99.83% | - | 88.83% | 72.71% |
| $Step0_{complete}$ | 12,391 | 99.58% | 76.77% | 74.83% | 67.59% |
| Corpus: MADAR+TArC | | | | | |
| Step1 | 17,261 (4,870) | 92.69% | - | 77.66% | 59.56% |
| Step2 | 22,173 (9,780) | 97.21% | - | 87.53% | 74.30% |
| Step3 | 27,270 (14,870) | 96.69% | - | 91.47% | 76.38% |
| Corpus: TArC | | | | | |
| Step4 | 22,150 | 96.83% | 75.30% | 73.38% | 69.76% |
| Step5 | 27,435 | 97.17% | 75.08% | 73.07% | 66.24% |
| $Step4_{smart\text{-}init}$ | 22,150 | 95.91% | 76.55% | 74.96% | 72.57% |
| $Step5_{smart\text{-}init}$ | 27,435 | 97.08% | 77.83% | 75.69% | 69.76% |
| Corpus: $MADAR_{Arabizi}$+TArC | | | | | |
| Step4 | 34,541 (22,150) | 96.59% | 78.94% | 77.38% | 74.54% |
| Step6 | 46,197 (33,806) | 97.63% | 83.29% | 81.94% | 81.02% |
| $Step6_{Arabic}$ | 46,197 (33,806) | 92.78% | 74.72% | - | - |
| $Step6_{Token}$ | 46,197 (33,806) | - | - | 89.54% | - |
| $Step6_{POS}$ | 46,197 (33,806) | - | - | - | 75.48% |
| Final $Step_{CODA^* input}$ | - | 98.67% | - | 96.78% | 86.31% |

*TABLE 17. SUMMARY OF RESULTS IN TERMS OF ACCURACY.*

Regarding the drop between *Steps4* and *Step5*, this is due to the presence of different text genres in block 5. In fact, *Step4* is mainly constituted of forum texts, while *Step5* texts came from blogs, which are quite different from the previous ones (see Section 2 for theoretical observation about the differences among textual genres and Chapter 4 for the analyses concerning this topic). As mentioned in Gugliotta et al. (forthcoming), another relevant observation that emerges from the performances in these two steps is that the most ambiguous annotation level, which uses Arabizi as an input, appears to be the Arabic-script level.

Therefore, we tried the second solution, represented by the experiments, which results are reported in the entries $Steps4_{smart-init}$ and $Steps5_{smart-init}$. This kind of initialisation, exploiting the model

trained at *Step0* on MADAR texts only, consists of making substantial improvements on each task (excluding the classification one, which has already been solved). Thanks to this experiment, it is possible to verify the usefulness of MADAR data for the multi-task system. In fact, the MADAR data, although generated through procedures that do not exactly coincide with dialectological methodology, represent a useful source for texts that are consistent in their morpho-syntactical structure.[21]

In light of the foregoing, a third method was also tried, with the intention of further improving the results of the three tasks thanks to a fuller exploitation of the MADAR data. Therefore, MADAR data has also been annotated through a semi-automatic procedure at the Arabizi script level. The latter simply consists of inverting the process employed to produce the CODA annotation level from an Arabizi input, and a subsequent manual correction of the Arabizi output obtained from the Arabic-script input. Once the MADAR was also annotated at the level of Arabizi, in order to compare the *Step0* results with a new model trained only on MADAR (but including the Arabizi level), we repeated the *Step0* annotation procedure. The result of this is defined as $Step0_{complete}$ in Table 17. Comparing *Step4* under the header $MADAR_{Arabizi}$+*TArC* and *Step4* produced only with TArC data and with the *smart-init* strategy, it is possible to observe an improvement in the accuracies of all the tasks. Using the MADAR provided with the Arabizi level is definitely the best solution in terms of our aims.

Therefore, TArC block annotation was continued, with the last step (*Step6*) being used to annotate the seventh and last block. This step shows further improvements on all levels. Related to the issue mentioned above and in Gugliotta et al. (forthcoming), namely, the fact that transliteration into CODA* is the most difficult task for the system, additional experiments were performed on individual tasks. These are presented in the table as $Step6_{Arabic}$, $Step6_{Token}$, $Step6_{POS}$. These mono-task experiments show us that all the tasks take advantage of the multi-task organisation, except for the *tokenisation* (89.54%). However, due to the fact that the latter is produced from an input consisting of the correct *CODAfication* level, it is not surprising that the mono-task result is so good in comparison to the multi-task-*Step6*. In fact, this result confirms that:

[21] The texts were first *translated* into Standard Arabic, which was the vehicular language for the retrieval of the Tunisian Arabic texts. This may generate some data bias in the collection phase.

1. The performance of the multi-task system on the *tokenisation* task is largely affected by errors in the *CODAfication* task.
2. The multi-task system helps improve both the *CODAfication* and the *POS-tagging* levels.
3. When given a correct *CODAfication* level, the *tokenisation* task is a relatively easy task to perform.
4. The *CODAfication* is the hardest-to-produce level for the multi-task system.
5. By improving the *CODAfication* task performances, a lot of the other tasks performances could also be improved.

The fact that the transliteration in CODA* represents the most difficult task for the multi-task system is not surprising at all, especially since it is based only on the selection made by the classification level and the Arabizi input. The first of these only selecting a token suitable for transliteration and the second being an informal and non-standardised encoding results in a noisy input for the system.

In order to definitely prove this consideration, the last experiment (*Final Step6*$_{CODA*input}$ has also been reported in Table 17), which presents the highest accuracy values on the different tasks which are jointly performed. For this experiment, the system received the Tunisian texts written in CODA* as an input, instead of their Arabizi encoding.

In order to achieve better multi-task system results on the *CODAfication* task, one last strategy was taken into consideration, which consisted of adding a lemmatisation level. We were already planning to add this layer, but due to time constraints we had considered finalising this issue at a later date.

## 5. TArC's data description

In the preceding chapters and sections, we have tried to express the principles that moved us to pursue the project of creating the TArC corpus. This section summarises the main features of the work in its final state.

As mentioned earlier, TArC is the result of multidisciplinary work with a hybrid approach, which seeks to make the most of the different forces that come together at the intersection of the disciplinary fields in which it arises. TArC was conceived with the intention of trying to respond to a need, i.e. that of making up for the lack of resources for an under-researched language. The entire project was in-

| | *Sentences* | *Words* | | |
|---|---|---|---|---|
| Total | 4,797 | 43,327 | | |
| | | *arabizi* | *foreign* | *emotag* |
| *forum* | 755 | 6,024 | 5,873 | 12 |
| *social* | 3,162 | 11,835 | 3,623 | 598 |
| *blog* | 366 | 5,970 | 694 | 7 |
| *rap* | 514 | 7,681 | 1,009 | 1 |

*Table 18. Some quantitative data of TArC*

spired by the centuries-old tradition of experts, scholars and experts in Semitic languages, and in particular in Arabic dialectology. The aim was to get as close as possible to the ranks of the great linguistic corpora, which were made so remarkable with respect of the linguistic *datum* and the methodological rigor. At the same time, this project also tries to embrace the most advanced techniques of deep learning and its revolutionary power, which today is the engine behind the greatest experiences with artificial intelligence when same is applied to linguistics. It is certain that these systems require an enormous amount of data, which implies an equal amount of effort in terms of quality control. However, by automating the building process, some problems related to the influence of the data during the collection phase can be easily bypassed (e.g. see the *observer paradox* of Labov (1972), which is a much discussed issue in dialectology (Boberg et al., 2018)).

The data collected in TArC, along with various metadata, provides a snapshot of Tunisian Arabic writing in Arabizi over the last ten years. In the following table (Table 18), some quantitative data is reported with respect to the texts collected, and same is organised according to textual genres (*forums*, *social* and *blogs*) and the musical genre of *rap*. The quantities of tokens according to their classification in *arabizi*, *foreign* and *emotag* are also reported.

As seen in Section 3, metadata is fundamental information for text data. As shown in Table 10, the metadata collected in TArC is regarding text genre, publication date, and user information, such as gender, age range and governorate of belonging. It is believed that the information collected could assist with different kind of linguistic research developed through TArC's data. In order to protect the privacy of our informants, only public information was collected. Sensitive data (such as proper names, phone numbers or links to personal pages) was also anonymised by covering same with a dedicated

token. This token had to be encoded in the Arabic script when hiding a proper noun in the *CODAfied* annotation level. For this reason, a Tunisian word, namely <m5abbi>, was directly employed for the level in Arabizi, while <مخبّي> is used for the levels in CODA*. This word means 'hidden' and is accompanied by a numerical code that always identifies the same token that has been covered.

Statistics about the metadata quantity is given in the following tables. Tables 19, 20, 21 and 22 are about the different genres collected in TArC and made available through different files as mentioned in Section 3.[22] The total number of tokens for each text genre is also given in the same tables. The tables show the percentages of information collected for each variable expected from our metadata collection system (i.e. four age ranges: under 25 years, between 25 and 35 years, between 35 and 50 years, and over fifties).

The total percentage is represented by the number of tokens for which information was collected. For the publication date information, the total number of tokens making up the genre collection always corresponds to the total number of tokens reported at the top of the table (i.e. for the Social Networks Table 19, *16,056* is the total number of Social Network tokens and is also the total number of tokens for which the publication date has been collected). It is not the same in the case of the users' metadata, considering that collecting their personal information was not always possible, depending of their profiles and the information which is shared publicly. As such, one an example, in Social Network data, users' provenience is a piece of information collected for 10,980 tokens, age range information for 11,313 tokens and gender information for 12,929 tokens. Percentages are given on the basis of these tokens numbers reported at the top of the table under the voice *tot*. With regard to the information on governorates to which users belong, sometimes users are from cities which are not in Tunisia, and in such cases, the information has been registered in TArC, among the governorate metadata. However, these non-Tunisian cities are not taken into consideration in these metadata tables (Social Networks: 19, Forums: 20, Blogs: 21 and Rap lyrics: 22), but they are included in Table 25, which is a specific table of the registered locations. Finally, as an example, *tot: 10,237* in Table 19 shows the total number of tokens in which a

[22] The reference website is: `https://github.com/eligugliotta`.

| | **Social Networks** | total tokens: 16,056 | | |
|---|---|---|---|---|
| **Years %** | **Years %** | **Govern. %** | **Govern. %** | **Age %** |
| | | tot: 10237 | | tot: 11313 |
| *2005*: *n.o* | *2013*: 9.04% | *Ariana*: 3.31% | *Zaghouan*: 0.66% | *-25*: 25.71% |
| *2006*: *n.o* | *2014*: 0.17% | *Béja*: 0.52% | *Kebili*: 0.12% | *25-35*: 52.21% |
| *2007*: *n.o* | *2015*: 6.12% | *Sousse*: 6.81% | *Mahdia*: 3.2% | *35-50*: 17.99% |
| *2008*: *n.o* | *2016*: 7.26% | *Bizerte*: 3.62% | *Manouba*: 0.81% | *50+*: 4.08% |
| *2009*: *n.o* | *2017*: 2.7% | *Gabès*: 2.13% | *Medenine*: 5.15% | |
| *2010*: 1.12% | *2018*: 14.95% | *Nabeul*: 4.69% | *Monastir*: 3.07% | |
| *2011*: *n.o* | *2019*: 53.25% | *Jendouba*: 1.52% | *Gafsa*: 2.32% | **Gender %** |
| *2012*: 5.38% | *2020*: *n.o* | *Kairouan*: 2.08% | *Sfax*: 5.57% | tot: 12929 |
| | | *Sidi Bouzid*: 1.83% | *Siliana*: 0.42% | ———— |
| | | *Ben Arous*: 2.08% | *Tataouine*: 0.62% | *F*: 42.6% |
| | | *Tozeur*: 0.46% | *Kasserine*: 1.12% | *M*: 57.4% |
| | | *Tunis*: 45.1% | *El Kef* : 2.78% | |

*Table 19. Social Network metadata.*

Tunisian governorate has been collected as the user provenience.[23] If the *tot* voice does not appear at the top of the table, as it is not for the voice, i.e. *Years* in Table 19 (or all the voices in Table 21), this means that the *tot* voice coincides with the *tokens* number reported at the top of the table, as in *16,056 tokens* in Table 19.

Therefore, Table 19 reports information about users whose texts are collected in TArC. The table also shows the percentage of female and male users, the percentage of each age range occurrence, and the percentage of tokens for each governorate.

Rap metadata can be registered in two ways. In fact, rap lyrics are collected and written down by fans, but the text of the lyrics is a product of the singer. In this case, the decision was made to register the information of the user, where it was made available, considering it to be more informative in comparison with information about the singer. This is because we collected rap lyrics taking into account

[23] The Tunisian governorates are considered to be the following: *Ariana, Béja, Sousse, Bizerte, Gabès, Nabeul, Jendouba, Kairouan, Zaghouan, Kebili, El Kef, Mahdia, Manouba, Medenine, Monastir, Gafsa, Sfax, Sidi Bouzid, Siliana, Ben Arous, Tataouine, Tozeur, Tunis, Kasserine*.

| | **Forum** | total tokens: 11,909 | | |
|---|---|---|---|---|
| **Years %** | **Years %** | **Govern. %** | **Age %** | **Gender %** |
| | | tot: 294 | tot: 514 | tot: 627 |
| *2005*: 14.28% | *2013*: 491 | *Bizerte*: 80.27% | *-25*: *n.o* | *M*: 43.06% |
| *2006*: 15.52% | *2014*: 5.27% | *Tunis*: 19.73% | *25-35*: 58.17% | *F*: 56.94% |
| *2007*: 5.38% | *2015*: 2.35% | | *35-50*: 41.83% | |
| *2008*: 2.65% | *2016*: *n.o* | | *50+*:*n.o* | |
| *2009*: 2.25% | *2017*: 0.78% | | | |
| *2010*: 1.75% | *2018*: 6.81% | | | |
| *2011*: *n.o* | *2019*: 18.75% | | | |
| *2012*: 5.42% | *2020*: 15.93% | | | |

*Table 20. Forum metadata.*

| | **Blog** | total tokens: 6,671 | | |
|---|---|---|---|---|
| **Years %** | **Years %** | **Govern. %** | **Age %** | **Gender %** |
| *2005*: *n.o* | *2013*: *n.o* | *Tunis*: 100% | *-25*: *n.o* | *M*: 47.95% |
| *2006*: *n.o* | *2014*: 52.05% | | *25-35*: 100% | *F*: 52.05% |
| *2007*: *n.o* | *2015*: *n.o* | | *35-50*: *n.o* | |
| *2008*: *n.o* | *2016*: *n.o* | | *50+*:*n.o* | |
| *2009*: 15.77% | *2017*: *n.o* | | | |
| *2010*: *n.o* | *2018*: *n.o* | | | |
| *2011*: 19.24% | *2019*: *n.o* | | | |
| *2012*: 12.93% | *2020*: *n.o* | | | |

*Table 21. Blog metadata.*

the possibility to study their orthography in comparison with the same song transcribed in Arabic script, thus making singer metadata mostly irrelevant.[24]

Finally, we have also reported some global information of TArC, such as the token percentages for each year, in Table 23, and the global percentage of tokens which have been registered as age range

[24] Furthermore, by collecting data in both Arabizi and Arabic characters, we could create a complete Language Model for written Tunisian Arabic, following a suggestion by Prof. Houda Bouamor whom we thank.

| | **Rap** | total tokens: 8,691 | | |
|---|---|---|---|---|
| **Years %** | **Years %** | **Govern. %** | **Age %** | **Gender %** |
| | | tot: 3086 | tot 4564 | tot 4564 |
| *2005*: 3.98% | *2013*: *n.o* | *Tunis*: 86.23% | *-25*: 9.31% | *M*: 100% |
| *2006*: *n.o* | *2014*: *n.o* | *Ben Arous*: 13.77% | *25-35*: 90.69% | *F*: *n.o* |
| *2007*: *n.o* | *2015*: 13.87% | | *35-50*: *n.o* | |
| *2008*: *n.o* | *2016*: 15.73% | | *50+*:*n.o* | |
| *2009*: *n.o* | *2017*: 10.08% | | | |
| *2010*: *n.o* | *2018*: 33.43% | | | |
| *2011*: 22.91% | *2019*: *n.o* | | | |
| *2012*: *n.o* | *2020*: *n.o* | | | |

*Table 22. Rap Lyrics metadata.*

| **tot:** *43,327* | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Years** | **%** | **Years** | **%** | **Years** | **%** | **Years** | **%** |
| *2005* | 4.52% | *2010* | 0.91% | *2015* | 5.72% | *2020* | 4.19% |
| *2006* : | 3.82% | *2011* : | 7.2% | *2016* : | 5.91% | | |
| *2007* : | 1.42% | *2012* : | 5.47% | *2017* : | 3.21% | | |
| *2008* : | 0.7% | *2013* : | 4.63% | *2018* : | 14.16% | | |
| *2009* : | 2.9% | *2014* : | 9.07% | *2019* : | 26.2% | | |

*Table 23. Percentage of tokens per year in TArC.*

and the gender information, in Table 24. The last table (25) shows the percentage of tokens for which a Tunisian governorate has been recorded. The number of tokens for which a Tunisian governorate has been registered is 20,595. In addition, in the same table, some of the most frequent foreign origins are also reported, which were recorded with the state name and which are a part of the number considered to be the total (i.e. 21,516 tokens).

As shown in these tables, despite our efforts to construct a corpus which is as representative as possible of the language in multiple of its aspects, it is not possible to obtain balanced percentages, except at the cost of forcing the data collection in a specific direction at the cost of the representativeness of the other language dimensions. Therefore, as stated in the previous chapters, we have opted for the

| **tot:** *22,213* | | **tot:** *25,157* | |
|---|---|---|---|
| **Age ranges** | **%** | **Gender** | **%** |
| *-25* : | 14.23% | *Male* | 62.13% |
| *25-35* : | 74.19% | *Female* : | 37.87% |
| *35-50* : | 9.6% | | |
| *50+* : | 1.97% | | |

*Table 24. Percentage of TArC tokens for which age range and gender has been registered.*

**tot:** *21,516*

| **Govern.** | **%** | **Govern.** | **%** | **Govern.** | **%** | **Govern.** | *%* |
|---|---|---|---|---|---|---|---|
| *Ariana* : | 1.53% | *Sfax* : | 2.64% | *Sidi Bouzid* : | 0.87% | *Siliana* : | 0.20% |
| *Béja* : | 0.2% | *Jendouba* : | 0.72% | *Mahdia* : | 1.53% | *Ben Arous* | 3.01% |
| *Sousse* : | 3.16% | *Kairouan* : | 0.99% | *Manouba* : | 0.38% | *Tataouine* : | 0.29% |
| *Bizerte* : | 2.85% | *Zaghouan* : | 0.39% | *Medenine* : | 2.45% | *Tozeur* : | 0.21% |
| *Gabès* : | 1.02% | *Kebili* : | 0.17% | *Monastir* : | 1.45% | *Tunis* : | 66.78% |
| *Nabeul* : | 2.24% | *El Kef* : | 1.28% | *Gafsa* : | 1.11% | *Kasserine*: | 0.53% |
| **For. St.** | **%** | **For. St.** | **%** | **For. St.** | **%** | **For. St.** | **%** |
| *Germany* : | 0.65% | *Algeria* : | 1.4% | *France* : | 0.86% | *Morocco* : | 0.35% |
| *Italy* : | 0.11% | *Switzerland* : | 0.05% | *Canada* : | 0.01% | *Qatar* : | 0.1% |
| *Libya* : | 0.3% | *Luxembourg* : | 0.06% | *UAE* : | 0.01% | | |

*Table 25. Percentage of tokens per governorate in TArC, excluding foreign ones. 'For. St.' stands for 'Foreign State'.*

least invasive choice, namely, a snapshot of reality, which is that of DNW in Tunisian Arabizi through different textual genres.

As far as the analyses are concerned, the solution imagined here in order to ensure the validity of the results found was to take into account this information regarding the percentages, and at the same time to use appropriate statistical tools. For example, Pearson's chi-squared test (Fisher, 1922; Pearson, 1900) is a statistic test that measures the statistical relationship between two categorical variables. In the simplest of terms, it is the comparison of the frequencies observed in certain categories with the frequencies we might expect to

get in those categories by chance (Field, 2009, 688).[25] It is then the standardisation of the deviation for each observation.

For each analysis carried out, the significance of same with the p-value set at a threshold of 0.05 was evaluated, as is quite common in linguistic analyses.[26]

The following section details the amount of data and metadata collected and presents the TArC-based linguistic analyses. Emphasis was placed on ensuring the reproducibility of each process carried out, both in the construction of the corpus and in the analyses performed on it. This was possible thanks to the building process adopted. Indeed, we made use of every available tool, including the scripts created for each analysis on the reference web sites.[27]

[25] These analyses are always based on absolute frequencies, even if we only report percentages in the tables in the next chapter, for the sake of synthesis. Absolute frequencies can be observed using the scripts created for the analyses and made available on the github page (see footnote 27).

[26] The smaller the p-value, the smaller the margin of error of analysis, and the exact p-value is then reported in the analysis.

[27] The multi-task architecture is available at the following link: `https://gricad-gitlab.univ-grenoble-alpes.fr/dinarelm/tarc-multi-task-system`, while TArC and the analyses scripts are available at the following link: `https://github.com/eligugliotta/tarc`.

# 4. ANALYSES DEVELOPED ON THE TARC

After the TArC data description given in Section 5, together with some information about the data and metadata quantity, this chapter will describe the analyses carried out on the corpus. Before entering into the outline of the analyses, some brief details will be given about the query tools employed in Section 1. The analyses realised are aimed at outlining the nature of the corpus itself, i.e. the description of the users, whose texts have been collected in TArC, and at investigating the linguistic reality of Tunisian Arabizi. Regarding the linguistic reality of Tunisian Arabizi, the analyses will move along three paths:

1. Quasi-orality traits (Section 2);
2. Spontaneous settling trends (Section 3);
3. Continuum of formality degree (Section 4).

Finally, in Section 4, we will discuss the possible conclusions that can be brought about in light of the phenomena highlighted throughout the analyses.

## 1. Query tools employed

The tools used to perform the analyses are simple query systems, made up of a set of Python functions, created specifically to observe the TArC data. In order to make these analyses reproducible, the query scripts created for each analysis have been made available at

the website where it is also possible to download the corpus data, along with instructions for their use.[1] In general, query tools are systems created in order to quickly identify and select the data that we want to observe. For instance, to conduct the initial type of analyses outlined in Section 2, it is essential to examine Prepositional Phrases (PPs) chosen based on their internal composition. The structure of PPs has been identified by iterating on the morpho-syntactic annotations, which TArC has been provided with (see Section 3). The script created for the quasi-orality analysis returns the frequency's percentage of the various orthographic realisations examined, following the instructions provided through the query functions. Moreover, depending on the level of depth of the data analysed, additional information can be requested from the system, such as the number of occurrences of prepositions involved in the various types of orthographic realisations. It is also possible to observe the textual data that makes up the different PPs, which are organised according to two types of structures, as shown in the examples 22 and 22. The representation of the two structures of a TArC PP, of the type [prep+n], used by the system in order to carry out the analyses, is provided below. The first of these (ex. 22) compares the token encoding in Arabizi (*arabizi*) with that in Arabic characters (*word*). The second (ex. 22) functions for the identification of the internal composition of the PP. This includes the alignment of the word in Arabic characters (*word*) and its reduction into the morphemes that compose it (*tokenisation*). Both structures associate the morpho-syntactic annotation produced on TArC data (*POS*) with the textual elements.

(22) Structure sample 1
('arabizi', 'word', 'POS') = [('bou9alb', 'بوقلب', '[PREP+NOUN]ADJ'), (...)]

(23) Structure sample 2
('tokenization', 'word', 'POS') = [('بو+قلب' , 'بوقلب', '[PREP+NOUN]ADJ'), (...)]

The sentences involved can also be written in a textual file or observed in the Python shell. The following example (22) shows an extract of the textual file written in order to observe the data

[1] Reference website: `https://github.com/eligugliotta`.

considered within the code-switching analyses outlined in Section 3.[2] The sentences reported in the example correspond to the first code-mixed NP sentence and the fifth NP sentence, which is also the first part for making a code-mixed PP. The output file reports the information about the selected code-mixed NPs, such as *El farine* in the first sentence, and the PPs, such as in the case of *lel famille* in the second sentence. The information given includes the corresponding POS tags and indexes in TArC, together with the whole sentence in Arabizi, the textual genre to which it belongs (*social*), and the users' information (*origin*, *age* and *gender*), if collected.

(24) 1. <Mixed NP: [El farine]> <POS: DET_foreign>
<TArC idx: 830-831>
<Within sentence: t7out El farine w mele7 w tsoub 3lihom 50 g zebda dhayba w t5alet>[3]
<Genre: social>
<Users' metadata:> <origin:/> <age:35-50> <gender:F>
5. PP(1). <Mixed PP: [lel famille]> <POS: PREP+DET_foreign>
<TArC idx: 1800-1801>
<Within sentence: ena houni mochkelti win todkol mra jdida lel famille nwali nakraha nakraha griba w blech sbab !!>[4]
<Genre: social>
<Users' metadata:> <origin:/> <age:/> <gender:F>

## 2. Quasi-orality traits

As seen in Chapter 1, Arabizi belongs to Computer Mediated Communication (CMC) and in particular to Digital Networked Writing (DNW) contexts. Indeed, it is a writing system which was born for digital communication. Its hybrid peculiarity as a quasi-oral system has already been discussed, including the characteristics of encoding an under-resourced language, i.e. Tunisian, which is mainly oral, but which uses the Arabic script when written down. As a result, Arabizi seems to be partially influenced by the writing practice of the Arabic script, and partially intended to better represent some

[2] The structure of the output files is a similar structure to that of XML files (*eXtensible Markup Language*) for data ordering reasons.

[3] 'Put the flour and salt in and pour on them 50 gr. of melted butter and mix'. My own translation.

[4] 'And here's my problem, when a new woman comes into the family, I start hating her, hating her strangely and for no reason!!'. My own translation.

oral peculiarities of Tunisian Neo-Arabic, which are not properly encoded by Arabic graphemes. In fact, in Section 1 of Chapter 1, it has already been stated that no system, whether based on Arabic or Latin characters, was created to encode Tunisian phonology. However, the Arabizi system seems to mirror Tunisian Neo-Arabic phonetics, while the Arabic-script-based system seems to be more oriented to reflecting the *morpho-phonological* structure of Arabic. Therefore, Subsection 2 observes that one aspect of Tunisian, which is more effectively represented using Arabizi compared to the Arabic script-based system, is the treatment of short vowels, including metathesis and diphthongs, as well as certain consonant realizations such as the loan phonemes /p/ and /v/, or the /q ~ g/ allophones (for detailed analyses on these points, refer to Section 2).[5] On the other hand, among the Arabizi writing features influenced by the Arabic script, hyper-correction phenomena such as the epenthetic vowel use at the beginning of #CC clusters need to be contemplated. As already observed, this phenomenon is typical of the Arabic-script orthography, where it is codified through the ʔalif grapheme, and same aims to transform the Tunisian #CC cluster into #VCC, which is easier to realise. This phenomenon has already been encountered in Chapter 2, and this feature is also present in Arabizi (as in the example given below, ex. (15), for example, in the Arabizi word *etkoun*, /tkūn/), in which an epenthetic vowel can be observed at the beginning of the same cluster. However, instances of this are found much less often in Arabizi in comparison with its occurrence in Tunisian encoded in Arabic script.[6]

(15) *t'es dans une ces regions donc yelzmek etkoun men mdina men ces regions*

*alors que t'es mehdwi*,

/t'es dans une ces regions donc *yəlzmək tkūn mən mdīna mən*ces regions alors que t'es *məhdwi/,*

'You are in one of these regions so you should come from a city of these regions,
that should make you a Mehdia citizen'

[5] Gibson (2002) states that /q/ and /g/ have phonemic status, considering the loan phoneme /g/; however, this is not taken into consideration in this analyses, which instead only focuses on the [g] as an allophone of /q/.

[6] It would be interesting, for future studies, to compare the two different types of encoding in terms of this phenomenon. The rap lyric corpus in TArC has been collected with this comparative aim.

Regarding Tunisian encoding in Arabic-script systems mirroring Arabic morphology, examples of this will be given in sentences (11) and (14).

However, considering Arabizi as a not standardised system, it is possible to suppose that it may allows its users to be more independent from strong writing traditions, such as those concerning the Arabic script. This is the cornerstone of our first hypothesis about Arabizi. What needs to be investigated precisely is whether this freedom of encoding is constrained by the fact that Arabizi belongs to a written context, and normally, before the invasion of Facebook, writing in Tunisian corresponded to writing in Arabic characters and according to some standardised rules of 'correct encoding' (see hypercorrection). The alternative instead is that Arabizi, by creating more freedom with which to render the characteristics of Tunisian, which do not coincide with those of Arabic, allows the user to get rid of some rules normally imposed by Arabic orthography and everything encompassed by the idea of 'correct encoding' in Arabic. This possibility includes Arabizi being influenced by French orthography, being encoded in Latin characters, and the ability to facilitates the use of French vocabulary in code-mixing. Moreover, in a completely different context, but which still involves the Arabic language, i.e. that of bilingual contexts and texts written in a system that is not the one usually intended for that language, it has been verified that the mismatch between the norms of the two languages can generate coding uncertainties that may lead to the spread of linguistic interference.[7]

However, in the case of Arabizi, different phenomena are involved. It is true that the Latin alphabet can be traced back to European languages, but in the operation of encoding Tunisian into Arabizi, there is no translation, but rather a transliteration, of the definite article or the inclusion of an article as a code-mixing element together with the name it defines, if this is an element belonging to another language system (this issue is discussed in Section 3). A similar case to Tunisian Arabizi is represented by the Maltese orthography, which is the only standardised Neo-Arabic variety, and is strongly related to Tunisian, but is encoded using the Latin alphabet. Maltese definite

[7] See, for example, the study of Metcalfe (2014) regarding Norman Sicilian Arabic and the practice of writing Arabic using another orthography, such as the Greek alphabet. See in particular the treatment of Arabic articles. Another example of this is Cypriot Arabic, and its definite article employments (Mion, 2017a,b).

NPs are represented by a geminated consonant in the case of nouns starting with coronal consonants, as in the NP /*is-soppa*/ of the following example (22) (Azzopardi-Alexander and Borg, 2013, 69).

(25) Il-kuċċarun li ħawwad is-soppa bih kollu tal-fidda,
det.N det.N

'The-ladle he stirred the soup with is made entirely of silver'.

We can consider Arabizi to be a similar case, in which a Neo-Arabic variety is encoded through a Latin script which is used for another language that is strongly present within the culture. In order to observe if Tunisian Arabizi orthography (Section 3) has a spontaneous settling tendency, we should consider the role of French in this process (Section 3), in addition to the user profile and the text register (sections 2 and 4).

The same hypothesis can be researched by observing the Arabizi orthographic realisation of the Prepositional Phrase (PP), which diverges from the Arabic script, and which may be considered to be similar to that of French. Indeed, in Standard Arabic, depending on the preposition, a PP can be represented as a unique orthographic word. If the PP involves an articled noun, the orthographic word includes the three following morphemes: preposition (prep), determiner (det), and noun (N). In Tunisian Arabizi, the orthographic realisation of this syntactic structure is different, most often being encoded through the linking of the det to the prep and the prep-det compound separation from the N through a white space, as in the examples (11) and (14), which are both seen in Chapter 1, which will be reported below. In both examples, it is possible to observe how the prep-det compounds (*bil* and *fil*) are separated from the Ns (respectively *dlel* and *jografia*), while the phonetic realisation corresponds to a unique compound, respectively: /b-əd-dlāl/ and /fi-ž-žuɣṛāfya/.

(11) *wa9t yji nghar9ou bil dlel*,

/waqt yži nɣarq-u b-əd- dlāl/,
Prep.det N

'When he comes I drown him with vices'.

(14) *entouma yelzemkom derss fil jografia*,

*/ntūma* yəlzəmkum dərs *fi-ž-* *žuɣṛāfya/*,
Prep.det N

'You (pl.) need a geography lesson'.

Regarding Example (11), the Arabizi prep-det compound *bil* is separated from the N *dlel* by a white space, while in the Arabic script, both prep and det should be written attached to the noun, i.e. بالدلال, /b-əd-dlāl/. The situation is similar to that in Example (14), where the prep-det compound *fil* is separated from the N *jografia*, while in the Arabic script, the PP should correspond to في الجغرافيا, (prep det+N), or at least to فيلجغرافيا, (prep+det+N), which are both pronounced /fi-ž-žuɣṛāfya/. The first Arabic-script realisation (prep det+N) reflects the Arabic morpho-syntactic system, while the second (prep+det+N) can mirror the pronunciation. The reason for *prep+det N* realisation, instead, could be the result of language contact, facilitated through the use of the Latin alphabet (Section 3).

The same situation is represented by Maltese PP, in which we can find the same compounds (Prep+det) as in Tunisian Arabizi, but joined to the noun through a hyphen (-), resulting in *Prep+det-N* as in the following example (22) (Azzopardi-Alexander and Borg, 2013, 70):

(26) Tlaqt fil-għodu mas-sebħ,
Prep.det-N Prep.det-N

'I left in the morning at dawn'.

As can be observed from the example, the second *prep+det* compound (*ma+s* in *mas-sebħ*) presents a determiner (**al-*), the /l/ of which is graphically represented as *s*, being assimilated with the noun's initial consonant (*sebħ*), which is a coronal. The similarity between Maltese and Tunisian Arabizi orthographies of the prep and det morphemes as a compound separated from the N allows us to suppose that the reason for this goes back to the multilingual environment facilitated by the Latin script. This clearly exists in the Maltese orthography, which still has clear traces of its Tunisian *stratum*, despite the Italian *adstratum* being immediately noticeable in sentences such as the following (example 22 (Caruana, 2009, 359)).

(27) Il-punt kollu tal-politika reġjonali hija biex
the-point.m all of.the-politics regional is-cop.f to
ittejjeb il-koeżjoni soċjali u ekonomika
improve.ip.f the-cohesion social and economic
bejn ir-reġjuni Ewropej.
between the-regions European.

'The most important point of regional policy is to improve social and economic cohesion between European regions'.

Furthermore, in Example (22), it is very clear that the nouns and adjectives used in this sentence are of Italian origin,[8] while the invariable terms (or what Myers-Scotton would probably call 'system morphemes'), such as prepositions and articles, are Tunisian. The next section shall proceed with an observation of the orthographic reality of the prepositional phrase in Arabizi Tunisian. Our objective will be to verify whether the encoding in Latin characters facilitates an encoding that reflects the oral reality of Tunisian, or whether Arabizi still remains anchored to the writing tradition in Arabic characters, being above all a written form of encoding. The role of the influence of contact languages, particularly French, will also be discussed.

### *Prepositional Phrase Schemes*

According to traditional Arabic grammar, prepositions are of two types, i.e. simple or lexically derived (Mion and D'Anna, 2021, 207). 'Simple' prepositions in Modern Standard Arabic can be proclitic, namely, bound to the noun which yields the PP, or independent:

1. بـ /*bi*-/, لـ /*l(a/i)*-/ and ك /*ka*-/ that link to the noun, or to the article that follows it;
2. في /fī/, عن /ʕan/, على /ʕalā/, من /*min*/, إلى /ʔilā/, منذ /munðu/ and حتّى /ħattā/ that do not connect.[9]

The PP in Tunisian Arabizi could be ideally realised through the following schemes, where the plus symbol (+) represents the orthographic concatenation:

1. [prep+det+n]
2. [prep+det n]
3. [prep det+n]
4. [prep det n]

Considering these possible combinations, our hypothesis is that, in the case of a Prepositional Phrase (PP) that includes the article

8 Italian translation given by Caruana (2009): 'Il punto di maggior rilievo della politica regionale è quello di migliorare la coesione sociale e economica tra le regioni europee.'

9 Except when followed by suffixed personal pronouns, as in فيّ lit. 'into me', منّي lit. 'from me' and عليك lit. 'on me'.

| | Patterns | Occ. | % |
|---|---|---|---|
| 1. | [prep+det n] | 320 | 58.3% |
| 2. | [prep+det+n] | 128 | 23.3% |
| 3. | [prep det+n] | 56 | 10.2% |
| 4. | [prep det n] | 45 | 8.2% |
| | **Tot** | 549 | 100% |

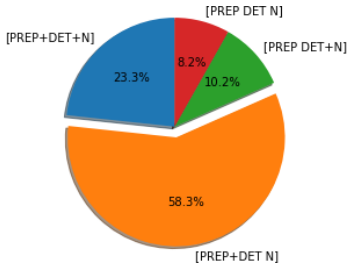



*Table 26. Distribution of Prepositional Phrase in case of articulated nouns.*

(det), the combinations are arranged along a continuum from the most representative of Tunisian orality ([prep+det n]) to the most representative of the orthographic system of Standard Arabic ([prep det+n]).

When analysing the TArC data in order to investigate this hypothesis, it becomes evident that the majority of prepositional syntagmas in Arabizi are structured as [prep+det n], whereas the [prep det+n] pattern is less common, as illustrated in Table 26.

By observing Table 26, it is possible to notice an orthographic structure of the PP that seems to be placed in an intermediate position with respect to the extremes of the continuum; this is the structure in which all the elements are attached into a single orthographic word ([prep+det+n]). It is our belief that this solution can be considered as being representative of Arabic orthography (in the case of proclitic prepositions of the first type), as well as representative of the oral realisation in Tunisian. On the other hand, as far as the fourth structure is concerned ([prep det n]), we do not believe this to be representative of either of the two realities, and given the scarcity of representation in percentage terms (8.2%), we consider same to be not very indicative for the purposes of our analysis. In order to verify whether the nature of the Tunisian preposition, matching Arabic proclitic ones, plays a role in the orthographic realisation of the syntagma in Arabizi, which is considered useful for observing the internal composition of these prepositional syntagmas. Taking into account what has been said about prepositions in MSA, two prepositions in particular, i.e. a proclitic one (بـ), and an independent one (في), which are among the most frequently used, were chosen as the focus for this analysis. In Table 27, the frequency of the two chosen prepositions and the percentage of their occurrences in the

| | Patterns | بـ | | في | | Other | Preps. |
|---|---|---|---|---|---|---|---|
| | | Occ. | % | Occ. | % | Occ. | % |
| 1. | [prep+det n] | 78 | 52.7% | 111 | 69% | 131 | 54.6% |
| 2. | [prep+det+n] | 66 | 44.6% | 21 | 13% | 41 | 17.1% |
| 3. | [prep det+n] | *n.o.* | *n.o.* | 15 | 9.3% | 41 | 17.1% |
| 4. | [prep det n] | 4 | 2.7% | 14 | 8.7% | 27 | 11.2% |
| | **Tot** | 148 | 100% | 161 | 100% | 240 | 100% |

*Table 27. Distribution of the two chosen prepositions following different schemes. N.o. stands for 'no occurrences'*

different types of orthographic realisations of prepositional phrases in TArC are reported.

Upon examining Table 27, it is evident that, in the case of prepositional phrases where the preposition does not attach to the NP, namely, the third and the fourth entries ([prep det+n] and [prep det n]), the preposition بـ never appears ([prep det+n]) or appears very rarely (2.7% in [prep det n]). This may seem obvious when reflecting on the fact that this preposition is always proclitic in MSA, but due to the fact that, in Tunisian Neo-Arabic, prepositions that are normally independent in MSA can behave as though they are proclitic, the opposite should also not be excluded. In addition, in Section 3, we will observe how this orthographic rule becomes less obvious in the case of code-mixed NPs.

Having previously assessed [prep det n] as being insignificant, in terms of both frequency and in terms of the representativeness of the Tunisian oral or Arabic orthographic system, the percentages of the most frequently occurring prepositions should be included herein, in order of the highest occurrence in the [prep det+n] type of syntagm.

1. /fi/ في: 36.6%;
2. /mtāʕ/ متاع: 24.4%;
3. /ʕala/ على: 21.9%;
4. /maʕ/ مع: 17.1%.

The type of prepositions that occur in the orthographic realisation type [prep det+n] seems to further confirm the fact that this realisation type tends to reproduce the orthographic system in Arabic characters. In fact, excluding the preposition متاع, which is typical of Neo-Arabic Tunisian but absent in MSA, the other prepositions in traditional Arabic orthography are encoded separately from the noun which they head within the prepositional phrase.[10]

In general, the previous observations regarding the distribution of prepositions in syntagmas in which the preposition is independent of the determiner confirm, on the one hand, that a certain amount of attention is also being paid to the orthographic rules of MSA in the written encoding of Tunisian in Arabizi, and on the other hand, also confirm the fact that these types of orthographic realisations of the prepositional syntagma, in particular ([prep det+n]), are representative of a certain tendency to respect the Arabic traditional orthography. It should be noted, however, that this is a less frequent structure, compared with the first two in Table 26 ([prep+det n] and [prep+det+n]), where the distribution of the two prepositions appears to be even more interesting.

In the case of the first structure ([prep+det n]), in which the preposition is linked to the determiner to form a sort of articulated preposition, and which is assumed to be more representative of the Tunisian oral system, the situation is almost reversed. Indeed, a significant proportion of instances of the proclitic preposition بـ (52.7%) are observed in this context. Although this percentage is consistently lower than that of the preposition في (69%), it remains notably frequent and typically independent. This factor can thus provide possible additional evidence to support our hypothesis. In addition, the preposition بـ in the traditional Arabic orthography should rather impose a realisation of a [prep+det+n] type, while the preposition في should impose the [prep det+n] structure. Therefore, when compared to the less frequent orthographic realisations ([prep det+n] and ([prep det n]), this type of realisation ([prep+det n]) seems to be less constrained by the orthographic traditions represented by Arabic characters.

[10] The prepositions على and مع can both be phonetically and orthographically joined with a suffix pronoun in MSA (Mion and D'Anna, 2021, 213).

Finally, it is necessary to examine the structure [prep+det+n], which has been hypothesized to be located at an intermediate point on the continuum between greater representativeness of Tunisian orality and Arabic orthography. In this case, since the percentage of occurrences of the clitic preposition (بـ, 44.6%) greatly exceeds the number of occurrences of the independent preposition (في, 13%), it can be supposed that, in a binary system of analysis that contrasts the Tunisian orality with the MSA orthography, this type of orthographic realisation ([prep+det+n]) is positioned at an intermediate point on the continuum.

However, we must also take into account the fact that the structure [prep+det+n] is more representative of the orthographic system of Standard Arabic than of Tunisian orality. Indeed, the structure type [prep+det+n] exactly represents the traditional structure that would be employed in prepositional phrases of written MSA involving the proclitic preposition بـ. As for the pattern [prep+det+n], this represents a middle ground, as a possible solution in both languages (MSA and Tunisian), but only when the preposition is proclitic in the case of MSA. In contrast, in the case of Tunisian, annexation is possible for a larger number of prepositions.[11]

However, this last observation forces us to add two additional considerations into the mix:

1. The first is that surely the very nature of some prepositions, such as بـ, imposes a specific realisation of the prepositional syntagm at the orthographic level, regardless of whether it is in Arabizi or in Tunisian encoded in Arabic characters.
2. The second is that considering the Tunisian linguistic reality to be a binary system is reductive in light of the points detailed in Chapter 1. It is necessary to consider the idea that the tendency to graphically join the preposition to the determiner, separating them from the noun, may be a consequence of the phenomenon of contact with languages that present articulated prepositions, such as Italian and French.

With regard to the first consideration, the question of which prepositions occur most frequently in the various types of orthographic

[11] These prepositions are mainly: ب ل في من مع م ع.

| | Patterns | Three most frequent prepositions % | Other preps. |
|---|---|---|---|
| **1.** | [prep+n] | بـ (57.6%) , لـ (17.7%), في (15.1%) | 9.6% |
| **2.** | [prep n] | في (42.2%), على (13.6%), من (13%) | 31.2% |
| **3.** | [prep+det n] | في (34.7%), بـ (24.4%), لـ (15.6%) | 25.3% |
| **4.** | [prep+det+n] | بـ (51.6%), في (16.4%), ع (13.3%) | 18.7% |
| **5.** | [prep det+n] | في (26.8%), متاع (17.8%), على (16.1%), | 39.3% |
| **6.** | [prep det n] | في (31.1%), على (22.2%), مع (11.1%) | 35.5% |

*Table 28. The three most frequent prepositions through different schemes.*

structures will now be investigated further; see the results of this investigation in the table below (Table 28).

As can be seen from Table 28, the most commonly recurring prepositions in the different orthographic schemes are:

1. /*mtāʕ*/ متاع, /*ʕala*/ على and /*min*/ من for the schemes [prep det n], [prep det+n] and [prep n];
2. /*b-*/ بـ, /*l-*/ لـ and /ʕa/ ع are instead the most frequent for [prep+det+n], [prep+det n] and [prep+n];
3. /*fi*/ في is highly frequent in all compounds, so we can say that it has a *super partes* status.

Except for the last of these, as expected, it appears that these prepositions have some influence on imposing an orthographic realisation for the PP. In patterns which tend to keep the preposition separate from what follows it, the prepositions are tonic, while conversely, in the other patterns, the prepositions are mainly proclitic. On the one hand, the role of the proclitic prepositions in restricting the possibilities of PP realisation in patterns 1, 3 and 4 seems evident. Nevertheless, in the case of a defined NP, the type of proclitic preposition does not really play a role in the choice of orthographic realisation among the possibilities represented by patterns 3 and 4. The same consideration applies to tonic prepositions employed in a PP involving a defined NP, i.e. in the choice between 5 and 6.

Regarding the second consideration, in addition to the phenomena of linguistic contact with European languages (an issue which will be discussed in Section 3), there is one further question which requires consideration. This is the fact that the orthographic realisation of [prep+det n] may be a scriptural fallout of a morpho-

| Patterns | Occ. | % |
|---|---|---|
| [prep+n] | 198 | 17% |
| [prep n] | 969 | 83% |
| **Tot** | 1,167 | 100% |

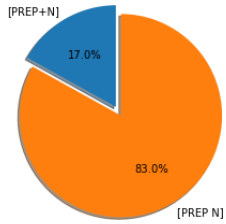



*Table 29. Distribution of Prepositional Phrase in case of not-articulated nouns.*

syntactic phenomenon, or rather of the nature of the determinative morpheme. This morpheme appears in some Semitic languages dating back to the first millennium BC as a phonetic simplification of a deictic theme (Garbini and Durand, 1994, 102). According to Pennacchietti (1968, 2005) and Loprieno (1995), the appearance of the determinative morpheme would have led to a reorganisation of the linguistic material (with relative and genitive functions) already present with consequent specialisation of the three functions (Eksell Harning, 1980; Versteegh, 1984, 164, 98-99).[12]

The article in MSA appears completely standardised in *al-* as a definiteness mark (def) and a genitive mark in the case of specification complements (known as *Construct State*). On the contrary, in Tunisian, the determinative morpheme may overlap with *əlli* in the case of a definite reduced relative clause (Jouini, 2012). Indeed, *əlli* may be a complementiser or a relativiser, and in the latter case tends to merge with the preposition that precedes it, whether proclitic or independent, to form a single relative particle (such as *m-əlli* 'from which').[13] Returning to the main topic of this thesis, it appears obvious that the role of the determiner seems to play a certain role in the orthographic realization of the Prepositional Phrase (PP).

In fact, looking at PPs lacking the determiner morpheme ([prep+n] and [prep n], in Table 29), the separation of the preposition from the noun, not their merging into a single compound, appears to be their main tendency.

It seems plausible, therefore, that it is the det mark that plays a key role in these types of orthographic formations. Consequently, there is a reason to believe that the higher frequency of the [prep+det n]

[12] The determinative morphemes are for Hebrew /*ha-*/, North-Arabic /*h(n)-*/ and Arabic /*ʔal-*/ (Garbini and Durand, 1994, 102).

[13] The close relationship between these elements requires much more in-depth investigations that we have chosen to reserve for further dedicated study (see Gugliotta et al. (forthcoming)).

| *Arithmogr.* | **Occurrences** | *Arithmogr.* | **Occurrences** |
|---|---|---|---|
| *2* | 69 | *6* | 1 |
| ***3*** | **2,108** | ***7*** | **1,463** |
| *4* | 13 | *8* | 72 |
| ***5*** | **598** | ***9*** | **1,125** |

*Table 30. Occurrences of arithmographemes in TArC.*

syntagm is due to the presence of det exerting an enclitic force on the preposition preceding it. This is exactly the same as in the case of *əlli*, and is probably caused by phonological reasons due to the shortness of their initial vowels, thus being related to the oral dimension of the language. What remains to be clarified is why Tunisian Arabizi instead prefers to leave the noun separated from the compound *prep+det*. In order to deepen the Arabizi encoding of the PP, a frequency analysis was developed in diachrony on TArC's data, allowing us to enter the second path of spontaneous settling tendencies.

## 3. Spontaneous Settling Trends

As outlined in Chapter 1, the Arabizi system is a recent invention, being born of the technological spread in Arabic countries at the end of the 1990s. However, since its inception until today, each Arabizi national variety should have been subject to a process of spontaneous settling, as it is a non-standardised system. This is also true for Tunisian Arabizi. Indeed, one type of evident spontaneous settling phenomena we can consider, as an example, is the use of *arithmographemes*. Regarding these, Bianchi (2012) explains there are numerals which are employed as letters for 'hard-to-transliterate' sounds. Since the inception of Arabizi encoding, all the numerals, except for the zero and the one, were employed to represent Tunisian phonemes.

Nowadays, the numerals that are still used are mostly *3*, *5*, *7*, and *9*, respectively for ع, خ, ح, and ق, as shown in Table 30.

On the other hand, the use of *2*, *4*, *6* and *8*, for ء, غ, ط and ه, is less consistent, as shown in Table 33. However, when observing the distribution of these arithmographemes over time, it is possible to affirm that *2*, *4*, *6* and *8* were not widely used since the beginning of Tunisian-Arabizi diffusion. It seems as though the most used arith-

| *Years* | **Number of tokens** | *Years* | **Number of tokens** |
|---|---|---|---|
| *2005* | 1,809 | *2013* | 2,229 |
| *2006* | 1,839 | *2014* | 4,363 |
| *2007* | 681 | *2015* | 2,751 |
| *2008* | 335 | *2016* | 1,982 |
| *2009* | 1,395 | *2017* | 1,290 |
| *2010* | 437 | *2018* | 4,430 |
| *2011* | 1,354 | *2019* | **12,680** |
| *2012* | 2,630 | *2020* | 2,017 |

*Table 31. Number of tokens per year within the first six blocks of TArC.*

mographemes in Arabizi (*3*, *5*, *7* and *9*) have also tended to follow the same distribution pattern as when this use first emerged.

In addition, in order to clearly understand the mentioned table (Table 33), it should be explained that the percentages are based on the exact number of tokens considered for the analyses (meaning 42,150 words, corresponding to the first six blocks of the seven, which constitute the TArC data. In fact, at the time of that analysis was carried out, six blocks of the were already annotated with all the linguistic information). Table 31 reports the information about the number of tokens contained in the first six blocks, divided by year.

Since each year is represented by a different amount of tokens, the following strategy was adopted to perform balanced analyses. First, the years were grouped into pairs of years contingent on time. In any case, we are interested in the evolution of this phenomenon over time, and certain writing habits are unlikely to change radically from one year to the next. The distribution of arithmographs employed by users in that two-year period (i.e. *2005-2006*) was then examined, the global number of arithmographs written in that time frame as a total (i.e. *arithmographs written in 2005 + arithmographs written in 2006 = 357 total arithmographs for the two-year period of 2005-06*). The total amount of arithmographs for each pair of years is as shown in Table 32.

Therefore, in Table 33, the relative frequency for each two-year subcorpus is shown so as to observe whether, at the quantitative level, these graphemes undergo changes over time.

| *Couple of Y.* | **Tot of arithm.** | *Couple of Y.* | **Tot of arithm.** |
|---|---|---|---|
| *2005-06* | 357 | *2007-08* | 158 |
| *2009-10* | 307 | *2011-12* | 754 |
| *2013-14* | 953 | *2015-16* | 871 |
| *2017-18* | 651 | *2019-20* | 1,398 |

*TABLE 32. NUMBER OF ARITHMOGRAPHS PER PAIR OF YEARS WITHIN THE FIRST SIX BLOCKS OF TARC.*

| *Num.* | **2005-06** | **2007-08** | **2009-10** | **2011-12** | **2013-14** | **2015-16** | **2017-18** | **2019-20** |
|---|---|---|---|---|---|---|---|---|
| *2* | 0.84% | 3.16% | *n.o.* | 1.33% | 0.84% | 1.38% | 1.38% | 1.14% |
| *3* | 31.09% | 33.54% | 25.41% | 16.45% | 12.17% | 40.53% | 37.48% | 41.56% |
| *4* | 0.28% | *n.o* | *n.o* | *n.o* | 0.1% | 0.46% | 0.31% | 0.36% |
| *5* | 13.17% | 12.66% | 14.98% | 2.65% | 3.15% | 4.36% | 0.92% | 13.88% |
| *6* | *n.o* | *n.o* | *n.o* | *n.o* | 0.1% | *n.o* | *n.o* | *n.o* |
| *7* | 31.65% | 29.75% | 17.92% | 10.21% | 9.86% | 27.67% | 32.1% | 24.03% |
| *8* | 0.56% | *n.o* | 6.51% | *n.o* | 0.21% | 3.67% | 0.15% | 0.86% |
| *9* | 22.41% | 20.89% | 16.94% | 13.13% | 18.57% | 21.93% | 5.53% | 18.17% |
| *Tot* | 100% | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

*TABLE 33. DISTRIBUTION OF THE USE OF ARITHMOGRAPHEMES THROUGH TIME IN TARC. N.O. STANDS FOR 'NO OCCURRENCES'.*

The percentages show trends indicating widespread fluctuations throughout the period studied. However, when observing the first column (*2005-2006*) in comparison with the last (*2019-20*), the values are quite similar. At the same time, however, after examining the central columns (*2011-12*) and (*2013-14*), in comparison with the first and the last, these show values that are diffusely lower, with the exception of the value reported for the arithmograph '2', the value of which is slightly higher than in the first column. In any case, although interesting to observe, the diachronic distribution of arithmographemes cannot provide an exhaustive answer to the question we posed at the beginning, that is, whether Tunisian Arabizi tends more towards the representation of orality or the respect of a graphic norm, the most influential exponent of which is represented by the encoding of Standard Arabic in Arabic characters. Therefore, in Section 3, we will reexamine issues raised in the previous sections, in reference to the phenomenon of the orthographic realisation of the Prepositional Phrase (PP), which seems to offer some evidence in this regard. Therefore, Section 3 will include observations on the code-switching phenomena in Arabizi, in order to understand if the

PP scheme [prep+det n] could have been influenced by contact with European languages. Finally, in Section 2, some diachronic and diastratic analyses will be presented, which were developed in order to understand if the data collected in TArC shows traces of the process described as koineisation by Miller (2004) and previously outlined in sections 1 and 2.

### *Prepositional Phrase distribution*

Regarding the linguistic elements of Arabizi that have been put under the magnifying glass of this quantitative analysis on TArC, it is now necessary to observe their diachronic distribution in Table 36. The purpose of this is to see whether prepositional phrase realisation in Tunisian Arabizi has varied over time, until we reach the distribution observed in the previous section (Table 26), or whether this distribution also reflects that of the early stages of Arabizi diffusion. The following table presents the percentage of each scheme occurrence in a specific year, divided into the total number of tokens of the same year (shown in Table 31). In the above-mentioned table, prepositional schemes are organised into the following types:

1. Type0 : [prep+n]
2. Type1 : [prep n]
3. Type2 : [prep+det+n]
4. Type3 : [prep+det n]
5. Type4 : [prep det+n]
6. Type5 : [prep det n]

As with all the analyses presented in Section 3, the analyses on TArC data have also been developed using the same query system, as outlined in Section 1. [14]

As shown in the previous section (Table 31), the data for each year is not balanced, and therefore, in order to make the presented analysis as balanced as possible, the years have been grouped together. In this case, as in the example provided regarding arithmographemes (Table 32), the aim is to observe the distribution of prepositional phrases over time, such that the most important factor is for the temporal groups to be contingent. The grouping performed is as shown in the following table (34).

[14] The query tools built for the spontaneous settling analyses are available at the same reference website: `https://github.com/eligugliotta`.

| *Group of Y.* | **Tot of tokens** | *Group of Y.* | **Tot of tokens** |
|---|---|---|---|
| *2005-10* | 6,496 | *2011-14* | 10,576 |
| *2015-17* | 6,023 | *2018-19* | 10,770 |
| *2019-20* | 8,357 | | |

*TABLE 34. GROUPS OF YEARS AND RESPECTIVE AMOUNT OF TOKENS.*

| *Group of Y.* | **PP tot** | *Group of Y.* | **PP tot** |
|---|---|---|---|
| *2005-10* | 156 | *2011-14* | 436 |
| *2015-17* | 267 | *2018-19* | 257 |
| *2019-20* | 215 | | |

*TABLE 35. TOTAL NUMBER OF PP OCCURRENCES PER GROUP OF YEARS.*

| **PP schemes** | **2005-2010** | **2011-14** | **2015-17** | **2018-19** | **2019-20** |
|---|---|---|---|---|---|
| **Type0** | 10.26% | 5.05% | 21.35% | 16.34% | 15.35% |
| **Type1** | 45.51% | 64.45% | 50.94% | 55.25% | 51.16% |
| Type2 | 12.18% | 4.59% | 8.99% | 9.34% | 11.16% |
| **Type3** | 22.44% | 22.02% | 16.48% | 14.79% | 17.67% |
| **Type4** | 4.49% | 2.52% | 2.25% | 3.5% | 3.26% |
| Type5 | 5.13% | 1.38% | *n.o.* | 0.39% | 0.93% |
| **TOT** | 100% | 100% | 100% | 100% | 100% |

*TABLE 36. DISTRIBUTION OF PP USAGE THROUGH TIME IN TARC. N.O. STANDS FOR 'NO OCCURRENCES'*

As can be seen in the above table (34), since the data for 2019 is much higher in quantity than the others, it was necessary to split it in half into two different groups. The following analyses will present the distribution of the different prepositional phrase patterns over time, such that the table below (35) respectively reports the total occurrences of prepositional phrases for the year groups identified in Table 34. These occurrences represent the total of the computations to be presented.

Pearson's chi-squared test (Fisher, 1922; Pearson, 1900) between the years and types of PPs, shows (with a p-value < 0.01) that the types are dependent from the years. In particular, by observing the distribution of Type3 and Type4 over the years, it is possible to notice that the percentages of occurrence of structures of Type3 are always greater than those of Type4. A similar observation can be seen

| *Total Nominal Phrases* | **1,336** | **Part of a PP** | **[det n]** |
|---|---|---|---|
| *NP where N starts with coronal lett.* | **480** | | |
| *NP where N does not start with coronal lett.* | **856** | | |
| *NP where Det is not present* | **21** | | |
| *NP where Det is assimilated* | **269** | **70** | **10 (3.7%)** |
| *NP where Det is not assimilated* | **190** | **65** | **169 (88.9%)** |

*Table 37. Distribution of PP usage through time in TArC.*

with regards to Type0 and Type1, where the second one (Type1) always accounts for a greater percentage in comparison to the first one (Type0). In both Type1 and Type3, the preposition is separated from the noun, and the only difference between them is that in the case of Type3, the determiner is present (and is merged with the preposition, rather than the noun). These could be categorized as two types, albeit belonging to the same category, with a variation at the morphological level depending on the presence or absence of the determiner.

Before the scope of the analysis is narrowed further in order to detect what, for native speakers and users, is more or less spontaneous/formal in the language (an issue we will address in the next section (4), there is an additional matter to address. This is the encoding of the assimilated article in the case of defined nouns beginning with a consonant of the coronal type. This matter pertains to both 'good' text encoding and encoding that leans towards representing either oral or written conventions in Arabic characters. Nevertheless, the preceding paragraph has shown that the most common pattern of prepositional phrases (PP) aligns with what was previously identified as the most indicative of Tunisian orality. However, it still remains to be understood why there is a tendency in Tunisian Arabizi to graphically separate the articulated preposition' from the defined noun with a space. We hypothesize that the phenomenon of determiner assimilation may be linked to this tendency.

Regarding article assimilation in the presence of nouns beginning with coronals, when assessing the data summarised in Table 37, it seems that whether NPs are part of a PP or not is irrelevant.

In fact, NPs that show assimilation and those that do not show assimilation have frequencies of participation in PPs which are not too dissimilar (70 occurrences, namely 26%, for NPs with assimilation, and 65 occurrences, or 34%, for NPs without assimilation). There-

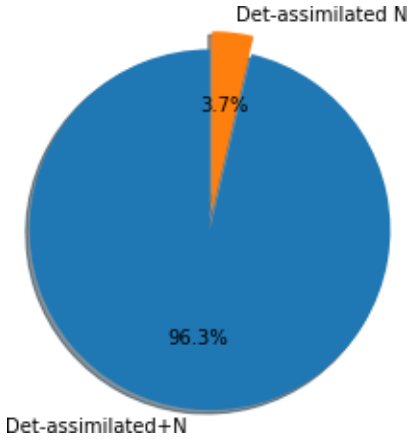



*FIGURE 1 NP WITH ASSIMILATED DET. THE N'S FIRST CONSONANT IS A CORONAL.*

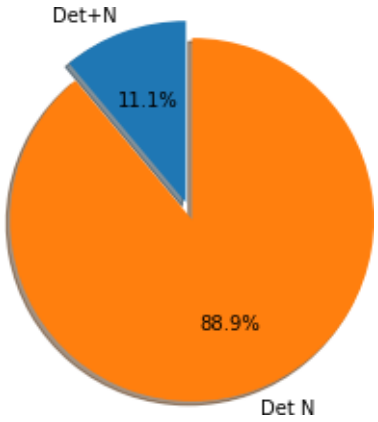



*FIGURE 2 NP WITH NOT-ASSIMILATED DET. THE N'S FIRST CONSONANT IS A CORONAL.*

fore, being part of a PP does not seem to have an influence on the realisation of assimilation. In contrast, when only electing NPs where nouns begin with a coronal (480 occurrences), it can be seen that the amount of noun-separated determiners is significantly higher in NPs that ignore assimilation on the orthographic level (almost 89%), compared to those that exhibit det's assimilation (3.7%). Pearson' chi-squared test, executed between NPs which include coronal phonemes at the beginning of an N (regardless of whether assimilation is rendered or not) and present in a space before the Ns, gives a p-value (< 0.01) which illustrates the relationship between the data. In the case of ignored-assimilation, the user therefore seems hardly inclined to orthographically respect the morpho-phonology of Tunisian, and in doing so, adopts a graphic coding strategy of NP and PP that includes it, namely, that of separating det from N with a space. However, from the number of occurrences of this, the tendency to respect the assimilation also at the orthographic level (269 occurrences) seems stronger than the tendency to ignore it, even if it is consistent (190 occurrences).

Finally, in order to discover whether this was a type of diachronic variation, i.e., if the two modes of defined NP writing code, with and without assimilation, represent a trend for spontaneous settling over time, a diachronic analysis and Pearson's chi-squared test were performed for NPs that appear with an N beginning with a coronal over time. Results of the diachronic analysis are as shown in Table 39,

| *NPs* | **2005-2008** | **2009-11** | **2012-13** | **2014-16** | **2017-19** | **2019-20** |
|---|---|---|---|---|---|---|
| *Tot* | 66 | 89 | 70 | 93 | 85 | 57 |

*Table 38. Distribution of assimilated det through time in TArC.*

| *NPs* | **2005-2008** | **2009-11** | **2012-13** | **2014-16** | **2017-19** | **2019-20** |
|---|---|---|---|---|---|---|
| *GR* | 81.82% | 5.62% | 50% | 80.65% | 82.35% | 52.63% |
| *not GR* | 18.18% | 94.38% | 50% | 19.35% | 17.365% | 47.37% |

*Table 39. Distribution of assimilated det through time in TArC. GR stands for 'Graphically Rendered'.*

while Pearson's chi-squared test of graphical assimilation and years gave a significant p-value.

Furthermore, the data was organised into years group with the aim of producing analyses which were as balanced as possible, and, in the following table (38), we report the amount of Nominal Phrases for each group of years.

When examining the data in the table above, it appears that user habits are not fairly stable in terms of converging towards a graphical rendering of article assimilation, with some periods of trend reversal. In fact, an entire period (*2009-2011*) can be observed, in which the tendency to prefer the graphical rendering of assimilation is reversed, with an abrupt decrease in *2009*, in correspondence with a peak in the values of the graphical non-representation of the article assimilation (from 2009 to 2011 inclusive). Immediately after the inverted-preference period, there is a notable period of balance, (*2012-2013*) and then a re-inversion in *2014*.

The fact that this data can be correlated with sociopolitical events linked to the revolution is a thought that had obviously crossed our minds, given the identity value that this system of writing assumes, especially among the new generations, during this delicate phase of Tunisia's history. However, in order to test this hypothesis, native speakers should be surveyed, in order to evaluate the way in which they perceive this variable; we have decided to consider this as a possible basis for future work.[15] At the same time, it is also plausible that another social event influenced the practices of informal writing, and this is also connected to the above-mentioned sociopolitical facts (see Chapter 1). The arrival of Facebook and its role in the contiguous mass diffusion of the Arabizi writing practice

[15] The authors of the blogs collected in TArC have already expressed their support in this regard.

| *Assimilation* | **-25** | **25-35** | **35-50** | **50+** |
|---|---|---|---|---|
| *GR* | 77.8% | 36.2% | **80%** | 100% |
| *not GR* | 22.2% | 63.8% | 20% | *n.o.* |

*TABLE 40. DISTRIBUTION OF DET'S ASSIMILATION THROUGH THE AGE RANGES. GR STANDS FOR 'GRAPHICALLY RENDERED'.*

| *Assimilation* | **M** | **F** |
|---|---|---|
| *GR* | 24.7% | **77.4%** |
| *not GR* | **75.3%** | 22.6% |

*TABLE 41. DISTRIBUTION OF DET'S ASSIMILATION THROUGH GENDER METADATA. GR STANDS FOR 'GRAPHICALLY RENDERED'.*

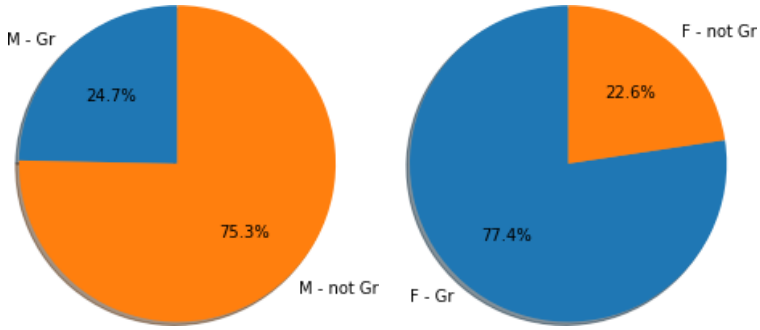



*FIGURE 3 ASSIMILATION THROUGH GENDER. GR STANDS FOR 'GRAPHICALLY RENDERED'.*

in those years was alluded to earlier. The fact that, in the last period alone, there seems to be an approximation of the percentages of these two modalities of graphical rendering for assimilation could be interpreted instead as a sign that the users consider it acceptable to have a system of written communication which is divergent from the 'scriptural norms' previously acquired. This could indicate that writing in Arabizi has become such a widespread practice that users no longer feel the need to eliminate any ambiguity from their texts and can ignore, for example, the 'correct' graphic rendering of the assimilated article.

By adding to the data in this table, as well as the data in tables 40 and 41 types of users can be attributed to the two methods of performing graphical encoding for assimilation.

In fact, the assimilation phenomenon seems to be diastratically influenced by both the users' age and their gender. Table 40 shows

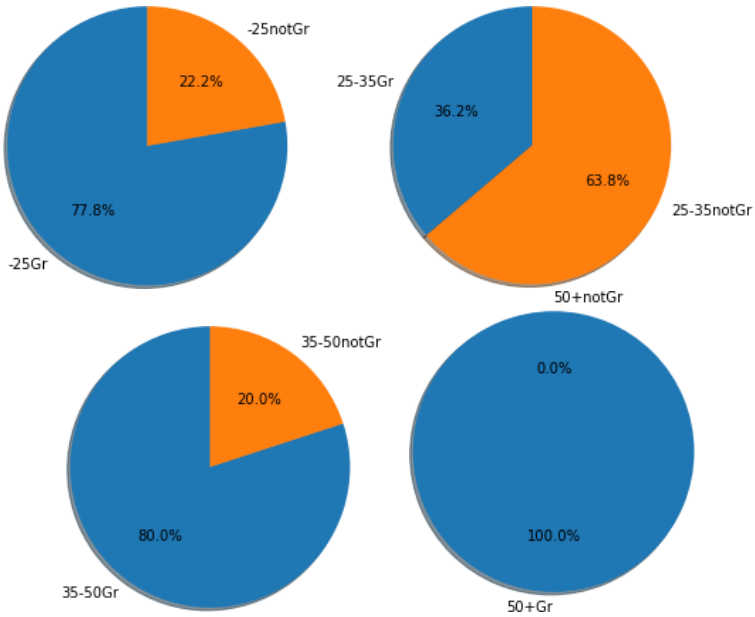



*FIGURE 4 ASSIMILATION THROUGH AGE RANGE. GR STANDS FOR 'GRAPHICALLY RENDERED'.*

a preference for orthographic rendering of assimilation in the age range *35-50*, and Table 41 indicates that the same preference is mainly attributable to female users. Pearson's chi-squared test confirms that the variable is not independent from the age or the gender (with a p-value < 0.01).

### *Code-switching distribution*

This work has documented how the variant of Tunisian Neo-Arabic, which has emerged as the national variety thanks to the mass media, coincides with that of the capital. We have also seen how this variety represents, at the diastratic level, a high social stratum with access to a French cultural and linguistic background, being rich in hybridisms or mechanism towards a French loan adaptation. As mentioned above in Chapter 1, Cerruti and Regis (2005) consider that it is better to keep hybridisms separate from code-switching (CS) phenomena, involving only the superficial linguistic system (i.e. words, morphemes and phonemes) and not the discourse, al-

though they are non-institutionalised manifestations of contact in use as much as CS (Cerruti and Regis, 2005, 193-194).[16] Considering the preliminary character of the analyses we are going to present, we thought it appropriate to take into account different possibilities of interpretation of the phenomena resulting from linguistic contact detected in Tunisian Arabic. In order to maintain a simple expositive line, we have chosen to adopt Myers-Scotton's terminology, given the diffusion that his studies represent. The definition of code-switching (CS) adopted here is that of Myers-Scotton (1993), who states that the term is used to refer to alternations of linguistic varieties within the same sentence (intrasentential).

> '*CS is the selection by bilinguals/multilinguals of forms from two or more linguistic varieties in the same conversation.* [...] *Stretches of CS material may be inter-sentential (switches from one language to the other between sentences) or intrasentential (within the same sentence, from the single morpheme level to higher levels).*'

Considering the two types of *asymmetry* involved in code-switching, i.e. the structural type and the content type, Myers-Scotton (2006) explains that, in a bilingual context, one language supplies the main grammatical frame for a clause containing words from both languages.

Myers-Scotton (1993) also defines this structural-driver-language as the *Matrix Language* (ML), while the content-driven-language is referred to as the *Embedded Language* (EL), as already mentioned in Section 2. The ML not only governs the form of a word selected from the EL lexicon, but also governs the structural relationships between words within the sentence. A code-switching model for bilingual speech is the Matrix Language Model (MLF), while a model specialising in the morpheme types is the 4-M model (Myers-Scotton, 2006). MLF makes a distinction between *content morphemes* and *system morphemes*, in which the former are those which assign or receive thematic roles,[17] while the latter are basically functional words, even if the two categories do not overlap completely. The 4-M model refines the MLF model by dividing system morphemes into three types: the *early system morphemes*, the *bridge late system morphemes*, and the *outsider late system morphemes*. In Section 2,

[16] We would like to thank the reviewers of this work, Simone Ciccolone and Luca D'Anna, for their helpful comments and corrections regarding these linguistic elements.

[17] Also called θ-roles, which are mainly carriers of semantic cores.

this topic was introduced through Examples (16), (17) and (18), which are included below. In fact, the three examples concern the Nominal Phrase in Tunisian Arabizi and present different encoding realisations of the determiner followed by a foreign noun. The Example (16) presents an *intrasentential code-switched defined NP*, where the determiner (*el*) corresponds to the Tunisian morpheme, while the N (*famille*) belongs to the French lexicon. Instead, in Example (17), the whole NP (*une zone*) is made of French elements.[18]

(16) *hatta el famille walit j evite tout le monde*,

/ħatta *l-* famille *wallīt* j'evite tout le monde/,
det$^{T}$ N$^{F}$

'Also the family I started to avoid everybody'.

(17) *S7i7, ama essa7el fi tounes hiya tasmya mta3 une zone, kima el cap bon…*,

/ṣħīħ, *āma əs-sāħəl fi tūnəs həya təsmya mtāʕ* une zone,
det$^{F}$ N$^{F}$
*kīma əl-*cap bon…/,

'That's right, but the Sahel in Tunisia, is a nomenclature of a zone, as the Cap Bon'.

These occurrences are particularly interesting when viewed in terms of the questions raised regarding determiner's role in the orthographic realisation of the Tunisian Arabizi NP and PP. Indeed, the occurrences of the Tunisian determiner followed by a foreign element were analysed in TArC. The aim was to observe the orthographic behaviour of the determiner, in case of a foreign N, and in particular when same is headed by a preposition. In fact, in the Examples (18), two PPs can be observed. The first of these (*f les chansons*) is headed by a Tunisian preposition (/*fi*/), which is graphically separated by the determiner (*les*) and the noun (*chansons*), which are French elements. The second of these also presents a Tunisian preposition (/*b-*/), separated from the French indefinite noun (*mention*) which is followed by:

[18] The word Sahel (sāħəl, 'coast'), is used in Tunisia specifically to refer to the eastern region of the coast and its main cities.

(18) *Nrapi manich conscient f les chansons Njib zéro vues ama naj7 b mention,*

| | | | |
|---|---|---|---|
| */nṛāpi ma-nī-š* conscient | *fi* | les | chansons |
| | prep$^{T}$ | det$^{F}$ | N$^{F}$ |
| *nžīb* zéro vues *āma nāža*ħ | *bi* | mention*/,* | |
| | prep$^{T}$ | N$^{F}$ | |

'I do rap without being aware of the songs I get zero views but I obtained a mention'.

The first consideration for example (18) is about the prepositions being employed in the PPs. As observed in Section 2, the /*fi*/ preposition tends to occur in each type of PP scheme, without any restrictions.[19] This is not the same as for the preposition occurring in the second PP (/*b-*/), for which a tendency to merge with the noun has been witnessed.[20] It can also be seen, in Table 27, that the preposition /*b-*/ rarely appears in schemes in which the preposition is separated by the det of the N, such as [prep det+n] and [prep det n]. The results shown in tables 28 and 27 have raised the issue of the preposition playing a role in selecting the orthography realisation of the PP scheme. The second consideration is that which was observed in Section 2, and which seems not to be respected in the case of a *code-switched* PP. Indeed, in example (18), the preposition /*b-*/ occurs separated from what follows it, which, in this specific case, is a foreign noun. It would be useful to deepen this consideration through quantitative analysis. In order to make this clear, in the analyses of the previous paragraphs, defined NPs containing nouns classified as *foreign* were purposely excluded. Instead, for these analyses, all the Tunisian det marks followed by a *foreign* noun in TArC (123 sentences) were selected, and 76 of these formed part of a PP. In 73 occurrences (namely 59.3% of the 123 sentences and 96% of the 76 PPs), the preposition appears to be graphically joined with the determiner ([prep+det n]). There were only 3 occurrences (namely 2.4% of all the sentences and 4.1% of the 76 PPs) in which the preposition appears to be separated from the determiner ([prep det n] or [prep det+n]). As reported in the file generated by the query script, these three sentences are:

[19] Table 28 highlighted the fact that /*fi*/preposition occurs with a high percentage in the following schemes: [prep+det n] (35%), but also in [prep det+n] (30%) and [prep det n] (33.3%).

[20] Table 28 highlighted the fact that /*b-*/ tends to occur only within the PP schemes where it is joined to the noun (namely: [prep+n] (59%), [prep+det n] (25.4%) and [prep+det+n] (52%).

(28) 50.PREP(33) <Mixed NP: [mte3 el GEOGRAPHIE]> <POS: PREP_DET_foreign>
<TArC idx: 26419-26420>
<Within sentence: chouf ya mte3 el GEOGRAPHIE el msetria wel mhedwia aktharhom mkach5in et je crois elli ennes el kol ta3ref hetha wé n7eb nkollek 7aja rahou wled el 3asma 9afzin 3lina barcha felli t7eb inti ma3ada ettourisme >
<Genre: forum>
<Users' metadata:> <origin:/> <age:/> <gender:/>[21]

(29) 93.PREP(56) <Mixed NP: [m3a l commandant]> <POS: PREP_DET_foreign>
<TArC idx: 36682-36683>
<Within sentence: fi salla mta3 l'équipage m3a l commandant bidou , expérience wehed may3ichHech martine >
<Genre: blog>
<Users' metadata:> <origin: Tunis> <age: 25-35> <gender: F>[22]

(30) 99.PREP(61) <Mixed NP: [ala l récup']> <POS: PREP_DET_foreign>
<TArC idx: 37358-37359>
<Within sentence: 3jebni fel galerie hethi li heyya feha barcha des œuvres basées essentiellement ala l récup' wel recyclage; Mathalan, blalet ma3moulin b mgharfa mta3 glaces, plastique ethika elli na7na nlaw7ouha, tableau yfatte9 ma3moul ala tarf hdid m5azza, prtrait afro w fazet, ala beb lou7 mdagdeg!>
<Genre: blog>
<Users' metadata:> <origin: Tunis> <age: 25-35> <gender: F>[23]

---

[21] My own translation: 'Look, (you who are) expert in geography, the people of Monastir and the people of Mahdia, most of them, are Taraji fans and I think everyone knows this thing. And I want to tell you something, right them, the children of the capital (they) are much smarter than us in everything you want, except tourism'.

[22] My own translation: 'In the crew room with the commander himself, unrepeatable experience'.

[23] My own translation: 'He brought me to a gallery where there are many works based mainly on recovery and recycling; for example, earrings made from ice cream spoons, the plastic ones that we throw away, a cool painting made on a piece of iron (m5azza), an Afro portrait and (other) stuff, on a damaged wooden door'.

It is evident that, in the case of a code-switched PP, the most highly preferred solution is the following PP scheme: [prep+det n]. This finding is in line with the hypothesis regarding the general preference in Tunisian Arabizi being for the same scheme, which, as observed in previous paragraphs, could be influenced by the fact that Arabizi uses the Latin script. Likewise, the Latin script is associated with the French language, and its usage could lead users to also extend this PP scheme to the Tunisian not code-switched PPs observed in the previous sections. Unfortunately, we are unable to compare our data with statistics on a Tunisian corpus in the Arabic script, in order to check if this scheme is used more in Arabizi in comparison to Tunisian encoded in the Arabic script.[24] However, in order to support our observations, there are two other types of analyses which can be performed. The first is to check the typology of users' involved in code-switching practices. The second one is to observe whether there is any variation in the code-switching quantity and in the PP schemes among different genres of text. This second question will be further traversed in Section 4, in order to verify whether the selection of an orthographic scheme is a matter regarding the degree of formality. Instead, we will outline the results of our diastratic question in the following table (42). However, since this also includes a diastratic analysis of code-switching, together with a diaphasic and diatopic one, it is impossible to proceed without first mentioning the studies that form the basis of the sociolinguistic approach to code-switching. The most prominent of these is the essay by Blom and Gumperz (1972), in which they address the issue of CS as a sophisticated strategy that speakers deploy to signal aspects of their ethnic and social identity. Among the various CS distinctions isolated by scholars, and destined to have great success, the 'we code' is particularly interesting. This represents the code used by speakers of a minority language for communication within their own ethnic community, and is opposed to the 'they code', which is used by the same speakers for communication with outsiders. However, Auer (1999) notes that the alternation of codes does not necessarily refer to a conversational-external opposition such as 'we-code/they-code', but that it is part of an emerging mixed code that exploits the dynamic relationship between the two source-languages (Auer, 1999, 119-120). When scrutinising the results of our analysis of prepositional phrases involving nominal elements that are the result of CS (Table 42), it appears that those making the

24 We may perform this analysis in the future.

| Govern. | % | Age | % | Gender | % | Text Genre | % |
|---|---|---|---|---|---|---|---|
| *Nabeul* | 2.63% | *-25* | 6.52% | *M* | 16.33% | *Blog* | 26.5% |
| *Monastir* | 5.26% | ***25-35*** | **86.96%** | ***F*** | **83.67%** | ***Forum*** | **51.3%** |
| *Tataouine* | 2.63% | *35-50* | 6.52% | | | *Social N.* | 22.1% |
| ***Tunis*** | **89.47%** | *50+* | *n.o.* | | | | |

*Table 42. Code-switched PPs in TArC. N.o. stands for 'no occurrences'.*

most use of intrasentential code-switching are female forums users aged between 25 and 35 years, mostly coming from Tunis. Below is a table summarising four independent trends, which are plausibly correlated, but are not to be interpreted in this way.[25]. For this specific data observation we considered only the sentences for which we had the gender information, namely 49 of the mentioned 123 sentences presenting code-switching.

However, it must be remembered that, as shown in tables 19, 20 and 21, data from Tunis and in the 25-35 age group are better represented in the corpus than the other governorates and age groups. Tunis represents 45.1% of the collected governorates for social networks, 19.73% of those for forums and 100% for blogs. The age range 25-35 represents 52.21% of the ages ranges collected for social networks users, 58.17% of those of on forums and 100% of the bloggers' ages. At the same time, however, the data on the gender of users remained quite balanced for all textual genres (again excluding rap, as a musical genre above all). Regarding the latter, social networks (with 16,056 tokens) were better represented than forums (11,909 tokens) and blogs (6,671 tokens), but despite this, in Table 42, the textual genre of forums stands out.

Regarding the percentages reported in Table 42 concerning gender, Pearson's chi-squared test was also performed (with a p-value < 0.01), which highlighted the relationship between gender and code-switching. In Table 43, the frequencies utilized for the calculations in a contingency table are presented.

Furthermore, regarding the textual genres, Pearson's chi-squared test show a dependence within data and the code-switching, with the p-value reaching 0.02. In Table 44, the frequencies of the data used in Pearson's chi-squared test are reported.

[25] Regarding Table 42, percentages are given based on the total number of sentences selected by our query tool, which was 123. In this case, the script is available on the reference website: `https://github.com/eligugliotta`

| | *M* | *F* | *Tot* |
|---|---|---|---|
| **Switched** | 8 | 41 | 49 |
| **Not Switched** | 130 | 107 | 237 |
| **Tot** | 138 | 148 | |

*TABLE 43. CONTINGENCY TABLE OF USERS' GENDER AND CS DATA.*

| | *Social N.* | *Forum* | *Blog* | *Tot* |
|---|---|---|---|---|
| **Switched** | 25 | 58 | 30 | 113 |
| **Not Switched** | 173 | 129 | 106 | 549 |
| **Tot** | 198 | 187 | 136 | |

*TABLE 44. CONTINGENCY TABLE OF TEXT GENRE AND CS DATA.*

| | *PPs without CS* | *PPs with CS* |
|---|---|---|
| **F** | 45.15% | 83.67 % |
| **M** | 54.85% | 16.33 % |
| **Tot** | 100% | 100% |

*TABLE 45. COMPARISON OF PERCENTAGES OF PPS WITH OR WITHOUT CS ACCORDING TO USERS' GENDER.*

However, as shown in Table 45, by also testing PPs that do not include any code-switching (237 sentences), the following percentages were obtained: 45.15% of the PPs were produced by female users and 54.85% were produced by male users.

These latter percentages definitely seem to confirm that in the case of prepositional phrases that do not include code-switching elements, there is no variation in the sexual gender of users, unlike in the case of prepositional phrases that include code-switching elements. This finding takes on particular interest when juxtaposed with the conclusions of the previous section, in which the tendencies of the feminine gender for the orthographic rendering of assimilation are documented. Women seems to be more 'precise' in encoding, making use of the morpho-phonological phenomena of Tunisian in Arabizi, while at the same time tending to use a lot of French terminology in their daily exchanges, in which case they use a French-style encoding of PP ([prep+det n]). In fact, we would like to stress that the forums that make up TArC are generic forums and do not represent a particular sector-specific language. The fact that Tunisian women, and in particular those from Tunis, have a preference for the French language is certainly not new (Daoud, 2011, 15).

Regarding the metaphorical-identity value that the French language and the Tunisian dialect may assume for this type of forum users, two hypotheses can be advanced. The first of these is based on the aforementioned study by Blom and Gumperz (1972), in which the two scholars identified a dual identity as the engine behind code-switching in a small Norwegian village. In this particular case, it was a urban identity in contrast to the identity of the inhabitants of a small village. However, the authors themselves quickly realised the impossibility of establishing a simple and direct biunivocal relationship between the two, predefined between the use of a language and the ascription to a particular identity (Pasquandrea, 2007, 42). As addressed in Section 1, it was unclear whether the status of the French language remained anchored to an elitist Bourguibian heritage or instead coincided with an everyday variation between urban and non-urban speeches. However, what our study has in common with the Norwegian case is that all participants in the interactions (in our case, young Tunisian women) seem to respect certain linguistic conventions imposed by a micro-social order. In contrast, the second hypothesis is that the two linguistic varieties used by a bilingual person go hand in hand and can be seen as mutually significant (Woolard, 1998) in creating a 'new' space for the bilinguals to occupy and utilise for self-positioning (Finnis, 2013; Georgakopoulou and Finnis, 2009; Nafa, 2015). In this specific situation, this idea of a new space may precisely exist as the space on women's forums, as anticipated in Chapter 1.[26] We will address this question more fully in the analyses in the following paragraph (Section 4).

### *Koineisation*

In Section 1, we introduced the topic of koineisation as a process that leads to the adoption of a national urban *koiné*, which tends to prevail in public spaces, while the vernacular which is typical of the place of origin is relegated to family communication (Miller, 2004). Koineisation is often a consequence of migrations and movements within the country. When referring to *koiné*, we imply contact between linguistic *subsystems*, although this contact does not always lead to the establishment of a *koiné*. As previously discussed

[26] Riegert and Ramsay (2013), regarding code-switching in Tunisian, showed that the mechanisms of language choice, code-switching, and linguistic variation that occur on digital platforms are the same as those that occur in offline interactions.

in Section 1, there are certain sociolinguistic conditions required for a *koiné* to emerge from this type of contact (Siegel, 1985). Section 2 details how communication technology is supporting the spread of this national *koiné*, whether through television, radio or online channels, the diffusion of which is supported by the widespread use of smartphones. We then delved into the writing systems used for informal online communication in Section 2, concluding with an overview, with no claim to exhaustiveness, of the Tunisian Arabizi system, in Section 2.

However, there were still some questions which remained unanswered; these are the questions we asked in Section 2, namely:

1. What *kind* of Tunisian idiom should we expect to find on social networks?
2. Will users employ their own local variant, or will they level out overly local features for the sake of inter-comprehension?
3. Will users choose an urban *koiné* specifically?

We will attempt to answer these questions through the TArC analyses outlined in this section. In order to do so, it is necessary to take, as discriminating macro-elements, those traditionally used for the classification between urban and Bedouin dialects, such as the double realisation of /q/, in [q] or [g], the evolution of diphthongs and some morpho-syntactic elements (such as the realisation of weak verbs, or variations within subject pronouns).

### *Qāf realization*

With regard to the distinctive phonological features presented in Table 1, the first point concerns the phonetic realisation of the uvular plosive phoneme /q/ in the voiceless [q] and/or the voiced [g]. This double pronunciation of /q/ is one of the main criteria for the classification of Tunisian idioms. As Skik (2003) reminds us:

> '*Cette double prononciation constitue le principal critère de distinction entre les parlers tunisiens, et ce, pour les usages eux-mêmes, qui font la différence entre « ceux qui parlent bil-qâla et ceux qui parlent bil-gâla », aussi bien que pour Ibn Khaldoun qui distinguait les « parlers des citadins » (à q) et « ceux des bédouins » (à g) et pour William Marçais qui avait fait de cette double prononciation un des principaux discrimina entre « parler citadins » (et villageois), d'une part et « parlers bédouins », d'autre part.*' (Skik, 2003, 637).[27]

[27] 'This double pronunciation constitutes the main criterion of distinction between the Tunisian languages, and this, for the uses themselves,

For further discussion of the typological category of village speech, reference should be made to Mion (2015), which highlights their mixed nature resulting from long interdialectal contact. As far as our study is concerned, what we want to verify is whether the diffusion of the urban dialect of Tunis has had an impact on the realisation of /q/ with consequent fallout at a graphemic level in the writing of Arabizi by users outside of Tunis. In order to carry out this data analysis, it is necessary to briefly describe how these two realisations of the same phoneme are distributed across the Tunisian territory. In the general classification of Tunisian idioms, village languages are found in the following areas: the Northwestern area (Bizerte countryside), the Sahel area and the Sfax countryside area (Mion, 2015, 270). According to the traditional classification, from the point of view of territorial distribution, the realisation of /q/ in [q] is a minority compared to that in [g], considering that all the rural Tunisia use the [g] phone, as well as some districts in cities on the coast, and Gabès (Saada, 1984, 28). The *bil-qala* governorates are mainly Bizerte, Tunis, Sousse, Monastir, Mahdia, Sfax, and Kairouan. However, it should also be noted that within these governorates, there are many places where the realisation in [g] also appears (Mion, 2015, 271).

Considering that the intention of this research is to identify evidence of the koineisation process, regarding the phono-graphic phenomenon in question, let us take the traditional realisation as described in (Skik, 2003, 637-641) as an example, in which Skik states that he favours informants aged around fifty and older, in order to investigate the situation prior to the changes which occurred in the country, especially following independence in 1956. Skik, while stressing the non-existence of clear boundaries and of governorates knowing a single pronunciation, identifies as *bil-qala* governorates: Bizerte, Ariana, Tunis, Zaghouan, Nabeul, Kairouan, Sousse, Monastir, Mahdia, and Sfax. The rest of the governorates are presented as *bil-gala*. Additionally, Skik (2003) states that the realisation in [q] is far less than that in [g], and that there are variations on the

---

which make the difference between « those who speak *bil-qâla* and those who speak *bil-gâla* », as well as for Ibn Khaldun who distinguished the « parlers des citadins » (with *q*) and « those of the Bedouins » (with *g*) and for William Marçais who had made of this double pronunciation one of the principal *discrimina* between « parlers citadins » (and village varieties), on the one hand and « parlers bédouins », on the other hand.' My own translation.

level of sexual gender, between the two realisations, the study of which should be investigated further. In spite of the imprecise nature of an analysis carried out by observing governorates instead of villages or towns, the result of our analysis, as shown in the table below (Table 46), reveals an important fact, i.e., that today's situation, from a quantitative point of view, of realisations in *q* or *g*, seems to be reversed, compared to that presented by Skik (2003). In fact, at the top of the column we can see that out of the total data taken into consideration (excluding Tunis, as the model of the *koiné* whose diffusion we want to observe), the percentage of realisations in *q* or *9* (81.03%) is significantly higher than the realisation in *g* (18.96%). Moreover, from the same table, it can also be observed that the largest contribution of realisations in *q* actually comes from the governorates classified by Skik (2003) as the urban type, as shown in the upper part of the table. However, in the governorates classified by Skik (2003) as predominantly *bil-gala* areas, the realisation in *q* is not negligible. On the contrary, with regard to the 18.96% of realisations in *g*, the majority of examples are from the governorate of Gabès (representative of the Bedouin Sulaymite typology, and well-known *bil-gala*-speech exponent (Saada, 1984, 28)), but 36.37% of that 18.96% is made up of realisations from the predominantly *bil-qala* type governorates. Concerning the situation in the capital, this seems to have instead reconfirmed the data which indicates that most citizens prefer the realisation in *q* or *9*. These results seem to confirm today's situation, compared to the pre-independence one, in which an ongoing process of koineization is being based on the model of the capital. In fact, Pearson's chi-squared test between urban (excluding Tunis) and non-urban varieties and the realisation of the phoneme /q/, according to our analysis shows a data relation, as well as for Tunis data (p-value).

The following analysis was carried out by taking up the suggestion of Skik (2003) regarding the need to investigate the gender issue relating to the realisation of /q/. Table 50 is considered to show all the instances of /q/ which are written as *q* or *9*, in all the governorates (except for Tunis, which is depicted in 51).

Regarding the distribution of the users who produced these graphemic realisations, their gender seems to be quite balanced, with a split of 43% of males and the 57% of females. Instead, the *g* grapheme seems to be mostly preferred by male users (90% of the *bil-gala* productions), while only 2.7% were produced by female users. The *bil-qala* realisation is produced equally by men and women, but only 26% of the male users wrote down the /q/ as *g* (correspond-

| Governorates | *bil-qala 81.03%* | *bil-gala 18.96%* |
|---|---|---|
| *Bizerte* | 19.79% | *n.o.* |
| *Ariana* | 7.29% | 4.35% |
| *Zaghouan* | 2.08% | *n.o.* |
| *Nabeul* | 5.21% | *n.o.* |
| *Kairouan* | 4.17% | 8.7% |
| *Sousse* | 19.79% | 8.7% |
| *Monastir* | 6.25% | 4.35% |
| *Mahdia* | 3.12% | *n.o.* |
| *Sfax* | 5.21% | 8.7% |
| *Béja* | 1.04% | 4.35% |
| *Gabès* | 3.12% | 39.13% |
| *Jendouba* | 5.21% | 4.35% |
| *Kebili* | 2.08% | *n.o.* |
| *El Kef* | *n.o.* | 4.35% |
| *Manouba* | 4.17% | *n.o.* |
| *Medenine* | 2.08% | 4.35% |
| *Gafsa* | 2.08% | 8.7% |
| *Sidi Bouzid* | 1.04% | *n.o.* |
| *Siliana* | *n.o.* | *n.o.* |
| *Ben Arous* | 2.08% | *n.o.* |
| *Tataouine* | 2.08% | *n.o.* |
| *Tozeur* | 2.08% | *n.o.* |
| *Kasserine* | *n.o.* | *n.o.* |
| *Tot* | 100% | 100% |

*Table 46. Diatopic analysis of /q/ realisation in Tunisia. N.o. stands for 'no occurrences'.*

ing to 90% of the *bil-gala* production in Table 50). In order to prove our analyses, Pearson's chi-squared test was performed, which gave a p-value of less than 0.01. Below is the contingency table for data excluding Tunis (48) and including Tunis (49). In fact, we have also added data coming from Tunis, to all the other data, and performed Pearson's chi-squared test between gender and /q/ realisation. The test again gave a positive p-value.

| Governorate | *bil-qala* | *bil-gala* |
|---|---|---|
| *Tunis* | 94.68% | 5.32% |

*TABLE 47. DIATOPIC ANALYSIS OF /Q/ REALISATION IN TUNIS.*

| Govs. except Tunis | *bil-qala* | *bil-gala* |
|---|---|---|
| **M** | 43.3% | 90.3% |
| **F** | 56.7% | 9.7% |
| **Tot** | 100% | 100% |

*TABLE 50. DIASTRATIC ANALYSIS OF /Q/ REALISATION IN TUNISIA.*

| Gov. of Tunis | *bil-qala* | *bil-gala* | Tot |
|---|---|---|---|
| **M** | 93.8% | 6.2% | 100% |
| **F** | 97.3% | 2.7% | 100% |

*TABLE 51. DIASTRATIC ANALYSIS OF /Q/ REALISATION IN TUNIS.*

| | *Male* | *Female* | *Tot* |
|---|---|---|---|
| **[q]** | 87 | 114 | 201 |
| **[g]** | 28 | 3 | 31 |
| **Tot** | 115 | 117 | |

*TABLE 48. CONTINGENCY TABLE OF /Q/ REALISATION IN TUNISIA (EXCEPT TUNIS).*

| | *Male* | *Female* | *Tot* |
|---|---|---|---|
| **[q]** | 283 | 147 | 430 |
| **[g]** | 104 | 118 | 222 |
| **Tot** | 387 | 265 | |

*TABLE 49. CONTINGENCY TABLE OF /Q/ REALISATION ALL OVER TUNISIA.*

In Table 51, in light of the fact that the majority of users prefer the *q* or *9* encoding (94.68% of Table 47), it can be observed that, among all of the male users from Tunis, most of them (approximately 94%) prefer the *q* or *9* graphemes. The situation appears similar for female users (approximately 97%).

These results are contradictory to the assertions of Skik (2003), despite it being well known that *bil-gala* speaking is a practice typically associated with men (and coming from the countryside). It is possible that this phonological phenomenon is more alive in rooted linguistic ideologies than in contemporary reality, but at the same time, the two may feed off of each other. As an example, it is worth mentioning the linguistic practices related to the world of rap, a predominantly masculine world, where, according to the linguistic ideology of native speakers, the realisation of *bil-gala* should be

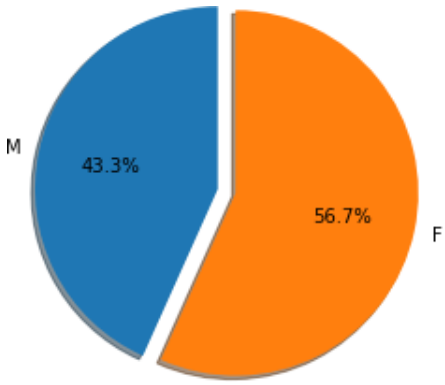



*Figure 5 Percentage of male and female users realising /q/ as [q] outside Tunis.*

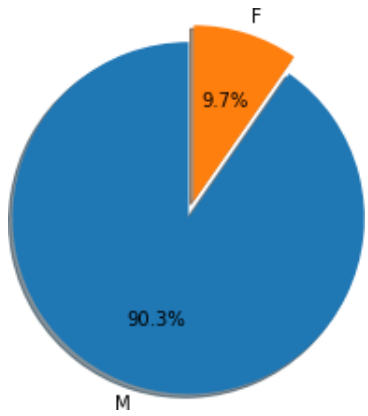



*Figure 6 Percentage of male and female users realising /q/ as [g] outside Tunis.*

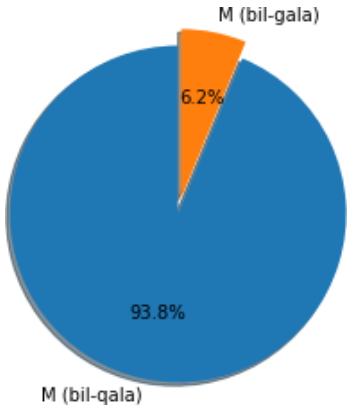



*Figure 7 Percentage of male users realising /q/ as [q] and [g] in Tunis.*

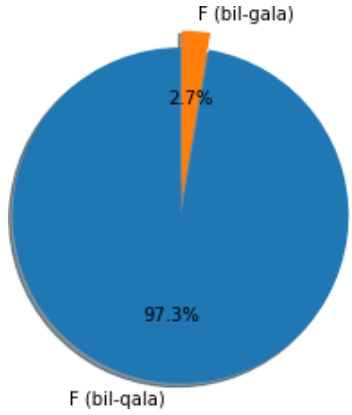



*Figure 8 Percentage of female users realising /q/ as [q] and [g] in Tunis.*

predominant. However, based on our analysis, it turns out that the exact opposite is in fact true.[28]

It will be interesting to observe the score assigned by Pearson's chi-squared test, to see if there is correlation between the realisation of

[28] However, it must be kept in mind that the songs that are part of our corpus are almost all by rappers from Tunis.

| ***bil-qala*** | ***bil-ɣala*** |
|---|---|
| 91.76% | 8.24% |

*TABLE 52. ANALYSIS OF /Q/ ENCODING IN RAP DATA.*

/q/ and textual genres (besides the musical genre of rap). This will be dealt with in the next section (4).

*Diphthongs*

Among the features to be observed, traditionally considered distinctive between conservative and innovative systems, is the evolution of etymological diphthongs. The diphthongs /ay/ and /aw/ are in fact preserved (in southern Tunisian dialects (Durand, 2007, 248)). However, there is a prevailing tendency for the monophtongisation of */ay/ into ī or ē and of */aw/ into ū or ō (Saada, 1984, 33-38). According to Gibson (2002), the most conservative systems, with diphthong realisations, are found in Sfax, Nabeul, and in the speech of old women in Tunis. Regarding the monophtongisation, the five-long-vowel system (ā, ē, ī, ō, ū) is found in several Hilali systems, but in particular in the Sahel village dialects (Mion, 2015, 271), while the three long vowel system (ā, ī, ū) is found in pre-Hilali systems, such as in Tunis and Sousse (Durand, 2007; Gibson, 2002; Mion, 2015).

In this regard, the analysis can inform us of certain aspects, such as the general percentage of the occurrence of diphthongs and monophthongation. However, the results of the distribution among the governorates cannot be considered to be perfectly informative; in fact, in order to simplify the automatic detection of patterns in the corpus, we adopted exclusion strategies that led to a reduction in the amount of data, particularly for diphthongs, and this amount cannot be considered sufficient for statistical analysis. Regarding the tokens that we excluded from our analysis, these are mainly the proper nouns (such as *Sayf* or *Nawres*), loanwords, and all verbs in order to avoid matches with first radical weak verbs, such as /ywarrī-hum/, 'he shows to them'.[29] As a result of these strategies, for example, the p-values in Pearson's chi-squared test between governorates (excluding Tunis) and the etymological diphthong reali-

[29] Past participles were not excluded. In addition, we decided not to exclude words containing gemination of the semivocal, given a substantial number of occurrences of reduction and diphthong preservation in cases such as /sīd-i/ and /sayyid-i/ for 'my lord'.

sation is necessarily insignificant, being 0.98. Instead, when using the same test, between governorates (including Tunis) and the etymological diphthong realisation, the value is lower than 0.01, which makes sense when considering the greater amount of data we dispose of for the capital than for other specific governorates, and the clear tendency that marks Tunis data (Table 55). In conclusion, what we can observe is a general trend by grouping together all the governorates excluding Tunis (as the model of the *koiné* hypothesis), in Table 53, and compare same to data from the capital, in Table 55, to see whether or not there is a convergence. Regarding Table 53, only governorates for which occurrences have been found have been reported on. Regarding the quantity of occurrences found for data in Table 53, 26.73% of the total data presents diphthongs, while 73.27% of the total data presents diphthong reductions to a long vowel, as shown in 54.[30]

| ***Diphthongisation*** | ***Monopht.*** |
|---|---|
| 26.73% | 73.27% |

*Table 54. Diphthong realisation in Tunisia.*

| ***Diphthongisation*** | ***Monopht.*** |
|---|---|
| 26.2% | 73.8% |

*Table 55. Diphthong realisation in Tunis data.*

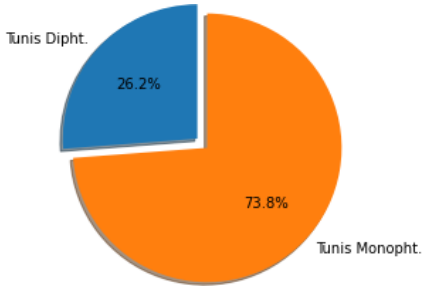
When examining the data in Table 54, it can be seen that most governorates (73.27%) exhibit forms of reduction of the original diphthongs */ay/ and */aw/. Furthermore, we can see that among the governorates that maintain etymological diphthong forms, alongside these, the governorates generally also present forms of monophtongisation. In particular, what is striking is the fact that the percentage of diphthong preservation is well distributed among governorates in the north (e.g., Ariana, Bizerte), the northwest (Jendouba), the Sahel (Sousse), and the south (Sfax, Gafsa, and Medenine). However, the same governorates also, almost always, show a good percentage of diphthong reduction, e.g., Bizerte (17.57%), Jendouba (5.41%), Sousse (10.81%), Sfax (6.76%), Gafsa (5.41%), and Medenine (6.76%). Finally, considering the high percentages of monophtongisation found for Bizerte, Sousse, Medenine, Mona-

[30] An interesting phenomenon to investigate in terms of this data, which has a good amount of monophtongisation data (approximately 650 occurrences), would be to observe the occurrences of monophtongisation in /ī/ vs. /ē/ and of /ū/ vs. /ō/, and the treatment of /ī/ before the 3rd s.f.p. pronoun suffix /-ha/. Regarding these phenomena, see Mion (2015). However, because of time constraints, this analysis will be postponed to future studies.

| Govs. except Tunis | *Diphthongisation* | *Monophtongisation* |
|---|---|---|
| *Ariana* | 14.81% | 1.35% |
| *Béja* | *n.o.* | 1.35% |
| *Sousse* | 7.41% | 10.81% |
| *Bizerte* | 7.41% | 17.57% |
| *Gabès* | 3.7% | 1.35% |
| *Nabeul* | 3.7% | 2.7% |
| *Jendouba* | 7.41% | 5.41% |
| *El Kef* | 3.7% | *n.o.* |
| *Kairouan* | *n.o.* | 4.05% |
| *Zaghouan* | *n.o.* | 1.35% |
| *Mahdia* | 3.7% | 2.7% |
| *Manouba* | 3.7% | 4.05% |
| *Medenine* | 7.41% | 6.76% |
| *Monastir* | 3.7% | 6.76% |
| *Gafsa* | 7.41% | 5.41% |
| *Sfax* | 11.11% | 6.76% |
| *Sidi Bouzid* | 3.7% | *n.o.* |
| *Siliana* | 3.7% | *n.o.* |
| *Ben Arous* | 3.7% | 18.92% |
| *Kasserine* | 3.7% | 2.7% |
| *Tot* | 100% | 100% |

*Table 53. Diatopic analysis of diphthong realisation in Tunisia (-Tunis). N.o. stands for 'no occurrences'.*

stir, Sfax and Ben Arous, we can conclude that, at least according to this data, it seems possible to hypothesise the diffusion of the monophtongisation system starting from the major urban centers of the country (Tunis, Bizerte, Sousse and Sfax).




These results invite us to at least leave open the question of the convergence of the Tunisian varieties on the model of Tunis (Table 55), which happens to present percentages of the distribution of diphthongs (preserved or reduced) which are very similar (if not almost identical) to those found for the sum of all the other governorates (excluding that of Tunis, Table 53). We must therefore consider it plausible that such a result can only be returned by data coming from social media, where evidently there is a certain tendency to speak according to a shared model, which seems to coincide with the Tunisian model (Table 55).

| Gov. of Tunis | *Diphthongisation* |
|---|---|
| **M** | 78.8% |
| **F** | 21.3% |

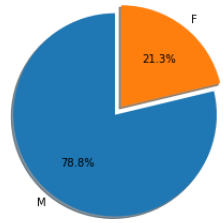



*Table 56. Diastratic analysis of diphthong realisation in Tunis data.*

| *Urban style* | *Gender opposition* |
|---|---|
| 98.32% | 1.68% |

*Table 57. Analysis of gender opposition in governorates except Tunis.*

| *Urban style* | *Gender opposition* |
|---|---|
| 98.48% | 1.52% |

*Table 58. Analysis of gender opposition in Tunis governorate.*

Finally, looking at the Tunis data in Table 56, it can be seen that if diastratic variation still exists, i.e., if the realisation of diphthongs is something typical for old women in Tunis, as noted by Singer (1984) and reiterated by Gibson (2002), this was not captured in our corpus, where the results indicate a rather clear male prevalence.

### *Gender opposition at 2nd singular person in verbs and pronouns*

There are two more analyses that need to be tackled in order to check the influence of Tunis Arabic on the rest of the country. In this case, two morphological issues were chosen to balance the analysis. The first of these is the gender opposition for the 2nd person singular, for both the independent personal pronoun system and verbal conjugation. In general, for urban languages (such as Sousse and Kairouane), the dialect of Tunis does not feature any gender opposition. However, across the rest of the country, according to the traditional description, it is possible to find, in the village or Bedouin dialects, oppositions both at the pronominal and verbal level (La Rosa, 2021; Mion, 2015, 273, 13-14)

The results have shown, in tables (57 and 58), a clear tendency for Tunisian systems to follow the Tunis model, where the occur-

| *Urban style* | *Village and Bedouin style* |
|---|---|
| 35.3% | 64.7% |

*TABLE 59. ANALYSIS OF PLURAL REALISATION IN GOVERNORATES EXCEPT TUNIS.*

| *Urban style* | *Village and Bedouin style* |
|---|---|
| 88.5% | 11.5% |

*TABLE 60. ANALYSIS OF PLURAL REALISATION IN TUNIS GOVERNORATE.*

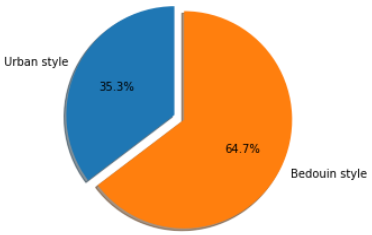



*FIGURE 9 PLURAL REALISATION OUTSIDE TUNIS.*

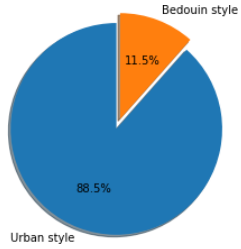



*FIGURE 10 PLURAL REALISATION IN TUNIS.*

rences of gender opposition are minimal (and restricted to the isolated pronominal), at least in the context of informal online writing.

*Realisation of -w plural morpheme in weak verbs*

The second phenomenon of a morphological nature is the realisation of the plural in weak verbs. In fact, the treatment of the plural morpheme -ū in the form of the perfect and imperfect is different in the pre-hilalic and hilalic systems, where, for the verb *mšā*, in the former case we have *mšāw-yəmšīw* 'to go', while in the latter we have *mšū-yəmšū* (Mion, 2015, 272). By inspecting the data and comparing the results obtained for the sum of all the governorates (excluding that of Tunis) with that of Tunis, a trend opposite to that of the previous section can clearly be seen. In the tables below (59 and 60), it can be seen that there is no convergence of the writing systems of the various governorates towards the Tunis model.

In the case of the gender opposition in the second person singular, users of online platforms, who are not from Tunis, are inclined to adopt an urban system. This tendency may be reinforced by the fact that they are addressing someone who is most likely not in their

circle of friends and whom they are addressing directly. Otherwise, the purpose of convergence towards a *koiné* might also be to reduce the ambiguity of discourse. The same need to tend toward a shared model does not seem to be felt in the case of using plural person verbs, perhaps because it is something that the user pays less attention to, is more difficult to 'level' on the *koiné* model, or is perhaps simply less perceived as a source of ambiguity for discussion in DNW. In conclusion, we can consider the hypothesis of the diffusion of the Tunis model in a koineisation process to be valid, which concerns the world of DNW. What remains to be clarified is the degree of spontaneity/formality that such a *koiné* represents, a topic which will be addressed in the next section.

## 4. Continuum of degree of formality

The last research path specifically concerns the topic of text specificity, and in particular aims to identify the differences between the genre of a text, namely social networks, forums and blogs, which are the three genres covered by TArC texts. In fact, the third hypothesis sees these three genres occupying different positions along a formality continuum. Typically, written speech is considered to be more formal than spoken speech, but, nowadays, with the advent of written messaging and social platforms such as Facebook, this paradigm appears to be obsolete. In the DNW, users employ writing, which is typically associated with formality, to communicate messages that are typically informal (Sullivan, 2017, 30). As already covered in Section 2, technology supports writing in whatever form is desired; users are not bound by any standards and are free to use whatever channels they choose to create and express social or cultural attitudes and identities, moving from standard to nonstandard varieties and back again in a more flexible way than is generally possible in a non-electronic written context. However, a text's formality is connected to the degree of spontaneity of the text, and accordingly social network texts are less formal in comparison to those of blogs, and that forums should find their place in the between. In order to analyse the formality degree we will observe some general phenomena typical of spontaneous text, such as punctuation, the use of emoticons and interjections, in the three textual genres in Section 4. Next, in Section 4, we will test whether our hypothesis about the distribution of prepositional patterns along a continuum of formality coincides with the distribution of the three textual gen-

| | *Social Networks* | *Forum* | *Blog* |
|---|---|---|---|
| **Interjection** | 46% | 32.5% | 18.4% |
| **Emotags** | 39.7% | 3.2% | 2.2% |
| **Final Punctuation** | 14.3% | 64.2% | 79.4% |
| **Tot** | 100% | 100% | 100% |

*TABLE 61. GENERAL ANALYSIS OF TEXT GENRE.*

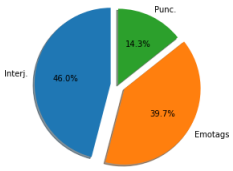



*FIGURE 11 SOCIAL NET.*

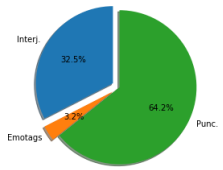



*FIGURE 12 FORUM*

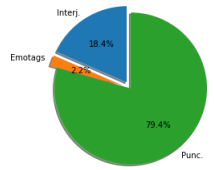



*FIGURE 13 BLOG*

res along the same continuum. Also in this case, the query tools used are available on the mentioned website.

### *Text genre general analysis*

In order to be able to base our vision on concrete data, of a continuum of formality on which to place the data of Social Networks and Blogs as extremes, and Forums as intermediate types of texts, we have identified three very simple parameters among the DNW characteristics presented by Thurlow and Poff (2013) and have discussed them in Section 2. These consist of the number of occurrences of interjections, the number of elements classified as emotags in our corpus (emoticons, smileys, etc.), and the number of occurrences of punctuation marks at the end of sentences. With regard to the first two, these are parameters, the size of which will be directly proportional to the value of the spontaneity of the text, while with regard to the last parameter, this will be inversely proportional to the degree of spontaneity.

The data presented in Table 61 fully supports the hypothesis regarding the positions that the three textual types take on the continuum of formality based on the degree of spontaneity of the implied texts. In fact, when observing the first two parameters, i.e. the number of interjections and emotags, it is possible to see that these are numerous in texts coming from social networks and forums. In the case of forums, the percentages decrease, getting closer to those found in blog texts. When instead observing the parameter consid-

ered in the last analysis, i.e. the punctuation at the end of a sentence, it is clear to see that, in the case of texts coming from social networks, there are very poor values (14.32%), while in the case of forums and blogs, the values are much higher (approximately 64% and 79% respectively). We can also say that forum texts are located almost halfway between the two extremes of the continuum in terms of the first parameter (the number of interjections), while the second and third are closer to the blog typology than to the social network typology.

### *Nominal Phrase distribution*

Considering the possible combinations mentioned at the end of the previous paragraph, our hypothesis is that, in the case of a Prepositional Phrase (PP), which includes the article (det), the combinations are arranged along a continuum from the most representative of Tunisian orality ([prep+det n]) to the most representative of the orthographic system of Standard Arabic ([prep det+n]).

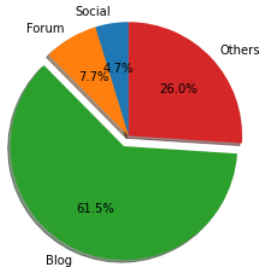
Therefore, Section 3 will rejoin the discussion on taking up the phenomenon of the orthographic realisation of the Prepositional Phrase (PP), which seems to offer some evidence in this regard. Therefore, in Section 3, code-switching phenomena was observed in Arabizi, in order to prove its role in the preference of the PP scheme [prep+det n]. Indeed, by observing the data, it was possible to assert that Arabizi encoded by the Latin alphabet seems to be have been influenced by contact with European languages.

In Table 37, it can be seen that among the defined NPs that witness a det's assimilation, only 10% of the defined NPs realise the pattern [det n]. Therefore, the tendency seems to be to graphically realise the defined NP as [det+n]. Instead, in the cases of defined NPs that do not respect article assimilation orthographic rules, the prevailing orthographic tendency seems to be to coincide with [det n]. In simple terms, it seems that if the article and noun do not 'merge' on the phonetic level, then they should not do so on the graphemic level either. Moreover, Table 62 shows that this 'rule' of encoding the defined NP of Arabizi seems to be more or less respected according to the degree of formality of the textual genre.

| *def.NP* | **Social** | **Forum** | **Blog** | **Others** |
|---|---|---|---|---|
| *With space* | 4.7% | 7.7% | 61.5% | 26% |

*TABLE 62. DISTRIBUTION OF SPACE IN NP THROUGH TARC GENRE. 'OTHERS' CONSISTS OF RAP LYRICS.*

In fact, the data confirms that separating det and N with a space is an orthographic choice which is more consistent (61.5%) in the contexts of more controlled writing contexts, such as blogs, but only occurs in around 4.7% on social networks and 7.7% in forums.




*DISTRIBUTION SPACED NP THROUGH TARC GENRES*

## 5. Conclusions

In order to carry out these analyses independently, but at the same time ensure that they are interrelated, a phenomenon of a syntactic nature with orthographic fallout, which is very frequent in Tunisian Arabizi, was selected. This is the definite prepositional phrase, the realisation of which has been defined, just for descriptive purposes, as divergent from that of Standard Arabic. In fact, in Tunisian Arabic, it is possible to find orthographic patterns that represent a kind of articulated preposition, where the article is graphically joined to the preposition that precedes it, whether it is proclitic or not. The orthographic rendering of the articulated preposition in Tunisian Arabic appears similar to the modality found in European languages and different from the typical realisation of it in Standard Arabic, where the two elements are graphically joined only in the presence of some proclitic prepositions. By analysing the frequency of the various prepositional patterns, we have come to the conclusion that the patterns are arranged along a continuum ranging from the maximum representation of the Tunisian oral system to that of the writing habits acquired by writing in Arabic characters.

At this point, the role that linguistic contact between French and Tunisian Arabic plays in these graphic realisations of the prepositional phrase have also been considered, as well as the very nature of the Tunisian definiteness mark, which, in some contexts, seems to overlap with the relative mark. Having observed a key role of the

definiteness mark in the formation of these prepositional patterns, we decided to investigate them further through diachronic analysis of the TArC data. This naturally led us to the scope of the second question related to settling trends in Arabizi writing practices. The TArC data contain texts ranging from 2005 to 2020; however, the amounts of data for each year are not equivalent to each other. As such, other strategies were put in place in order to ensure balanced analyses, for example, through the use of statistical tests such as Pearson's chi-squared test. In order to give an example in this regard, we carried out a small analysis on one of the most friendly features of Arabizi, namely the use of numerals as graphemes, which were previously defined as arithmographs. We then continued our preliminary diachronic analyses of the distribution of prepositional phrase patterns. Through these analyses, we found a correlation between two types of prepositional patterns, the [prep n] type and the [prep+det n] type. The second type is what we have defined as the representational pole of the Tunisian oral system in the analyses related to quasi-orality. Given the correlations found between the two patterns in the analyses related to settling trends, we were also able to define these two patterns as being the same type of prepositional pattern (where the noun is separated from the preposition) with variations on the morphological level based on the presence or absence of the definiteness mark.

In order to investigate this issue further by scrutinising the morpho-syntactic intersection with the phonetic plane, we then looked at whether the presence of coronal phonemes played a role in the realisation of one pattern rather than another. In fact, we have often defined orthographic variation within the Arabizi system as providing a certain degree of freedom for the writer in terms of representing or not representing certain linguistic phenomena. One of the phenomena in question is that of article assimilation in the presence of nouns beginning with a coronal phoneme. Tunisian Arabizi writers sometimes do not represent article assimilation. Therefore, we observed that in the distribution of prepositional phrases containing nouns that should assimilate the article (because they begin with a coronal phoneme), when assimilation is not graphically represented, the noun is often separated from what precedes it, regardless of whether this is an article or preposition (with a frequency of 88.9%). One of the plausible reasons for this is that the writer, in not rendering the assimilation graphically, feels the need to adopt a coherent behaviour at the orthographic level, keeping separate, at the graphic level, the elements that would not produce assimilation

at the phonological level, which he has, for some reason, chosen not to render at the orthographic level.

Diachronic analyses of the graphical rendering of assimilation were also conducted, so as to find trends that could only be explained by interviewing native speakers. In this regard, we have already stated our interest in conducting interviews dedicated to the topic with bloggers with whom we have already made contact. In fact, the tendency in research over the years seems to be mainly towards graphically rendering the assimilation; however, this trend has seemed to undergo a sudden reversal over recent years, due to the spread of social networks coinciding with the political upsets of the Arab revolutions. With regard to the graphic rendering of assimilation, we also found a tendency to maintain a type of handwriting which is coherent with the phonetic plan in the 35-50 age group (80%) and among female writers (77.4%). On the other hand, those of the male gender seemed to care less (24.7%) about a 'phonetically correct' graphic rendering. We have therefore concluded that the graphical rendering of assimilation seems to be a diastratically influenced issue.

Finally, we have also observed the influence of the French language on these spelling practices. Indeed, it is clear to see the distribution of prepositional phrases that include marked nominal syntagmas at the code-switching level. Again, and especially in this case, the [prep+det n] pattern was found with a very high frequency. We then conducted these analyses by taking into consideration the type of users involved, as well as the textual genres. It appeared that those making the most use of intrasentential code-switching are female forums users.

As mentioned earlier in this chapter, these findings reinforce the previous observation regarding the female gender's preference for an orthographic rendering of assimilation. Women seem to be more 'precise' in encoding the morpho-phonological phenomena of Tunisian, and at the same time, also tend to use a lot of French terminology in their everyday exchanges; in this case, using a 'French-style' encoding of the prepositional phrase ([prep+det n]). We also mentioned that this is in line with already widespread linguistic observations regarding the language of women, particularly from Tunis, and their predilection for French. We also speculated about the identity significance that this space, i.e. the forum, has for women's social groups, referring to the meaning of 'new communicative space' as exposed by Woolard (1998). Undoubtedly this issue deserves further investigation which we reserve for the future, indeed,

the kind of contact phenomena discussed here could fall within the sphere of code-switching, but it will be interesting to separate such elements from nonce borrowing and the occasional insertion of isolated lexical elements.

With regard to the questions posed about the type of language that can be encountered on social networks, and from which the path of analysis inherent to the settling trends of Tunisian Arabizi was then born, we chose to observe the distribution of some specific features in particular. This included those traditionally considered to be characteristics of diatopic variation within Tunisian Arabic. In fact, we also discussed the hypothesis regarding whether or not, within Arabizi practices, a language influenced by the *koiné* of Tunis can be identified. We then observed the realisation of the phoneme /qāf/ in the data from the different governorates collected in TArC, before comparing them with those from Tunis. Our data seemed to confirm the presence of an ongoing process of koineisation based on the model of the capital, which have already been attested to in previous studies to which we refer in the chapter.

Given that hypotheses have been raised in the past about a diastratic variation of this phoneme realisation, an analysis was also performed for another issue by observing the variations in sexual genders. These analyses highlighted the strong preference (of both genders) for the orthographic realisation of /q/ in *q* or *9* in the governorate of Tunis, and a clear preference of the male gender for the orthographic realisation in *g* in all other governorates. The result obtained in Tunis was at odds with the fact that, in general, the realisation in *q* is considered to be a feminine practice. This led us to a theory that perhaps this phonological phenomenon is more alive in rooted linguistic ideologies than in contemporary reality. Another of the characteristics observed in the analyses related to the issue of the *koiné* was that of the realisation of diphthongs. As stated earlier in the chapter, we quickly realized, through our analysis, that this can only inform us of certain aspects such as the general percentage occurrence of diphthongs and monotone. However, we cannot consider the results of the distribution across governorates as being completely accurate. In fact, in order to simplify the automatic detection of examples in the corpus, we adopted exclusion strategies that led to a reduction in the amount of data, particularly for diphthongs, and this amount was not considered sufficiently large for statistical analysis. However, we report the results of these analyses, as well as our intention to extend this field of study in the future. It appears that the trend for the most part, both in Tunis and in

all the other governorates, is that of reducing the diphthong to a long vowel (monophtongisation). What left us a little surprised was the fact that the diastratic analysis shows the clear preference of the male gender (78.8%) compared to the female gender (21.3%) for the maintenance of the diphthong. It must be specified that the conservation of diphthongs by women in the old city of Tunis is a phenomenon described many years ago, which would certainly have to be verified to be sure that it still exists. Thus, the data may not be surprising at all, since women may have changed to the reduction of diphthongs a while ago and the graphic rendering of diphthongs (not equivalent to the phonetic one) may have entirely different reasons. This intriguing discovery is worthy of further study in the future. In order to conclude the analysis on the process of koineisation, we have finally observed two other variables, namely that of gender opposition and that of the realisation of the morpheme -w of the plural in weak verbs. The first shows the clear tendency of the varieties of Tunisian excluding that of Tunis to overlap with the Tunis model, the second instead seems to maintain a certain divergence between the two, in which Tunis remains clearly urban, while the sum of the other governorates exhibits the opposite trend.

Once we had highlighted these dynamics, we wanted to analyse the formality continuum theory exposed above as the third analysis path. In a very simple way, we identified informative text features about the degree of formality and then performed analysis on them. These were the presence of interjections, smileys or emoticons, and final punctuation marks. The first two are typical of informal texts, while the latter is typical of more controlled texts. Our analyses confirmed the distribution of the three textual genres on the continuum as assumed, i.e. the blog genre being the more formal and the opposite pole of the social network genre.

Finally, once this distribution of textual genres was confirmed, we could reconnect to the observations made about the most frequent prepositional pattern in Arabizi Tunisian ([prep(+det) n]), on the basis of its distribution in these three textual genres. What we had previously observed with regards to this pattern was its diastratic influenced variation, as it is very much in use among female forum users, in the case of prepositional phrases marked by code-switching. We had therefore hypothesised its use as an orthographic choice, even in the case of nominal and prepositional phrases that include nouns that should have produced an assimilation of the article. In fact, in the case where such assimilation was not rendered graphically, with

a high percentage (approximately 89%), we had encountered precisely this type of prepositional pattern with an isolated noun. In the end, the analysis on textual genres, which showed a high percentage of the same scheme in texts coming from blogs (61.5%), seems to definitively confirm the hypothesis that this scheme presupposes a conscious choice of the writer, i.e. an orthographic choice made in a controlled textual context.

# BIBLIOGRAPHY


1. Abainia, Kheireddine; Ouamour, Siham, and Sayoud, Halim. A novel robust Arabic light stemmer. *Journal of Experimental & Theoretical Artificial Intelligence*, 290 (3):0 557–573, 2017.
2. Abd-El-Jawad, Hassan R. Cross-dialectal variation in Arabic: Competing prestigious forms. *Language in Society*, pages 359–367, 1987.
3. Abdelali, Ahmed; Darwish, Kareem; Durrani, Nadir, and Mubarak, Hamdy. Farasa: A fast and furious segmenter for Arabic. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*, pages 11–16, 2016.
4. Abdelali, Ahmed; Mubarak, Hamdy; Samih, Younes; Hassan, Sabit, and Darwish, Kareem. Arabic dialect identification in the wild. *arXiv preprint arXiv:2005.06557*, 2020.
5. Abdelkader, Ben. *Peace Corps English-Tunisian Arabic Dictionary*. ERIC Cleringhouse, Washington D.C., 1977.
6. Abdul-Mageed, Muhammad and Diab, Mona. Awatif: A multi-genre corpus for modern standard Arabic subjectivity and sentiment analysis. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC)*, volume 515, pages 3907–3914, 2012.
7. Abdul-Mageed, Muhammad and Diab, Mona. Sana: A large scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, pages 1162–1169, 2014.
8. Abdul-Mageed, Muhammad; Alhuzali, Hassan, and Elaraby, Mohamed. You tweet what you speak: A city-level dataset of Arabic dialects. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, 2018.
9. Abdul-Mageed, Muhammad; Zhang, Chiyu; Elmadany, Abdel-Rahim, and Ungar, Lyle. Toward micro-dialect identification in diaglossic and code-switched environments. *arXiv preprint arXiv:2010.04900*, 2020.

10. Abdul-Mageed, Muhammad; Zhang, Chiyu; Elmadany, Abdel-Rahim; Bouamor, Houda, and Habash, Nizar. Nadi 2021: The second nuanced Arabic dialect identification shared task. *arXiv preprint arXiv:2103.08466*, 2021.
11. Abidi, Karima and Smaïli, Kamel. An automatic learning of an algerian dialect lexicon by using multilingual word embeddings. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, 2018.
12. Abidi, Karima; Menacer, Mohamed Amine, and Smaili, Kamel. Calyou: A comparable spoken algerian corpus harvested from youtube. In *18th Annual Conference of the International Communication Association (Interspeech)*, 2017.
13. Adouane, Wafia; Semmar, Nasredine; Johansson, Richard, and Bobicev, Victoria. Automatic detection of arabicized berber and Arabic varieties. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 63–72, 2016.
14. Akbar, Rahima. Arabizi among kuwaiti youths: Reshaping the standard Arabic orthography. *International Journal of English Linguistics*, 90 (1):0 301–323, 2019.
15. Al-Badrashiny, Mohamed; Eskander, Ramy; Habash, Nizar, and Rambow, Owen. Automatic transliteration of romanized dialectal Arabic. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 30–38, 2014.
16. Al-Sabbagh, Rania and Girju, Roxana. Yadac: Yet another dialectal Arabic corpus. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC)*, pages 2882–2889, 2012.
17. Al Sallab, Ahmad; Hajj, Hazem; Badaro, Gilbert; Baly, Ramy; El-Hajj, Wassim, and Shaban, Khaled. Deep learning models for sentiment analysis in Arabic. In *Proceedings of the second workshop on Arabic natural language processing*, pages 9–17, 2015.
18. Al-Shargi, Faisal and Rambow, Owen. Diwan: A dialectal word annotation tool for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 49–58, 2015.
19. Al-Shargi, Faisal; Kaplan, Aidan; Eskander, Ramy; Habash, Nizar, and Rambow, Owen. Morphologically annotated corpora and morphological analyzers for moroccan and sanaani yemeni Arabic. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC)*, 2016.
20. Al-Twairesh, Nora; Al-Khalifa, Hend; Al-Salman, AbdulMalik, and Al-Ohali, Yousef. Arasenti-tweet: A corpus for Arabic sentiment analysis of saudi tweets. *Procedia Computer Science*, 117:0 63–72, 2017.

21. Al-Twairesh, Nora; Al-Matham, Rawan; Madi, Nora; Almugren, Nada; Al-Aljmi, Al-Hanouf; Alshalan, Shahad; Alshalan, Raghad; Alrumayyan, Nafla; Al-Manea, Shams; Bawazeer, Sumayah, and others, . Suar: Towards building a corpus for the saudi dialect. *Procedia computer science*, 142:0 72–82, 2018.
22. Albirini, Abdulkafi. *Modern Arabic sociolinguistics: Diglossia, variation, codeswitching, attitudes and identity*. Routledge, 2016.
23. Alghamdi, Hamdah and Petraki, Eleni. Arabizi in saudi arabia: A deviant form of language or simply a form of expression? *Social Sciences*, 70 (9):0 155, 2018.
24. Alharbi, Randah; Magdy, Walid; Darwish, Kareem; Abdelali, Ahmed, and Mubarak, Hamdy. Part-of-speech tagging for Arabic gulf dialect using bi-lstm. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, 2018.
25. Ali, Ahmed; Dehak, Najim; Cardinal, Patrick; Khurana, Sameer; Yella, Sree Harsha; Glass, James; Bell, Peter, and Renals, Steve. Automatic dialect detection in Arabic broadcast speech. *arXiv preprint arXiv:1509.06928*, 2015.
26. Ali, Mohamed. Character level convolutional neural network for Arabic dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 122–127, 2018.
27. Allal, Amin and Geisser, Vincent. *Tunisie. Une démocratisation au-dessus de tout soupçon?* Cnrs, 2018.
28. Allehaiby, Wid H. Arabizi: An analysis of the romanization of the Arabic script from a sociolinguistic perspective. *Arab World English Journal*, 40 (3), 2013.
29. Almeman, Khalid and Lee, Mark. Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words. In *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6. IEEE, 2013.
30. Alonso, Héctor Martínez and Plank, Barbara. When is multitask learning effective? semantic sequence prediction under varying data conditions. *arXiv preprint arXiv:1612.02251*, 2016.
31. Alsarsour, Israa; Mohamed, Esraa; Suwaileh, Reem, and Elsayed, Tamer. Dart: A large dataset of dialectal Arabic tweets. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, 2018.
32. Aly, Mohamed and Atiya, Amir. Labr: A large scale Arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 494–498, 2013.

33. Ameur, Hanen; Jamoussi, Salma, and Hamadou, Abdelmajid Ben. Exploiting emoticons to generate emotional dictionaries from facebook pages. In *Intelligent Decision Technologies 2016*, pages 39–49. Springer, 2016.
34. Androutsopoulos, Jannis. Language change and digital media: a review of conceptions and evidence. *Standard languages and language standards in a changing Europe*, 1:0 145–159, 2011.
35. Androutsopoulos, Jannis. "Greeklish": Transliteration practice and discourse in the context of computer-mediated digraphia. *Orthography as social action: Scripts, spelling, identity and power*, pages 359–392, 2012.
36. Androutsopoulos, Jannis and Schmidt, Gurly. Sms-kommunikation: Ethnografische gattungsanalyse am beispiel einer kleingruppe. *Zeitschrift für angewandte Linguistik*, 36:0 49–80, 2002.
37. Anis, Jacques. Pour une graphématique des usages: le cas de la ponctuation dans le dialogue télématique. *Linx*, 310 (2):0 81–97, 1994.
38. Anis, Jacques. *Texte et ordinateur: l'écriture réinventée?* De Boeck Supérieur, 1998.
39. Anis, Jacques. Communication électronique scripturale et formes langagières. *Actes des Quatrièmes rencontres Réseaux humains/Réseaux technologiques*, 31, 2003.
40. Anis, Jacques. La dynamique discursive d'une liste de diffusion: analyse d'une interaction sur «typographie@ irisa. fr». *Les Carnets du Cediscor. Publication du Centre de recherches sur la didacticité des discours ordinaires*, 8:0 39–56, 2004.
41. Anis, Jacques. Neography: Unconventional spelling in french sms text messages. *The multilingual internet: Language, culture, and communication online*, 87:0 115, 2007.
42. Antoun, Wissam; Baly, Fady, and Hajj, Hazem. Arabert: Transformer-based model for Arabic language understanding. *arXiv preprint arXiv:2003.00104*, 2020.
43. Aridhi, Chaima; Achour, Hadhemi; Souissi, Emna, and Younes, Jihene. Word-level identification of romanized Tunisian dialect. In *International Conference on Applications of Natural Language to Information Systems*, pages 170–175. Springer, 2017.
44. Arts, Tressy; Belinkov, Yonatan; Habash, Nizar; Kilgarriff, Adam, and Suchomel, Vit. artenten: Arabic corpus and word sketches. *Journal of King Saud University-Computer and Information Sciences*, 260 (4):0 357–371, 2014.
45. Attia, Abdelmajid. Differents registres de l'emploi de l'arabe en tunisie. *Revue Tunisienne des Sciences Sociales*, 8:0 115–150, 1966.

46. Auer, Peter. Führt dialektabbau zur stärkung oder schwächung der standardvarietät? zwei phonologische fallstudien. In Klaus, Mattheier J. and Radtke, Edgar (ed.), *Standardisierung und Destandardisierung europäischer Nationalsprachen*, pages 135–140. M. u.a.: Lang, Frankfurt, 1997.
47. Auer, Peter. From codeswitching via language mixing to fused lects: Toward a dynamic typology of bilingual speech. *International journal of bilingualism*, 30 (4):0 309–332, 1999.
48. Azouaou, Faical and Guellil, Imane. Alg/fr: A step by step construction of a lexicon between algerian dialect and french. In *The 31st Pacific Asia Conference on Language, Information and Computation PACLIC*, volume 31, 2017.
49. Azzopardi-Alexander, Marie and Borg, Albert. *Maltese*. Routledge, 2013.
50. Babaei, Arsam. Farsi or pinglish; effective factors on script selection in user's chats in telegram and whatsapp messengers. *Media Studies*, 160 (4):0 35–47, 2022.
51. Baccouche, Taïeb. *L'emprunt en arabe moderne*. Académie tunisiennes des sciences des lettres et des arts, Beït al-Hikma, 1994.
52. Badaro, Gilbert; Jundi, Hussein; Hajj, Hazem; El-Hajj, Wassim, and Habash, Nizar. Arsel: A large scale Arabic sentiment and emotion lexicon. *OSACT*, 3:0 26, 2018.
53. Badawī, El-Saʕīd Muḥammad. Levels of modern Arabic in egypt. *Cairo: dar al-ma'arif [In Arabic]*, 1973.
54. Bahdanau, Dzmitry; Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
55. Bahl, Lalit R; Brown, Peter F; de Souza, Peter V, and Mercer, Robert L. A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 370 (7):0 1001–1008, 1989.
56. Baker, Philip. Developing ways of writing vernaculars: problems and solutions in a historical perspective. In Tabouret-Keller, Andree; LePage, Robert; Gardner-Chloros, Penelope, and Varro, Gabrielle (ed.), *Vernacular Literacy*. Clarendon, New York, 1997.
57. Bannour, Abderrazak. Brève mise au point sur la lingua franca en méditerrannée. *Les langues en Tunisie: Etat des lieux et perspectives*, pages 241–259, 2000.
58. Baron, Naomi S. Computer mediated communication as a force in language change. *Visible language*, 180 (2):0 118, 1984.
59. Baron, Naomi S. Letters by phone or speech by other means: The linguistics of email. *Language & Communication*, 180 (2):0 133–170, 1998.

60. Baron, Naomi S. *Always on: Language in an online and mobile world*. Oxford University Press, 2010.
61. Bashraheel, Laura. 9aba7 2l5air! texting, arab style. *Arab News.(26 August 2008).[Online] Available at: http://www. arabnews. com/node/315362 (Accessed: 15/03/2021)*, 2008.
62. Bassiouney, Reem. *Arabic sociolinguistics: Topics in Diglossia, Gender, Identity, and Politics*. Georgetown University Press, 2009.
63. Belgacem, Mohamed. Construction d'un corpus robuste de différents dialectes arabes. *Actes des 8emes Rencontres Jeunes Chercheurs en Parole*, 33, 2009.
64. Belgacem, Mohamed; Antoniadis, Georges, and Besacier, Laurent. Automatic identification of Arabic dialects. In *LREC*, 2010.
65. Bender, Emily M. and Lascarides, Alex. *Linguistic fundamentals for natural language processing II: 100 essentials from semantics and pragmatics*. Morgan & Claypool Publishers series, 2019.
66. Bengio, Yoshua; Ducharme, Réjean, and Vincent, Pascal. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, pages 932–938, 2001.
67. Bengio, Yoshua; Ducharme, Réjean; Vincent, Pascal, and Janvin, Christian. A neural probabilistic language model. *The journal of machine learning research*, 3:0 1137–1155, 2003.
68. Bengio, Yoshua; LeCun, Yann, and others, . Scaling learning algorithms towards ai. *Large-scale kernel machines*, 340 (5):0 1–41, 2007.
69. Benkato, Adam. From medieval tribes to modern dialects: On the afterlives of colonial knowledge in Arabic dialectology. *Philological Encounters*, 40 (1-2):0 2–25, 2019.
70. Benkato, Adam. Maghrebi Arabic. *Arabic and contact-induced change*, 1:0 197, 2020.
71. Benmamoun, Elabbas and Bassiouney, Reem. *The Routledge handbook of Arabic linguistics*. Routledge, 2017.
72. Benmayouf, Yamina-Chafia. *L'arabe parlé par les cadres algériens (ou l'arabe algérien médian). Description linguistique.* PhD thesis, Paris 5, 2003.
73. Berruto, Gaetano. *Fondamenti di sociolinguistica*. Laterza, 1995.
74. Bettaïeb, Viviane. *Dégage: la révolution tunisienne, 17 décembre 2010-14 janvier 2011*. Editions du Layeur, Paris, 2011.
75. Biadsy, Fadi and Hirschberg, Julia. Using prosody and phonotactics in Arabic dialect identification. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.
76. Biadsy, Fadi; Hirschberg, Julia, and Habash, Nizar. Spoken Arabic dialect identification using phonotactic modeling. In *Proceedings of the eacl 2009 workshop on computational approaches to semitic languages*, pages 53–61, 2009.

77. Bianchi, Robert Michael. 3arabizi-when local Arabic meets global english. *Acta Linguistica Asiatica*, 20 (1):0 89–100, 2012.
78. Bianchi, Robert Michael. Arab english: The case of 3arabizi/arabish on mahjoob. com. *Voices in Asia Journal*, 10 (1):0 82–96, 2013.
79. Bies, Ann; Song, Zhiyi; Maamouri, Mohamed; Grimes, Stephen; Lee, Haejoong; Wright, Jonathan; Strassel, Stephanie; Habash, Nizar; Eskander, Ramy, and Rambow, Owen. 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, 2014.
80. Bingel, Joachim and Søgaard, Anders. Identifying beneficial task relations for multi-task learning in deep neural networks. *arXiv preprint arXiv:1702.08303*, 2017.
81. Blanc, Haim. *Style variations in spoken Arabic: A sample of interdialectical educated conversation*. Harvard University Press, 1960.
82. Blau, Joshua. Classical Arabic, middle Arabic, middle Arabic literary standard, neo-arabic, judaeo-arabic and related terms. *Joshua Finkel Festschrift*, pages 37–40, 1974.
83. Blau, Joshua. The state of research in the field of the linguistic study of middle Arabic. *Arabica*, 280 (2):0 187–203, 1981.
84. Blom, Jan-Petter and Gumperz, John J. Social meaning in linguistic structure: Code-switching in norway. In *Directions in sociolinguistics: The ethnography of communication*, pages 407–434. Holt, Rinehart and Winston, 1972.
85. Boberg, Charles; Nerbonne, John A, and Watt, Dominic James Landon. *The handbook of dialectology*. Wiley Online Library, 2018.
86. Bou Tanios, Jennifer. *Language Choice and Romanization Online by Lebanese Arabic Speakers*. Master's thesis, University of Pompeu Fabra, Barcelona, 2016.
87. Bouamor, Houda; Habash, Nizar, and Oflazer, Kemal. A Multidialectal Parallel Corpus of Arabic. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, pages 1240–1245, 2014.
88. Bouamor, Houda; Habash, Nizar; Salameh, Mohammad; Zaghouani, Wajdi; Rambow, Owen; Abdulrahim, Dana; Obeid, Ossama; Khalifa, Salam; Eryani, Fadhl; Erdmann, Alexander, and Oflazer, Kemal. The madar Arabic dialect corpus and lexicon. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, 2018.
89. Bouchlaghem, Rihab; Elkhlifi, Aymen, and Faiz, Rim. Tunisian dialect wordnet creation and enrichment using web resources

and other wordnets. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 104–113, 2014.

90. Boudlal, Abderrahim; Lakhouaja, Abdelhak; Mazroui, Azzeddine; Meziane, Abdelouafi; Bebah, MOAO, and Shoul, Mostafa. Alkhalil morpho sys1: A morphosyntactic analysis system for Arabic texts. In *International Arab conference on information technology*, pages 1–6. Elsevier Science Inc New York, NY, 2010.
91. Boujelbane, Rahma. Génération des corpus en dialecte tunisien pour la modélisation de langage d'un système de reconnaissance. In *Proceedings of RECITAL, Les Sables d'Olonne*, pages 206–216, 2013.
92. Boujelbane, Rahma; Khemekhem, Mariem Ellouze, and Belguith, Lamia Hadrich. Mapping rules for building a Tunsian dialect lexicon and generating corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 419–428, 2013a.
93. Boujelbane, Rahma; Khemekhem, Mariem Ellouze; BenAyed, Siwar, and Belguith, Lamia Hadrich. Building bilingual lexicon to create dialect Tunsian corpora and adapt language model. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 88–93, 2013b.
94. Boujelbane, Rahma; Mallek, Mariem; Ellouze, Mariem, and Belguith, Lamia Hadrich. Fine-grained pos tagging of spoken Tunsian dialect corpora. In *International Conference on Applications of Natural Language to Data Bases/Information Systems*, pages 59–62. Springer, 2014.
95. Boujelbane, Rahma; Ellouze, Mariem; Béchet, Frédéric, and Belguith, Lamia. De l'arabe standard vers l'arabe dialectal: projection de corpus et ressources linguistiques en vue du traitement automatique de l'oral dans les médias tunisiens. *Revue TAL*, 550 (2):0 73–96, 2015.
96. Boussofara, Naima. Learning the 'linguistic habitus' of a politician: A presidential authoritative voice in the making. *Journal of Language and Politics*, 50 (3):0 325–358, 2006. doi: https://doi.org/10.1075/jlp.5.3.04bou.
97. Boussofara-Omar, Naima. Neither third language nor middle varieties but diglossic switching. *Zeitschrift für arabische Linguistik*, pages 55–80, 2006.
98. Bridle, John S. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, pages 227–236. Springer, 1990.
99. Brill, Eric; Magerman, David; Marcus, Mitchell, and Santorini, Beatrice. Deducing linguistic structure from the statistics of large

corpora. In *Proceedings of the 5th Jerusalem Conference on Information Technology, 1990.'Next Decade in Information Technology'*, pages 380–389. IEEE, 1990.
100. Bromley, Myron H. *The Phonology of Lower Grand Valley Dani*. Brill, 1961.
101. Bronckart, Jean-Paul and Schneuwly, Bernard. *Vygotsky aujourd'hui*. Delachaux et Niestlé, Lausanne, 1985.
102. Brown, Peter F; Cocke, John; Della Pietra, Stephen A; Della Pietra, Vincent J; Jelinek, Frederick; Lafferty, John; Mercer, Robert L, and Roossin, Paul S. A statistical approach to machine translation. *Computational linguistics*, 160 (2):0 79–85, 1990.
103. Brown, Peter F; Della Pietra, Stephen A; Della Pietra, Vincent J, and Mercer, Robert L. Word-sense disambiguation using statistical methods. In *29th Annual meeting of the Association for Computational Linguistics*, pages 264–270, 1991.
104. Brugnatelli, Vermondo. Il berbero di jerba: rapporto preliminare. *Incontri Linguistici*, 21:0 115–128, 1998.
105. Brustad, Kristen. *The syntax of spoken Arabic: A comparative study of Moroccan, Egyptian, Syrian, and Kuwaiti dialects*. Georgetown University Press, 2000.
106. Buckwalter, T. Buckwalter Arabic morphological analyzer (bama) version 2.0. linguistic data consortium (ldc) catalogue number ldc2004l02. Technical report, ISBN1-58563-324-0, 2004.
107. Buckwalter, Tim. Buckwalter Arabic morphological analyzer version 1.0. linguistic data consortium. *University of Pennsylvania, LDC Catalog No.: LDC2002L49*, 2002.
108. Buckwalter, Tim and Parkinson, Dilworth. *A frequency dictionary of Arabic: Core vocabulary for learners*. Routledge, 2014.
109. Burnard, Lou. Metadata for corpus work. In Wynne, Martin (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, pages 41–58. Oxbow Books for the Arts and Humanities Data Service, Oxford, 2004.
110. Campanini, Massimo. *Ibn Khaldûn e la Muqaddima. Passato e futuro del mondo arabo*. La Vela, Viareggio, 2019.
111. Camps, Gabriel. Comment la berbérie est devenue le maghreb arabe. *Revue des mondes musulmans et de la Méditerranée*, 350 (1):0 7–24, 1983.
112. Canavan, Alexandra; Zipperlen, George, and Graff, David. *CALLHOME Egyptian Arabic Speech LDC97S45*. Linguistic Data Consortium, Philadelphia, 1997.
113. Caruana, Rich. Multitask learning. *Machine learning*, 280 (1):0 41–75, 1997.
114. Caruana, Sandro. Terminology of italian origin used in eu maltese. In *Introducing Maltese Linguistics: Selected papers from the 1st*

*International Conference on Maltese Linguistics, Bremen, 18 20 October, 2007*, volume 113, page 355. John Benjamins Publishing, 2009.

115. Cassola, Arnold. A note on the dating of h, gh and x in maltese. In *Perspective on Maltese Linguistics*, volume 14, pages 13–22. Verlag, Berlin, 2014.
116. Caubet, Dominique. Propositions concernant la notation usuelle de l'arabe maghrébin: graphie arabe et graphie latine. *Paris, INALCO*, 2000.
117. Caubet, Dominique. Apparition massive de la darija à l'écrit à partir de 2008-2009: sur le papier ou sur la toile: quelle graphie? quelles régularités? *De los manuscritos medievales a internet: la presencia del árabe vernáculo en las fuentes escritas. En: Colección Estudios de Dialectología Árabe*, 6:0 377–402, 2012.
118. Caubet, Dominique. Morocco: An informal passage to literacy in dārija (moroccan Arabic). In *The Politics of Written Language in the Arab World*, pages 116–141. Brill, 2017a.
119. Caubet, Dominique. Darija and the construction of "moroccanness". *Identity and Dialect Performance. A Study of Communities and Dialects. London and New York: Routledge/Taylor*, pages 99–124, 2017b.
120. Caubet, Dominique. New Elaborate Written Forms in Darija: Blogging, Posting and Slamming in Morocco. In Benmamoun, Elabbas and Bassiouney, Reem (ed.), *The Routledge Handbook of Arabic Linguistics*, pages 387–406. Routledge, 2018.
121. Caubet, Dominique. Vers une littératie numérique pour la darija au maroc, une démarche collective. In Miller, Catherine; Barontini, Alexandrine; Germanos, Marie-Aimée; Guerrero, Jairo Guerrero, and Pereira, Christophe (ed.), *Studies on Arabic Dialectology and Sociolinguistics. Proceedings of the 12th International Conference of AIDA*. Livres de l'IREMAM, 2019.
122. Cerruti, Massimo and Regis, Riccardo. Code switching'e teoria linguistica: la situazione italoromanza. *Italian Journal of Linguistics*, 170 (1):0 179, 2005.
123. Chalabi, Achraf and Gerges, Hany. Romanized Arabic transliteration. In *Proceedings of the Second Workshop on Advances in Text Input Methods*, pages 89–96, 2012.
124. Charniak, Eugene; Gavin, Michael, and Hendler, James A. *The Frail/NASL reference manual*. Brown University, Department of Computer Science, 1983.
125. Chen, Danqi and Manning, Christopher D. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014.

126. Chitrao, Mahesh V and Grishman, Ralph. Statistical parsing of messages. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.
127. Cho, Kyunghyun; Van Merriënboer, Bart; Bahdanau, Dzmitry, and Bengio, Yoshua. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014a.
128. Cho, Kyunghyun; Van Merriënboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger, and Bengio, Yoshua. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014b.
129. Chowdhary, KR. *Fundamentals of Artificial Intelligence*. Springer, 2020.
130. Cifoletti, Guido. *La lingua franca barbaresca*. Il calamo, 2004.
131. Cohen, David. *Le parler arabe des juifs de Tunis: textes et documents linguistiques et ethnographiques*, volume 7. Mouton & Co, Paris, 1964.
132. Cohen, David. *Etudes de linguistique semitique et arabe (Studies of Semitic and Arabic Linguistics)*. ERIC, 1970.
133. Collobert, Ronan and Weston, Jason. Fast semantic extraction using a novel neural network architecture. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 560–567, 2007.
134. Collobert, Ronan and Weston, Jason. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, 2008.
135. Collobert, Ronan; Weston, Jason; Bottou, Léon; Karlen, Michael; Kavukcuoglu, Koray, and Kuksa, Pavel. Natural language processing (almost) from scratch. *Journal of machine learning research*, 120 (ARTICLE):0 2493–2537, 2011.
136. Corriente, Federico. From old Arabic to classical Arabic through the pre-islamic koine: Some notes on the native grammarians'sources, attitudes and goals. *Journal of Semitic studies*, 210 (1-2):0 62–98, 1976.
137. Cotterell, Ryan; Renduchintala, Adithya; Saphra, Naomi, and Callison-Burch, Chris. An algerian Arabic-french code-switched corpus. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme*, page 34, 2014.
138. Coulmas, Florian. Development of orthographies. *Literacy. An International Handbook*, pages 137–142, 1999.
139. Coulmas, Florian. *Writing and society: An introduction*. Cambridge University Press, 2013.

140. Crystal, David. Refining stylistic discourse categories. *English linguistics in honour of Magnus Ljung*, pages 35–46, 1994.
141. Crystal, David. *Language and the Internet*. Cambridge University Press, Cambridge, 2004.
142. Crystal, David. *A Dictionary of Linguistics and Phonetics (6th edn.)*. Blackwell Publishing, 2008.
143. D'Anna, Luca. The Arabic dialect of chebba. preliminary data and historical considerations. *Zeitschrift für Arabische Linguistik*, 0 (72):0 80–100, 2020.
144. Daoud, Mohamed. The language situation in Tunisia. *Current Issues in Language Planning*, 20 (1):0 1–52, 2001.
145. Daoud, Mohamed. The sociolinguistic situation in Tunisia: language rivalry or accommodation? *International journal of the sociology of language*, 20110 (211):0 9–33, 2011.
146. Darwish, Kareem. Arabizi detection and conversion to Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 217–224, 2014.
147. Darwish, Kareem; Sajjad, Hassan, and Mubarak, Hamdy. Verifiably effective Arabic dialect identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1465–1468, 2014.
148. Darwish, Kareem; Mubarak, Hamdy; Abdelali, Ahmed, and Eldesouki, Mohamed. Arabic pos tagging: Don't abandon feature engineering just yet. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 130–137, 2017.
149. Darwish, Kareem; Mubarak, Hamdy; Abdelali, Ahmed; Eldesouki, Mohamed; Samih, Younes; Alharbi, Randah; Attia, Mohammed; Magdy, Walid, and Kallmeyer, Laura. Multi-dialect Arabic pos tagging: A crf approach. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, 2018.
150. Darwish, Kareem; Habash, Nizar; Abbas, Mourad; Al-Khalifa, Hend; Al-Natsheh, Huseein T; Bouamor, Houda; Bouzoubaa, Karim; Cavalli-Sforza, Violetta; El-Beltagy, Samhaa R; El-Hajj, Wassim, and others, . A panoramic survey of natural language processing in the arab world. *Communications of the ACM*, 640 (4):0 72–81, 2021.
151. David, Chiang; Diab, Mona; Habash, Nizar; Rambow, Owen, and Shareef, Safiullah. Parsing Arabic dialects. In *Proceedings of EACL*, pages 369–376, 2006.
152. Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton, and Toutanova, Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

153. Diab, Mona; Habash, Nizar; Rambow, Owen; Altantawy, Mohamed, and Benajiba, Yassine. Colaba: Arabic dialect annotation and processing. In *Lrec workshop on semitic language processing*, pages 66–74. Citeseer, 2010.
154. Dinarelli, M. and Grobol, L. Seq2biseq: Bidirectional output-wise recurrent neural networks for sequence modelling. *CoRR*, abs/1904.04733, 2019a. URL `http://arxiv.org/abs/1904.04733`.
155. Dinarelli, M. and Grobol, L. Hybrid neural models for sequence modelling: The best of three worlds. *CoRR*, 2019b. URL `https://arxiv.org/abs/1909.07102`.
156. Dinarelli, Marco and Dupont, Yoann. Modélisation de dépendances entre étiquettes dans les réseaux neuronaux récurrents. *Traitement Automatique des Langues*, 580 (1), 2017.
157. Dinarelli, Marco and Grobol, Loïc. Seq2biseq: Bidirectional output-wise recurrent neural networks for sequence modelling. *arXiv preprint arXiv:1904.04733*, 2019c.
158. Dinarelli, Marco; Moschitti, Alessandro, and Riccardi, Giuseppe. Discriminative reranking for spoken language understanding. *IEEE transactions on audio, speech, and language processing*, 200 (2):0 526–539, 2011.
159. Duchet, Jean-Louis; Kraif, Olivier, and Torrellas Castillo, Manuel. Corpus massifs et corpus bilingues alignés: leur impact sur la recherche linguistique. *Bulletin de la Société de Linguistique de Paris*, 1030 (1):0 129–150, 2008.
160. Duh, Kevin and Kirchhoff, Katrin. Pos tagging of dialectal Arabic: a minimally supervised approach. In *Proceedings of the acl workshop on computational approaches to semitic languages*, pages 55–62, 2005.
161. Dupont, Yoann; Dinarelli, Marco, and Tellier, Isabelle. Réseaux neuronaux profonds pour l'étiquetage de séquences (deep neural networks for sequence labeling). In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2-Articles courts*, pages 19–27, 2017.
162. Durand, Olivier. L'arabo di tunisi. note di dialettologia comparata. *Dirāsāt Aryūliyya. Studi in onore di Angelo Arioli*, pages 241–272, 2007.
163. Durand, Olivier. *Dialettologia araba*. Carocci editore, Roma, 2009.
164. Durand, Olivier. Voyelles tunisoises. In Barontini, Alexandrine; Pereira, Christophe; Vicente, Ángeles, and Ziamari, Karima (ed.), *Dynamiques langagières en Arabophonies. Variations, contacts, migrations et créations artistiques. Hommage offert à Dominique Caubet*

*par ses élèves et ses collègues, Saragosse*, pages 65–76. Universidad de Zaragoza, Zaragoza, 2012.

165. Eid, Mushira. The non-randomness of diglossic variation in Arabic. *GLOSSA-AN INTERNATIONAL JOURNAL OF LINGUISTICS*, 160 (1):0 54–84, 1982.
166. Eid, Mushira. Principles for code-switching between standard and egyptian Arabic. *al-'Arabiyya*, pages 51–79, 1988.
167. Eksell Harning, Kerstin. *The analytic genitive in the modern Arabic dialects*. 1980.
168. El-Haj, Mahmoud. Habibi - a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1318–1326, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.165.
169. El-Haj, Mahmoud and Koulali, Rim. Kalimat a multipurpose Arabic corpus. In *Second workshop on Arabic corpus linguistics (WACL-2)*, pages 22–25, 2013.
170. El-Haj, Mahmoud; Rayson, Paul, and Aboelezz, Mariam. Arabic dialect identification in the context of bivalency and code-switching. In *Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan.*, pages 3622–3627. European Language Resources Association, 2018.
171. Elfardy, Heba and Diab, Mona. Aida: Automatic identification and glossing of dialectal Arabic. In *Proceedings of the 16th eamt conference (project papers)*, pages 83–83, 2012a.
172. Elfardy, Heba and Diab, Mona. Simplified guidelines for the creation of large scale dialectal Arabic annotations. In *Proceedings of the 8th Language Resources and Evaluation Conference (LREC)*, pages 371–378. Citeseer, 2012b.
173. Elfardy, Heba and Diab, Mona. Token level identification of linguistic code switching. In *Proceedings of COLING 2012: Posters*, pages 287–296, 2012c.
174. Elfardy, Heba and Diab, Mona. Sentence level dialect identification in Arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461, 2013.
175. Elfardy, Heba; Al-Badrashiny, Mohamed, and Diab, Mona. Code switch point detection in Arabic. In *International Conference on Application of Natural Language to Information Systems*, pages 412–416. Springer, 2013.
176. Elfardy, Heba; Al-Badrashiny, Mohamed, and Diab, Mona. Aida: Identifying code switching in informal Arabic text. In *Proceedings*

*of The First Workshop on Computational Approaches to Code Switching*, pages 94–101, 2014.

177. Embarki, Mohamed. Les dialectes arabes modernes: état et nouvelles perspectives pour la classification géo-sociologique. *Arabica*, 550 (5):0 583–604, 2008.
178. Erdmann, Alexander; Zalmout, Nasser, and Habash, Nizar. Addressing noise in multidialectal word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 558–565, 2018.
179. Eryani, Fadhl; Habash, Nizar; Bouamor, Houda, and Khalifa, Salam. A spelling correction corpus for multiple Arabic dialects. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4130–4138, 2020.
180. Eskander, Ramy and Rambow, Owen. Slsa: A sentiment lexicon for standard Arabic. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2545–2550, 2015.
181. Eskander, Ramy; Al-Badrashiny, Mohamed; Habash, Nizar, and Rambow, Owen. Foreign words and the automatic processing of Arabic social media text written in roman script. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 1–12, 2014.
182. Esuli, Andrea and Sebastiani, Fabrizio. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Language Resources and Evaluation Conference (LREC)*, volume 6, pages 417–422. Citeseer, 2006.
183. Farha, Ibrahim Abu and Magdy, Walid. Mazajak: An online Arabic sentiment analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 192–198, 2019.
184. Ferguson, Charles A. Diglossia. *word*, 150 (2):0 325–340, 1959.
185. Ferguson, Charles A. Epilogue: diglossia revisited. *Understanding Arabic, essays in contemporary Arabic linguistics in honor of El-Said Badawi*, pages 49–67, 1996.
186. Ferguson, Charles A. The Arabic koine. In *Structuralist Studies in Arabic Linguistics*, pages 50–68. Brill, 1997.
187. Field, Andy. *Discovering statistics using SPSS*. Sage publications, 2009.
188. Finnis, Katerina. Creating a 'new space': Code-switching among british-born greek-cypriots in london. *Pragmatics and Society*, 40 (2):0 137–157, 2013.
189. Fisher, Ronald A. On the interpretation of $\chi$ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 850 (1):0 87–94, 1922.

190. Fishman, Joshua A. Bilingualism with and without diglossia; diglossia with and without bilingualism. *Journal of social issues*, 230 (2):0 29–38, 1967.
191. Fishman, Joshua A. *Handbook of language and ethnic identity*. Oxford University Press, USA, 1999.
192. Fourati, Chayma; Messaoudi, Abir, and Haddad, Hatem. Tunizi: a Tunsian arabizi sentiment analysis dataset. *arXiv preprint arXiv:2004.14303*, 2020.
193. Gadalla, H; Kilany, H; Arram, H; Yacoub, A; El-Habashi, A; Shalaby, A, and McLemore, C. Callhome egyptian Arabic transcripts ldc97t19. *Web Download. Philadelphia: Linguistic Data Consortium*, 1997.
194. Gadet, François. Une distinction bien fragile: écrit/oral. *Revue Tranel (Travaux neuchâtelois de linguistique)*, 25:0 13–27, 1996.
195. Gadet, Françoise. *Ubi scripta et volant et manent*. Gunter Narr Verlag, 2008.
196. Garbini, Giovanni and Durand, Olivier. *Introduzione alle lingue semitiche*. Paideia, 1994.
197. Georgakopoulou, Alexandra. Computer-mediated communication. *Pragmatics in practice*, 9:0 93, 2011.
198. Georgakopoulou, Alexandra and Finnis, Katerina. Code-switching 'in site' for fantasizing identities: A case study of conventional uses of london greek cypriot. *Pragmatics*, 190 (3):0 467–488, 2009.
199. Ghoul, Hasna. Variations linguistiques dans le marquage du territoire dans la ville de tunis. *Synergies Tunisie*, 1:0 119–124, 2009.
200. Gibson, Maik. Dialect levelling in Tunisian Arabic: towards a new spoken standard. In Rouchdy, A. (ed.), *Language contact and language conflict in Arabic – variations on a sociolinguistic theme*, pages 24–40. Routledge, New York, 2002.
201. Gibson, Maik. Tunis Arabic. In Versteegh, Kees; Eid, Mushira, and Elgibali, Alaa (ed.), *Encyclopedia of Arabic Language and Linguistics*, page 563–571. Brill, Leiden, 2008.
202. Goodfellow, Ian; Bengio, Yoshua; Courville, Aaron, and Bengio, Yoshua. *Deep learning*, volume 1. MIT press Cambridge, 2016.
203. Gorgis, Dinha T. Transliterating Arabic: The nuisance of conversion between romanisation and transcription schemes. In *Paper presented at The International Symposium on Arabic Transliteration Standards: Challenges and Solutions, Abu Dhabi*, 2010.
204. Graja, Marwa; Jaoua, Maher, and Hadrich Belguith, L. Lexical study of a spoken dialogue corpus in Tunisian dialect. In *Proceedings of The International Arab Conference on Information Technology, benghazi–libya*. Citeseer, 2010.

205. Graja, Marwa; Jaoua, Maher, and Belguith, Lamia Hadrich. Building ontologies to understand spoken Tunsian dialect. *arXiv preprint arXiv:1109.0624*, 2011a.
206. Graja, Marwa; Jaoua, Maher, and Belguith, Lamia Hadrich. Towards understanding spoken Tunsian dialect. In *International Conference on Neural Information Processing*, pages 131–138. Springer, 2011b.
207. Graja, Marwa; Jaoua, Maher, and Belguith, Lamia Hadrich. Discriminative framework for spoken Tunisian dialect understanding. In *International Conference on Statistical Language and Speech Processing*, pages 102–110. Springer, 2013.
208. Graja, Marwa; Jaoua, Maher, and Belguith, L Hadrich. Statistical framework with knowledge base integration for robust speech understanding of the Tunsian dialect. *IEEE/ACM transactions on audio, speech, and language processing*, 230 (12):0 2311–2321, 2015.
209. Grand'Henry, Jacques. Eléments du système consonantique préhilâlien en arabe maghrébin: perspective historique. *Quaderni di Studi Arabi*, pages 93–98, 1992.
210. Granger, Sylviane; Kraif, Olivier; Ponton, Claude; Antoniadis, Georges, and Zampa, Virginie. Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *RECALL-HULL THEN CAMBRIDGE-*, 190 (3):0 252, 2007.
211. Grenoble, Lenore A. and Whaley, Lindsay J. *Saving languages: An introduction to language revitalization*. Cambridge University Press, 2005.
212. Grosz, Barbara J; Appelt, Douglas E; Martin, Paul A, and Pereira, Fernando CN. Team: An experiment in the design of transportable natural-language interfaces. *Artificial Intelligence*, 320 (2):0 173–243, 1987.
213. Guellil, Imane and Faical, Azouaou. Bilingual lexicon for algerian Arabic dialect treatment in social media. *WiNLP: Women & Underrepresented Minorities in Natural Language Processing (Co-located with ACL 2017)*, 2017.
214. Guellil, Imane; Azouaou, Faiçal, and Abbas, Mourad. Comparison between neural and statistical translation after transliteration of algerian Arabic dialect. *WiNLP: women & underrepresented minorities in natural language processing (co-located with ACL 2017)*, pages 1–5, 2017a.
215. Guellil, Imane; Azouaou, Faical, and Abbas, Mourad. Neural vs statistical translation of algerian Arabic dialect written with arabizi and Arabic letter. In *The 31st pacific asia conference on language, information and computation paclic*, volume 31, page 2017, 2017b.

216. Guellil, Imane; Adeel, Ahsan; Azouaou, Faical, and Hussain, Amir. Sentialg: Automated corpus annotation for algerian sentiment analysis. In *International Conference on Brain Inspired Cognitive Systems*, pages 557–567. Springer, 2018.
217. Guellil, Imane; Saâdane, Houda; Azouaou, Faical; Gueni, Billel, and Nouvel, Damien. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 2019.
218. Guellil, Imène and Azouaou, Faiçal. Arabic dialect identification with an unsupervised learning (based on a lexicon). application case: Algerian dialect. In *2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES)*, pages 724–731. IEEE, 2016.
219. Guellil, Imène and Azouaou, Faiçal. Asda: Analyseur syntaxique du dialecte alg é rien dans un but d'analyse s é mantique. *arXiv preprint arXiv:1707.08998*, 2017.
220. Guerrero, Jairo. The jbala-villageois dialects. a grammatical study of a rural typology of colloquial maghrebi Arabic. *DIALECTOLOGIA*, 0 (20):0 85–105, 2018.
221. Gugliotta, Elisa and Dinarelli, Marco. Tarc. Un corpus d'arabish tunisien. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2: Traitement Automatique des Langues Naturelles*, pages 232–240, 2020a.
222. Gugliotta, Elisa and Dinarelli, Marco. Tarc: Incrementally and semi-automatically collecting a Tunsian arabish corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6279–6286, 2020b.
223. Gugliotta, Elisa; Dinarelli, Marco, and Kraif, Olivier. Multi-task sequence prediction for Tunsian arabizi multi-level annotation. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 178–191, 2020.
224. Gugliotta, Elisa; Massaro, Angelapia; Mion, Giuliano, and Dinarelli, Marco. Definiteness in Tunisian arabizi: Some data from statistical approaches. forthcoming.
225. Habash, Nizar and Rambow, Owen. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pages 573–580, 2005.

226. Habash, Nizar and Rambow, Owen. Magead: A morphological analyzer and generator for the Arabic dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 681–688, 2006.
227. Habash, Nizar and Roth, Ryan. Catib: The columbia Arabic treebank. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 221–224, 2009.
228. Habash, Nizar; Rambow, Owen; Diab, Mona, and Kanjawi-Faraj, Reem. Guidelines for annotation of Arabic dialectness. In *Proceedings of the LREC Workshop on HLT & NLP within the Arabic world*, pages 49–53, 2008.
229. Habash, Nizar; Rambow, Owen, and Roth, Ryan. Mada+ tokan: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt*, volume 41, page 62, 2009.
230. Habash, Nizar; Diab, Mona, and Rambow, Owen. Conventional orthography for dialectal Arabic. In *Proceedings of the 8th Language Resources and Evaluation Conference (Proceedings of the 8th Language Resources and Evaluation Conference (LREC))*, pages 711–718, 2012a.
231. Habash, Nizar; Eskander, Ramy, and Hawwari, Abdelati. A morphological analyzer for egyptian Arabic. In *Proceedings of the twelfth meeting of the special interest group on computational morphology and phonology*, pages 1–9, 2012b.
232. Habash, Nizar; Mohit, Behrang; Obeid, Ossama; Oflazer, Kemal; Tomeh, Nadi, and Zaghouani, Wajdi. Qalb: Qatar Arabic language bank. In *Proceedings of Qatar Annual Research Conference (ARC-2013), pages ICTP–032, Doha, Qatar*, 2013.
233. Habash, Nizar; Eryani, Fadhl; Khalifa, Salam; Rambow, Owen; Abdulrahim, Dana; Erdmann, Alexander; Faraj, Reem; Zaghouani, Wajdi; Bouamor, Houda; Zalmout, Nasser, and others, . Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, 2018.
234. Habash, Nizar Y. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 30 (1):0 1–187, 2010.
235. Hamdi, Ahmed; Gala, Nuria, and Nasr, Alexis. Automatically building a Tunsian lexicon for deverbal nouns. In *Proceedings of the first workshop on applying NLP tools to similar languages, Varieties and Dialects*, pages 95–102, 2014.

236. Hamdi, Ahmed; Nasr, Alexis; Habash, Nizar, and Gala, Núria. Pos-tagging of Tunsian dialect using standard Arabic resources and tools. In *Workshop on Arabic Natural Language Processing*, pages 59–68, 2015.
237. Harrat, Salima; Meftouh, Karima; Abbas, Mourad, and Smaili, Kamel. Building resources for algerian Arabic dialects. In *15th Annual Conference of the International Communication Association Interspeech*, 2014.
238. Harrat, Salima; Meftouh, Karima, and Smaïli, Kamel. Creating parallel Arabic dialect corpus: pitfalls to avoid. In *18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, 2017.
239. Hary, Benjamin. Middle Arabic: Proposals for new terminology. *al-'Arabiyya*, pages 19–36, 1989.
240. Hary, Benjamin. The importance of the language continuum in Arabic multiglossia. In Elgibali, Alaa (ed.), *Understanding Arabic: Essays in Contemporary Arabic Linguistics in Honor of El-Saif Badawi*, pages 69–90. American University in Cairo Press, Cairo, 1996.
241. Hashimoto, Kazuma; Xiong, Caiming; Tsuruoka, Yoshimasa, and Socher, Richard. A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv preprint arXiv:1611.01587*, 2016.
242. Haspelmath, Martin and Tadmor, Uri. The loanword typology meaning list: Electronic databases of 29 languages. *A collaborative project coordinated by the Max Planck Institute for Evolutionary Anthropology, Department of Linguistics*, 2009.
243. Hassine, Mohamed; Boussaid, Lotfi, and Messaoud, Hassani. Maghrebian dialect recognition based on support vector machines and neural network classifiers. *International Journal of Speech Technology*, 190 (4):0 687–695, 2016.
244. Hassine, Mohamed; Boussaid, Lotfi, and Massaoud, Hassani. Tunisian dialect recognition based on hybrid techniques. *Int. Arab J. Inf. Technol.*, 150 (1):0 58–65, 2018.
245. Heath, Maria. Orthography in social media: Pragmatic and prosodic interpretations of caps lock. *Proceedings of the Linguistic Society of America*, 30 (1):0 55–1, 2018.
246. Herring, Susan C. *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*, volume 39. John Benjamins Publishing, 1996.
247. Herring, Susan C. A faceted classification scheme for computer-mediated discourse. *Language@internet*, 40 (1), 2007.
248. Herring, Susan C. and Stoerger, Sharon. Gender and (a)nonymity in computer-mediated communication. *The handbook of language, gender, and sexuality*, page 567, 2014.

249. Hert, Philippe. Quasi-oralité de l'écriture électronique et sentiment de communauté dans les débats scientifiques en ligne. *Réseaux*, 170 (97), 1999.
250. Hirst, Graeme. A semantic process for syntactic disambiguation. In *AAAI*, pages 148–152, 1984.
251. Hirst, Graeme. *Semantic interpretation and the resolution of ambiguity*. Cambridge University Press, 1987.
252. Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural computation*, 90 (8):0 1735–1780, 1997a.
253. Hochreiter, Sepp and Schmidhuber, Jürgen. Lstm can solve hard long time lag problems. *Advances in neural information processing systems*, pages 473–479, 1997b.
254. Holes, Clive. Community, dialect and urbanization in the Arabic-speaking middle east. *Bulletin of the School of Oriental and African Studies, University of London*, pages 270–287, 1995.
255. Hovy, Dirk and Spruit, Shannon L. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, 2016.
256. Hymes, Dell. *Foundations in sociolinguistics: An ethnographic approach*. University of Pennsylvania Press, Philadelphia, 1974.
257. Inoue, Go; Shindo, Hiroyuki, and Matsumoto, Yuji. Joint prediction of morphosyntactic categories for fine-grained Arabic part-of-speech tagging exploiting tag dictionary information. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 421–431, 2017.
258. Inoue, Go; Habash, Nizar; Matsumoto, Yuji, and Aoyama, Hiroyuki. A parallel corpus of Arabic-japanese news articles. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, 2018.
259. Iskra, Dorota J; Siemund, Rainer; Borno, Jamal; Moreno, Asuncion; Emam, Ossama; Choukri, Khalid; Gedge, Oren; Tropf, Herbert S; Nogueiras, Albino; Zitouni, Imed, and others, . OrienTel-Telephony Databases Across Northern Africa and the Middle East. In *Proceedings of the 4th Language Resources and Evaluation Conference (LREC)*. Citeseer, 2004.
260. Ivković, Dejan. Cyber-latinica: A comparative analysis of latinization in internet slavic. *Language@ Internet*, 120 (2), 2015.
261. Jaffe, Alexandra; Androutsopoulos, Jannis; Sebba, Mark, and Johnson, Sally. *Orthography as social action: Scripts, spelling, identity and power*, volume 3. Walter de Gruyter, 2012.
262. Jarrar, Mustafa; Habash, Nizar; Alrimawi, Faeq; Akra, Diyam, and Zalmout, Nasser. Curras: an annotated corpus for the palestinian

Arabic dialect. *Language Resources and Evaluation*, 510 (3):0 745–775, 2017.

263. Jouini, Kemel. Dependency relations in the syntactic structure of Tunisian Arabic. *Arab World English Journal*, 30 (4):0 36–57, 2012.
264. Kallel, Myriam Achour. Choix langagiers sur la radio mosaïque fm. dispositifs d'invisibilité et de normalisation sociales. *Langage et société*, 4:0 77–96, 2011.
265. Karmani, Ben Moussa Nadia and Alimi, Adel M. Construction d'un wordnet standard pour l'arabe tunisien. In *Proceedings of Colloque pour les Étudiants Chercheurs en Traitement Automatique du Langage naturel et ses applications, Sousse, Tunisia*, 2015.
266. Karmani, Nadia BM; Soussou, Hsan, and Alimi, Adel M. Building a standardized wordnet in the iso lmf for aeb language. In *Proceedings of the Seventh Global Wordnet Conference*, pages 71–77, 2014.
267. Karoui, Jihen; Graja, Marwa; Boudabous, Mohamed Mahdi, and Belguith, Lamia Ha-Drich. Domain ontology construction from a Tunsian spoken dialogue corpus. In *International Conference on Web and Information Technologies*, 2013a.
268. Karoui, Jihen; Graja, Marwa; Boudabous, Mohamed Mahdi, and Belguith, Lamia Hadrich. Semi-automatic domain ontology construction from spoken corpus in Tunsian dialect: Railway request information. *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, 10 (1), 2013b.
269. Khalifa, Salam; Habash, Nizar; Abdulrahim, Dana, and Hassan, Sara. A large scale corpus of Gulf Arabic. *Proceedings of the 10th Language Resources and Evaluation Conference (LREC)*, 2016a.
270. Khalifa, Salam; Zalmout, Nasser, and Habash, Nizar. Yamama: Yet another multi-dialect Arabic morphological analyzer. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 223–227, 2016b.
271. Khalifa, Salam; Hassan, Sara, and Habash, Nizar. A morphological analyzer for gulf Arabic verbs. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 35–45, 2017.
272. Khalifa, Salam; Habash, Nizar; Eryani, Fadhl; Obeid, Ossama; Abdulrahim, Dana, and Al Kaabi, Meera. A morphologically annotated corpus of emirati Arabic. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, 2018.
273. Khalifa, Salam; Zalmout, Nasser, and Habash, Nizar. Morphological analysis and disambiguation for gulf Arabic: The interplay between resources and methods. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3895–3904, 2020.
274. Kraif, Olivier and Tutin, Agnès. Collocations et linguistique de corpus: l'intuition des linguistes et les critères quantitatifs

convergent-ils? *Le Français Moderne-Revue de linguistique Française*, 1:0 p–84, 2020.

275. Kumar, Gaurav; Cao, Yuan; Cotterell, Ryan; Callison-Burch, Chris; Povey, Daniel, and Khudanpur, Sanjeev. Translations of the callhome egyptian Arabic corpus for conversational speech translation. In *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*. Citeseer, 2014.
276. Kwaik, Kathrein Abu; Saad, Motaz; Chatzikyriakidis, Stergios, and Dobnik, Simon. Shami: A corpus of levantine Arabic dialects. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, 2018.
277. La Rosa, Cristina. Mahdia dialect: An urban vernacular in the Tunisian sahel context. *Languages*, 60 (3):0 145, 2021.
278. Labov, William. *Sociolinguistic Patterns*, volume 10. University of Pennsylvania Press, 1972.
279. Lachachi, Nour-Eddine and Adla, Abdelkader. Gmm-based maghreb dialect identification system. *JIPS (Journal of Information Processing Systems)*, 110 (1):0 22–38, 2015.
280. Lachachi, Nour-Eddine and Adla, Abdelkader. Identification automatique des dialectes du maghreb. *Revue Maghrébine des Langues*, 100 (1):0 85–101, 2016a.
281. Lachachi, Nour-Eddine and Adla, Abdelkader. Two approaches-based l2-svms reduced to meb problems for dialect identification. *International Journal of Computational Vision and Robotics*, 60 (1-2):0 1–18, 2016b.
282. Lacquaniti, Luce. *I muri di Tunisi: segni di rivolta*. Exòrma, Rome, 2015.
283. Lafferty, John D.; McCallum, Andrew, and Pereira, Fernando C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.
284. Lai, Rosangela. Language planning and language policy in sardinia. *Language Problems and Language Planning*, 420 (1):0 70–88, 2018.
285. Lan, Wuwei; Chen, Yang; Xu, Wei, and Ritter, Alan. Gigabert: Zero-shot transfer learning from english to Arabic. In *Proceedings of The 2020 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2020.
286. Lancioni, Giuliano. Ordini lineari marcati in arabo un'analisi generativa. *Rivista degli studi orientali*, 69:0 1–154, 1996.
287. Lancioni, Giuliano. Encodings, genres, texts: Issues in Arabic corpus linguistics. In Bohas, Georges and Kouloughli, Djamel

Eddine (ed.), *Langues et Littératures du Monde Arabe*, volume 9, pages 84–93, 2011.

288. Lancioni, Giuliano. Insegnamento dell'arabo e certificazione: una panoramica. In Lancioni, G. and Solimando, C. (ed.), *Didattica dell'arabo e certificazione linguistica: riflessioni e iniziative*. RomaTre-Press, Roma, 2018.
289. Langone, Angela Daiana and Mion, Giuliano. Le journal de la médina. un récent projet éditorial en arabe tunisien. In Catherine, Miller; Alexandrine, Barontini; Marie-Aimée, Germanos; Jairo, Guerrero, and Pereira, Christophe (ed.), *Studies on Arabic Dialectology and Sociolinguistics Proceedings of the 12th International Conference of AIDA held in Marseille from 30th May- 2nd June 2017*. Institut de recherches et études sur les mondes arabes et musulmans, 2019.
290. Laroussi, Foued. Les politiques linguistiques des pays maghrébins un essai d'évaluation. *Iles d imesli*, 2:0 183–196, 2010.
291. Leech, Geoffrey. Chapter 2. adding linguistic annotation. In Wynne, Martin (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, pages 26–39. Oxbow Books for the Arts and Humanities Data Service, Oxford, 2004.
292. Leech, Geoffrey; Garside, R; McEnery, Tony, and others, . Corpus annotation: Linguistic information from computer text corpora. *chapter Introducing corpus annotation*, pages 1–18, 1997.
293. Léglise, Isabelle; Caubet, Dominique; Bulot, Thierry; Billiez, Jacqueline, and Miller, Catherine. *Parlers jeunes ici et là-bas*. L'Harmattan, 2004.
294. Lei, Yun and Hansen, John HL. Dialect classification via text-independent training and testing for Arabic, spanish, and chinese. *IEEE Transactions on Audio, Speech, and Language Processing*, 190 (1):0 85–96, 2010.
295. Lesk, Michael. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, 1986.
296. Lewis, Mary Dewhurst. *Divided Rule: Sovereignty and Empire in French Tunisia, 1881Ð1938*. Univ of California Press, 2013.
297. Liberman, Mark Y. The trend towards statistical models in natural language processing. 1991.
298. Lison, Pierre and Tiedemann, Jörg. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In Chair), Nicoletta Calzolari (Conference; Choukri, Khalid; Declerck, Thierry; Goggi, Sara; Grobelnik, Marko; Maegaard, Bente;

Mariani, Joseph; Mazo, Helene; Moreno, Asuncion; Odijk, Jan, and Piperidis, Stelios (ed.), *Proceedings of the 10th Language Resources and Evaluation Conference (LREC)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

299. Loprieno, Antonio. *Ancient Egyptian: a linguistic introduction*. Cambridge University Press, 1995.
300. Maamouri, Mohamed; Bies, Ann; Buckwalter, Tim, and Mekki, Wigdan. The Penn Arabic Treebank: Building a large-scale annotated Arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467. Cairo, 2004.
301. Maamouri, Mohamed; Buckwalter, Tim; Graff, Dave, and Jin, Hubert. Fisher levantine Arabic conversational telephone speech. *Linguistic Data Consortium, University of Pennsylvania, LDC Catalog No.: LDC2007S02*, 2007.
302. Maamouri, Mohamed; Bies, Ann; Kulick, Seth; Ciul, Michael; Habash, Nizar, and Eskander, Ramy. Developing an egyptian Arabic treebank: Impact of dialectal morphology on annotation and tool development. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, pages 2348–2354, 2014.
303. Marçais, William. *Le diglossie arabe*. librairie Delagrave, 1930.
304. Marçais, William. Comment l'afrique du nord a été arabisée. In *Articles et conférences*. Adrien-Maisonneuve, Paris, 1961.
305. Marcoccia, Michel. *Analyser la communication numérique écrite*. Armand Colin, 2016.
306. Marçais, Philippe. Les parlers arabes. In Alazard, Jean and al., et (ed.), *Initiation à l'Algérie*, pages 215–237. Adrien-Maisonneuve, Paris, 1957.
307. Marçais, William. D. Le parlers Arabes et Berbéres, I. Les parles arabes. In Basset, A. (ed.), *Initiation à la Tunisie*, pages 195–219. Adrien-Maisonneuve, Paris, 1950.
308. Masmoudi, Abir; Estève, Yannick; Khmekhem, Mariem Ellouze; Bougares, Fethi, and Belguith, Lamia Hadrich. Phonetic tool for the Tunsian Arabic. In *Spoken Language Technologies for Under-Resourced Languages*, 2014a.
309. Masmoudi, Abir; Khmekhem, Mariem Ellouze; Esteve, Yannick; Belguith, Lamia Hadrich, and Habash, Nizar. A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, pages 306–310, 2014b.
310. Masmoudi, Abir; Habash, Nizar; Ellouze, Mariem; Estève, Yannick, and Belguith, Lamia Hadrich. Arabic transliteration of romanized Tunisian dialect text: A preliminary investigation. In

*International Conference on Intelligent Text Processing and Computational Linguistics*, pages 608–619. Springer, 2015.

311. Masmoudi, Abir; Ellouze, Mariem; Bougares, Fethi; Estève, Yannick, and Belguith, Lamia Hadrich. Conditional random fields for the Tunsian dialect grapheme-to-phoneme conversion. In *INTERSPEECH*, pages 1457–1461, 2016.
312. Masmoudi, Abir; Bougares, Fethi; Ellouze, Mariem; Estève, Yannick, and Belguith, Lamia. Automatic speech recognition system for Tunsian dialect. *Language Resources and Evaluation*, 520 (1):0 249–267, 2018.
313. McNeil, Karen. Tunisian Arabic morphological parser. *Ling-420*, 11, 2012.
314. McNeil, Karen. Tunisian Arabic corpus: Creating a written corpus of an 'unwritten' language. In *International Symposium on Tunsian and Libyan Arabic Dialects*. University of Vienna, 2015.
315. McNeil, Karen and Faiza, Miled. Tunisian Arabic corpus: Creating a written corpus of an 'unwritten' language. In *Proceeding of the Workshop on Arabic Corpus Linguistics*. Lancaster University, UK, 2011.
316. Mdhaffar, Salima; Bougares, Fethi; Esteve, Yannick, and Hadrich-Belguith, Lamia. Sentiment analysis of Tunisian dialects: Linguistic ressources and experiments. In *Third Arabic Natural Language Processing Workshop (WANLP)*, pages 55–61, 2017.
317. Meftah, Sara; Semmar, Nasredine; Tahiri, Mohamed-Ayoub; Tamaazousti, Youssef; Essafi, Hassane, and Sadat, Fatiha. Multi-task supervised pretraining for neural domain adaptation. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 61–71, 2020.
318. Meftouh, Karima; Harrat, Salima; Jamoussi, Salma; Abbas, Mourad, and Smaili, Kamel. Machine translation experiments on padic: A parallel Arabic dialect corpus. In *The 29th Pacific Asia conference on language, information and computation*, 2015.
319. Meftouh, Karima; Harrat, Salima, and Smaïli, Kamel. Padic: extension and new experiments. In *7th International Conference on Advanced Technologies ICAT*, 2018.
320. Meiseles, Gustav. Educated spoken Arabic and the Arabic language continuum. *Archivum linguisticum*, 110 (2):0 117–148, 1980.
321. Mejdell, Gunvor. Diglossia. In Benmamoun, E. and Bassiouney, R. (ed.), *The Routledge Handbook of Arabic Linguistics*, pages 332–344. Routledge, NY, 2017.
322. Mejri, Salah; Said, Mosbah, and Sfar, Inès. Pluringuisme et diglossie en tunisie. *Synergies Tunisie*, 1:0 53–74, 2009.

323. Mekki, Asma; Zribi, Inès; Ellouze, Mariem, and Belguith, Lamia Hadrich. Syntactic Analysis of the Tunisian Arabic. In *LPKM*, 2017.
324. Metcalfe, Alexander. *Muslims and Christians in Norman Sicily: Arabic-Speakers and the End of Islam*. Routledge, 2014.
325. Mikolov, Tomáš. Language modeling for speech recognition in czech. Master's thesis, Brno University of Technology, Czech Republic, 2007.
326. Mikolov, Tomas; Kopecky, Jiri; Burget, Lukas; Glembek, Ondrej, and others, . Neural network based language models for highly inflective languages. In *2009 IEEE international conference on acoustics, speech and signal processing*, pages 4725–4728. IEEE, 2009.
327. Mikolov, Tomáš; Karafiát, Martin; Burget, Lukáš; Černockỳ, Jan, and Khudanpur, Sanjeev. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*, 2010.
328. Mikolov, Tomas; Chen, Kai; Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
329. Miller, Catherine. Variation and changes in Arabic urban vernaculars. In Haak, M.; Versteegh, K., and Dejong, R. (ed.), *Approaches to Arabic Dialects : Collection of Articles presented to Manfred Woidich on the Occasion of his Sixtieth Birthday*. Brill, Amsterdam, 2004.
330. Miller, Catherine. Langues et médias dans le monde arabe/arabophone. entre idéologie et marché, convergences dans la glocalisation. In Lachkar, A. (ed.), *Langues et médias en méditerranée*, pages 157–171. L'Harmattan, Langue et parole, Paris, 2012.
331. Miller, Catherine. Contemporary dārija writings in morocco: ideology and practices. In *The politics of written language in the Arab world*, pages 90–115. Brill, 2017.
332. Miller, Gerald R. On defining communication: another stab. *Journal of Communication*, 1966.
333. Mimouna, Zitouni. Is english there?: Investigating language use among young algerian users of internet. *Unpublished Doctorate thesis in Sociolinguistics*, 2013.
334. Minsky, Marvin L and Papert, Seymour A. Perceptrons: an introduction to computational geometry, 1969.
335. Mion, Giuliano. La versione del piccolo principe in arabo tunisino. *Ricerca e didattica tra due sponde, atti della Convenzione Internazionale tra l'Università 'G. d'Annunzio di Chieti-Pescara' e l'Université '7 Novembre à Carthage' di Tunisi*, 2007.

336. Mion, Giuliano. Le patah furtivum en sémitique. remarques de phonétique et phonologie. 2008a.
337. Mion, Giuliano. Le vocalisme et l'imāla en arabe tunisien. *Between the Atlantic and Indian Oceans. Studies on Contemporary Arabic Dialects. Proceedings of the 7th AIDA Conference*, 2008b.
338. Mion, Giuliano. *Sociofonologia dell'arabo. Dalla ricerca empirica al riconoscimento del parlante*. Edizioni Nuova Cultura, Roma, 2010.
339. Mion, Giuliano. Réflexions sur la catégorie des «parlers villageois» en arabe tunisien. *Romano Arabica XV. Bucharest, Editura Universităţii din Bucureşti*, pages 269–279, 2015.
340. Mion, Giuliano. *La lingua araba. Nuova edizione*. Carocci, Roma, 2016.
341. Mion, Giuliano. Un arabo cipriota romanizzato? distacchi e identità fra variazione scrittoria e confessione religiosa. In C, Cosani (ed.), *Aspetti della variazione linguistica. Discorso, sistema, repertori*. LED Edizioni, Milano, 2017a.
342. Mion, Giuliano. 18. cypriot Arabic between orality and literacy in o typos ton maroniton. In *Language and Identity in Multilingual Mediterranean Settings*, pages 325–340. De Gruyter Mouton, 2017b.
343. Mion, Giuliano. Pré-hilalien, hilalien, zones de transition. relire quelques classiques aujourd'hui. *Mediterranean Contaminations. Middle East, North Africa, and Europe in Contact*, pages 102–125, 2018.
344. Mion, Giuliano. Pré-hilalien, hilalien, zones de transition.nrelire quelques classiques aujourd'hui. In Mion, Giuliano (ed.), *Mediterranean Contaminations: Middle East, North Africa, and Europe in Contact*, volume 31. Walter de Gruyter, Berlin, 2020.
345. Mion, Giuliano and D'Anna, Luca. *Grammatica di arabo standard moderno: Fonologia, morfologia e sintassi*. HOEPLI EDITORE, 2021.
346. Mion, Giuliano and others, . Éléments de description de l'arabe parlé à mateur (tunisie). 2014.
347. Mion, Gluliano. Osservazioni sul sistema verbale dell'arabo di tunisi. *Rivista degli studi orientali*, 780 (Fasc. 1/2):0 243–255, 2004.
348. Mitchell, Terence F. Educated spoken Arabic in egypt and the levant, with special reference to participle and tense. *Journal of Linguistics*, 140 (2):0 227–258, 1978.
349. Mitchell, Terence F. What is educated spoken Arabic? *International Journal of the sociology of language*, 610 (1):0 7–32, 1986.
350. Monroe, Will; Green, Spence, and Manning, Christopher D. Word segmentation of informal Arabic with domain adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 206–211, 2014.

351. Morel, Etienne and Doehler, Simona Pekarek. Les 'textos' plurilingues: l'alternance codique comme ressource d'affiliation à une communauté globalisée. *Revue française de linguistique appliquée*, 180 (2):0 29–43, 2013.
352. Mori, Laura. La sociolinguistica dei corpora per lo studio della lingua inclusiva di genere nelle varietà legislative dell'eurolect observatory multilingual corpus (francese, inglese, italiano, spagnolo, tedesco). *Cavagnoli, Stefania & Mori, Laura (a cura di), Gender in legislative languages. From EU to national law in English, French, German, Italian and Spanish*, pages 39–65, 2019.
353. Moussa, Nadia Karmani Ben; Soussou, Hsan, and Alimi, Adel Mohamed. Tunisian Arabic aeb wordnet: Current state and future extensions. In *2015 First International Conference on Arabic Computational Linguistics (ACLing)*, pages 3–8. IEEE, 2015.
354. Mubarak, Hamdy and Darwish, Kareem. Using twitter to collect a multi-dialectal corpus of Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 1–7, 2014.
355. Murray, Denise E. When the medium determines turns: Turntaking in computer conversation. *Working with language: A multidisciplinary consideration of language use in work contexts*, 319:0 37, 1989.
356. Myers-Scotton, Carol. Common and uncommon ground: Social and structural factors in codeswitching. *Language in society*, pages 475–503, 1993.
357. Myers-Scotton, Carol. Carol myers-scotton: Multiple voices: An introduction to bilingualism., 2006.
358. Myers-Scotton, Carol and Jake, Janice L. Testing the 4-m model: An introduction. *International journal of bilingualism*, 40 (1):0 1–8, 2000.
359. Nafa, Hanan Omar A Ben. Code-switching and social identity construction among Arabic-english bilinguals: A stance perspective. In *Papers from the 10th Lancaster University Postgraduate Conference in Linguistics & Language Teaching 2015*, page 1, 2015.
360. Neifar, W; Bahou, Y; Graja, M, and Jaoua, M. Implementation of a symbolic method for the Tunsian dialect understanding. In *Proceedings of 5th International Conference on Arabic Language Processing (CITALA 2014). Oujda, Maroc*, 2014.
361. Obeid, Ossama; Khalifa, Salam; Habash, Nizar; Bouamor, Houda; Zaghouani, Wajdi, and Oflazer, Kemal. Madari: A web interface for joint Arabic morphological annotation and spelling correction. *arXiv preprint arXiv:1808.08392*, 2018.
362. Obeid, Ossama; Zalmout, Nasser; Khalifa, Salam; Taji, Dima; Oudah, Mai; Alhafni, Bashar; Inoue, Go; Eryani, Fadhl;

Erdmann, Alexander, and Habash, Nizar. Camel tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th language resources and evaluation conference*, pages 7022–7032, 2020.

363. Oberlander, Jon. Cognitive science: Overview. In Brown, Keith (ed.), *Encyclopedia of Language and Linguistics*, volume 1, pages 562–568. Elsevier Science, 2005.
364. Ong, Walter J. *Oralità e scrittura (Original Title: Orality and Literacy. The Technologizing of The Word)*. Bologna, Il Mulino, 1986.
365. Ouerhani, Béchir. La morphologie verbale du dialecte tunisien: Repères méthodologiques pour un traitement systématique. *Salah Mejri (coordonator)*, pages 333–344, 2006.
366. Palfreyman, David and Khalil, Muhamed al. "a funky language for teenzz to use:" representing gulf Arabic in instant messaging. *Journal of computer-mediated communication*, 90 (1):0 JCMC917, 2003.
367. Panckhurst, Rachel. Le discours électronique médié: bilan et perspectives., 2006.
368. Panckhurst, Rachel. Short message service (sms): typologie et problématiques futures., 2009.
369. Parker, Robert; Graff, David; Chen, Ke; Kong, Junbo, and Maeda, Kazuaki. Arabic gigaword fifth edition ldc2011t11. *Philadelphia: Linguistic Data Consortium*, 2011.
370. Pasha, Arfath; Al-Badrashiny, Mohamed; Diab, Mona; El Kholy, Ahmed; Eskander, Ramy; Habash, Nizar; Pooleery, Manoj; Rambow, Owen, and Roth, Ryan. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, volume 14, pages 1094–1101. Citeseer, 2014.
371. Pasquandrea, Sergio. *Code-switching e identità: pratiche discorsive di famiglie italiane in Paesi anglofoni*. PhD thesis, 2007.
372. Pearson, Karl. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 500 (302):0 157–175, 1900.
373. Pennacchietti, Fabrizio A. Studi sui pronomi determinativi semitici. *Pubblicazioni del Seminario di Semitistica/Ricerche*, 1968.
374. Pennacchietti, Fabrizio A. Ripercussioni sintattiche in conseguenza de'llintroduzione de'll articolo determinativo proclitico in semitico. *Aula orientalis: revista de estudios del Próximo*

*Oriente Antiguo*, 230 (1):0 175–184, 2005.

375. Pennington, Jeffrey; Socher, Richard, and Manning, Christopher D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
376. Picard, Jean. Le passé et ses monuments, A. La Tunisie antique. In Basset, A. (ed.), *Initiation à la Tunisie*, pages 33–72. Adrien-Maisonneuve, Paris, 1950.
377. Poplack, Shana and Meechan, Marjory. Introduction: How languages fit together in codemixing. *International journal of bilingualism*, 20 (2):0 127–138, 1998.
378. Psichari, J. (1886). Essais de grammaire historique néo-grecque (Vol. 1). Ernest Leroux.
379. Qi, Peng; Zhang, Yuhao; Zhang, Yuhui; Bolton, Jason, and Manning, Christopher D. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*, 2020.
380. Regis, Riccardo. *Appunti grammaticali sull'enunciazione mistilingue*. Lincom Europa, 2005.
381. Regis, Riccardo. ibridismi. 2010.
382. Riahi, Zohra. Emploi de l'arabe et du français par les élèves du secondaire. *Cahiers du CERES*, 3:0 99–165, 1970.
383. Riegert, Kristina and Ramsay, Gail. Activists, individualists, and comics: The counter-publicness of lebanese blogs. *Television & New Media*, 140 (4):0 286–303, 2013. doi: 10.1177/1527476412463447. URL `https://doi.org/10.1177/1527476412463447`.
384. Riesbeck, Christopher K. From conceptual analyzer to direct memory access parsing: An overview. *Advances in cognitive science*, 1:0 236–258, 1986.
385. Ritt-Benmimoun, Veronika. The Tunsian hilāl and sulaym dialects. a preliminary comparative study. In *Alf lahǧa wa lahǧa, Proceedings of the 9th Aida Conference*, pages 351–360. Lit Verlag, Berlin, 2014.
386. Rush, Alexander M; Reichart, Roi; Collins, Michael, and Globerson, Amir. Improved parsing and pos tagging using inter-sentence consistency constraints. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1434–1444, 2012.
387. Rushdi-Saleh, Mohammed; Martín-Valdivia, M Teresa; Ureña-López, L Alfonso, and Perea-Ortega, José M. Oca: Opinion corpus for Arabic. *Journal of the American Society for Information Science and Technology*, 620 (10):0 2045–2054, 2011.

388. Russel, Stuart; Norvig, Peter, and others, . *Artificial intelligence: a modern approach*. Pearson Education Limited London, 2009.
389. Saad, Motaz and Alijla, Basem O. Wikidocsaligner: An off-the-shelf wikipedia documents alignment tool. In *2017 Palestinian International Conference on Information and Communication Technology (PICICT)*, pages 34–39. IEEE, 2017.
390. Saada, Lucienne. *Eléments de description du parler arabe de Tozeur, Tunisie: phonologie, morphologie, syntaxe*, volume 2. Librairie Orientaliste Paul Geuthner, 1984.
391. Saadane, Houda and Habash, Nizar. A conventional orthography for algerian Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 69–79, 2015.
392. Saadane, Houda; Nouvel, Damien; Seffih, Hosni, and Fluhr, Christian. Une approche linguistique pour la détection des dialectes arabes. In *2017-06-26*, 2017.
393. Saadane, Houda; Seffih, Hosni; Fluhr, Christian; Choukri, Khalid, and Semmar, Nasredine. Automatic identification of maghreb dialects using a dictionary-based approach. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, 2018.
394. Sadat, Fatiha; Kazemi, Farnazeh, and Farzindar, Atefeh. Automatic identification of Arabic dialects in social media. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media*, pages 22–27. Dublin, Ireland, 2014a.
395. Sadat, Fatiha; Mallek, Fatma; Boudabous, Mohamed Mahdi; Sellami, Rahma, and Farzindar, Atefeh. Collaboratively constructed linguistic resources for language variants and their exploitation in nlp application–the case of Tunsian Arabic and the social media. In *Proceedings of workshop on Lexical and grammatical resources for language processing*, pages 102–110, 2014b.
396. Safaya, Ali; Abdullatif, Moutasem, and Yuret, Deniz. Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, 2020.
397. Salama, Ahmed; Bouamor, Houda; Mohit, Behrang, and Oflazer, Kemal. YouDACC: the Youtube dialectal Arabic comment corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1246–1251, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).
398. Salameh, Mohammad; Bouamor, Houda, and Habash, Nizar. Fine-grained Arabic dialect identification. In *Proceedings of the*

*27th International Conference on Computational Linguistics*, pages 1332–1344, 2018.

399. Salem, Fadi. Social media and the internet of things towards data-driven policymaking in the arab world: potential, limits and concerns. *The Arab Social Media Report, Dubai: MBR School of Government*, 7, 2017.
400. Salloum, Wael and Habash, Nizar. Adam: Analyzer for dialectal Arabic morphology. *Journal of King Saud University-Computer and Information Sciences*, 260 (4):0 372–378, 2014.
401. Samarin, William J. Salient and substantive pidginization. In Hymes, Dell (ed.), *Proceedings of a Conference Held at The University of the West Indies Mona, Jamaica, April 1968*. Cambridge University Press, 1971.
402. Samih, Younes; Attia, Mohammed; Eldesouki, Mohamed; Abdelali, Ahmed; Mubarak, Hamdy; Kallmeyer, Laura, and Darwish, Kareem. A neural architecture for dialectal Arabic segmentation. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 46–54, 2017.
403. Sampson, Geoffrey. *Writing systems*. London: Hutchinson, 1985.
404. Sayadi, Karim; Liwicki, Marcus; Ingold, Rolf, and Bui, Marc. Tunisian dialect and modern standard Arabic dataset for sentiment analysis: Tunisian election context. In *Second International Conference on Arabic Computational Linguistics, ACLING*, pages 35–53, 2016.
405. Sayahi, Lotfi. Code-switching and language change in Tunisia. *International Journal of the sociology of language*, 20110 (211):0 113–133, 2011.
406. Sayahi, Lotfi. *Diglossia and language contact: Language variation and change in North Africa*. Cambridge University Press, 2014.
407. Schäfer, Roland and Bildhauer, Felix. Web corpus construction. *Synthesis Lectures on Human Language Technologies*, 60 (4):0 1–145, 2013.
408. Schiffman, Harold F. Standardization or restandardization: the case for" standard" spoken tamil. *Language in Society*, pages 359–385, 1998.
409. Schmitz, Ulrich. auswirkungen elektronischer medien und neuer kommunikationstechniken auf das sprachverhalten von individuum und gesellschaft. In Werner, Besch; Anne, Betten; Oskar, Reichmann, and Stefan, Sonderegger (ed.), *sprachgeschichte. ein handbuch zur geschichte der deutschen sprache und ihrer erforschung. 2. aufl., 2. teilband [hsk 2.2].*, pages 2168–2175. De Gruyter, Berlin, 2001.

410. Schuster, Mike and Paliwal, Kuldip K. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 450 (11):0 2673–2681, 1997.
411. Schwenk, Holger. Continuous space language models. *Computer Speech & Language*, 210 (3):0 492–518, 2007.
412. Selab, Essma and Guessoum, Ahmed. Building talaa, a free general and categorized Arabic corpus. In *ICAART (1)*, pages 284–291, 2015.
413. Sghaier, Mohamed Ali and Zrigui, Mounir. Tunisian dialect-modern standard Arabic bilingual lexicon. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 973–979. IEEE, 2017.
414. Shahrour, Anas; Khalifa, Salam; Taji, Dima, and Habash, Nizar. Camelparser: A system for Arabic syntactic analysis and morphological disambiguation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 228–232, 2016.
415. Shaw, Thomas. *Travels, Or Observations Relating to Several Parts of Barbary and the Levant: Illustrated with Cuts*. A. Millar in the Strand and W. Sandby in Fleet-Street, London, 1757.
416. Shon, Suwon; Ali, Ahmed, and Glass, James. Convolutional neural networks and language embeddings for end-to-end dialect recognition. *arXiv preprint arXiv:1803.04567*, 2018.
417. Shortis, Tim. Gr8 txtpectations. *English Drama Media*, 2126, 2007.
418. Siegel, Jeff. Koines and koineization. *Language in society*, pages 357–378, 1985.
419. Simons, Gary F. Principles of multidialectal orthography design. *Workpapers in Papua New Guinea Languages*, 21:0 325–342, 1977.
420. Sinclair, John. Developing linguistic corpora: A guide to good practice corpus and text–basic principles. In Wynne, Martin (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, pages 6–23. Oxbow Books for the Arts and Human-ities Data Service, Oxford, 2004.
421. Singer, Hans R. *Grammatik der arabischen Mundart der Medina von Tunis*. Walter de Gruyter, 1984.
422. Skik, Hichem. La prononciation de qâf arabe en tunisie. *Jérome lentin& Antoine Lonnet (éds), mélanges David cohen, études sur le langage, les langues, les dialectes, les littératures*, pages 635–642, 2003.
423. Smrz, Otakar; Šnaidauf, Jan, and Zemánek, Petr. Prague dependency treebank for Arabic: Multi-level annotation of Arabic corpus. In *Proc. of the Intern. Symposium on Processing of Arabic*, pages 147–155, 2002.

424. Søgaard, Anders and Goldberg, Yoav. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–235, 2016.
425. Soliman, Abu Bakr; Eissa, Kareem, and El-Beltagy, Samhaa R. Aravec: A set of Arabic word embedding models for use in Arabic nlp. *Procedia Computer Science*, 117:0 256–265, 2017.
426. Stumme, Hans. *Tunisische märchen und gedichte: Eine sammlung prosaischer und poetischer stücke im arabischen dialecte der stadt Tunis nebst einleitung und übersetzung*. JC Hinrichs, 1893.
427. Sullivan, Natalie. *Writing Arabizi: Orthographic Variation in Romanized Lebanese Arabic on Twitter*. PhD thesis, The University of Texas at Austin, 2017.
428. Sutskever, Ilya; Vinyals, Oriol, and Le, Quoc V. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.
429. Taboada, Maite; Brooke, Julian; Tofiloski, Milan; Voll, Kimberly, and Stede, Manfred. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 370 (2):0 267–307, 2011.
430. Tachicart, Ridouane; Bouzoubaa, Karim, and Jaafar, Hamid. Building a moroccan dialect electronic dictionary (mded). In *5th International Conference on Arabic Language Processing*, pages 216–221, 2014.
431. Tachicart, Ridouane; Bouzoubaa, Karim; Aouragh, Si Lhoussaine, and Jaafa, Hamid. Automatic identification of moroccan colloquial Arabic. In *International Conference on Arabic Language Processing*, pages 201–214. Springer, 2017.
432. Taine-Cheikh, Catherine. *L'arabe médian parlé par les Arabophones de Mauritanie: étude morpho-syntaxique*. PhD thesis, Paris V – René Descartes, 1978.
433. Taine-Cheikh, Catherine. La classification des parlers bédouins du maghreb: revisiter le classement traditionnel. In Ritt-Benmimoun, V. (ed.), *Tunisian and Libyan Arabic Dialects: Common Trends - Recent Developments - Diachronic Aspects*, pages 15–42. Prensas de la Universidad de Zaragoza, Zaragoza, 2017.
434. Taji, Dima; Habash, Nizar, and Zeman, Daniel. Universal dependencies for Arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 166–176, 2017.
435. Taji, Dima; Khalifa, Salam; Obeid, Ossama; Eryani, Fadhl, and Habash, Nizar. An Arabic morphological analyzer and generator with copious features. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*,

pages 140–150, 2018.

436. Taji, Dima; Gizuli, Jamila El, and Habash, Nizar. An Arabic dependency treebank in the travel domain. *arXiv preprint arXiv:1901.10188*, 2019.

437. Takezawa, Toshiyuki; Kikui, Genichiro; Mizushima, Masahide, and Sumita, Eiichiro. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324, 2007.

438. Tannen, Deborah; Li, Charles N; Thompson, Sandra A; Heath, Shirley Brice; Green, Georgia M; Goody, Jack; Hildyard, Angela; Olson, David R; Chafe, Wallace L; Clancy, Patricia M, and others, . *Spoken and written language: Exploring orality and literacy*, volume 9. Praeger, 1982.

439. Tarquini, Maura. Ḥūma. musica rap e convergenza linguistica in arabo tunisino. *Folia Orientalia*, pages 335–362, 2019.

440. Tesnière, Lucien. *Elements of structural syntax*. John Benjamins Publishing Company, 2015.

441. Thurlow, Crispin. From statistical panic to moral panic: The metadiscursive construction and popular exaggeration of new media language in the print media. *Journal of computer-mediated communication*, 110 (3):0 667–701, 2006.

442. Thurlow, Crispin and Poff, Michele. The language of text-messaging. *Handbook of the Pragmatics of CMC*, 01 2013.

443. Toutanova, Kristina; Klein, Dan; Manning, Christopher D, and Singer, Yoram. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259, 2003.

444. Turki, Houcemeddine; Adel, Emad; Daouda, Tariq, and Regragui, Nassim. A conventional orthography for maghrebi Arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC), Portoroz, Slovenia*, 2016.

445. van der Wees, Marlies; Bisazza, Arianna, and Monz, Christof. A simple but effective approach to improve arabizi-to-english statistical machine translation. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 43–50, 2016.

446. Vanhove, Martine. La langue maltaise et le passage à l'écriture. In *Caubet, D., Chaker, S. et Sibille, J.(éds.), Codification des langues de France*, pages 369–382. L'Harmattan, 2003.

447. Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Lukasz, and Polosukhin,

Illia. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

448. Versteegh, Kees. *Pidginization and creolization: The case of Arabic*, volume 33. John Benjamins Publishing, 1984.
449. Versteegh, Kees. *Arabic language*. Edinburgh University Press, 2014.
450. Vygotsky, Lev Semënovič. Pensée et langage, trad. frse. *Paris, Editions Sociales*, 1985.
451. Walters, Keith. Gender, identity, and the political economy of language: Anglophone wives in Tunisia. *Language in society*, 250 (4):0 515–555, 1996.
452. Walther, Joseph B. and Burgoon, Judee K. Relational communication in computer-mediated interaction. *Human communication research*, 190 (1):0 50–88, 1992.
453. Warschauer, Mark; El Said, Ghada R., and Zohry, Ayman. Language choice online: Globalization and identity in egypt. *Journal of Computer-Mediated Communication*, 70 (4):0 JCMC744, 2002.
454. Woolard, Kathryn A. Language ideology: Issues and approaches. *Pragmatics*, 20 (3):0 235–249, 1992.
455. Woolard, Kathryn A. Simultaneity and bivalency as strategies in bilingualism. *Journal of linguistic anthropology*, 80 (1):0 3–29, 1998.
456. Wynne, Martin. Archiving, distribution and preservation. In Wynne, Martin (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, pages 88–96. Oxbow Books for the Arts and Humanities Data Service, Oxford, 2004.
457. Yaghan, Mohammad Ali. "Arabizi": A contemporary style of Arabic slang. *Design issues*, 240 (2):0 39–52, 2008.
458. Yates, Simeon J. Oral and written linguistic aspects of computer conferencing. *Pragmatics and beyond New Series*, pages 29–46, 1996.
459. Younes, Jihen; Achour, Hadhemi, and Souissi, Emna. Constructing linguistic resources for the Tunsian dialect using textual user-generated contents on the social web. In *International Conference on Web Engineering*, pages 3–14. Springer, 2015.
460. Younes, Jihene and Souissi, Emna. A quantitative view of Tunsian dialect electronic writing. In *5th International Conference on Arabic Language Processing*, pages 63–72, 2014.
461. Younes, Jihene; Souissi, Emna, and Achour, Hadhemi. A hidden markov model for the automatic transliteration of romanized Tunsian dialect. In *Proceedings of the 2nd international conference on Arabic computational linguistics*, 2016.
462. Younes, Jihene; Achour, Hadhemi; Souissi, Emna, and Ferchichi, Ahmed. 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, 2018a.

463. Younes, Jihene; Souissi, Emna; Achour, Hadhemi, and Ferchichi, Ahmed. A sequence-to-sequence based approach for the double transliteration of Tunsian dialect. *Procedia computer science*, 142:0 238–245, 2018b.
464. Younes, Jihene; Achour, Hadhemi; Souissi, Emna, and Ferchichi, Ahmed. Romanized Tunisian dialect transliteration using sequence labelling techniques. *Journal of King Saud University-Computer and Information Sciences*, 2020.
465. Youssi, Abderrahim. *L'arabe marocain médian: analyse fonctionaliste des rapports syntaxiques: de la synchronie dynamique dans les corrélations de normes linguistiques et des formes phonologiques, morphosyntaxiques et lexicales*. PhD thesis, Paris 3, 1986.
466. Zaghouani, Wajdi and Charfi, Anis. Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. *arXiv preprint arXiv:1808.07674*, 2018a.
467. Zaghouani, Wajdi and Charfi, Anis. Guidelines and annotation framework for Arabic author profiling. *arXiv preprint arXiv:1808.07678*, 2018b.
468. Zaidan, Omar and Callison-Burch, Chris. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, 2011.
469. Zaidan, Omar F and Callison-Burch, Chris. Arabic dialect identification. *Computational Linguistics*, 400 (1):0 171–202, 2014.
470. Zalmout, Nasser and Habash, Nizar. Don't throw those morphological analyzers away just yet: Neural morphological disambiguation for Arabic. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 704–713, 2017.
471. Zalmout, Nasser and Habash, Nizar. Adversarial multitask learning for joint multi-feature and multi-dialect morphological modeling. *arXiv preprint arXiv:1910.12702*, 2019a.
472. Zalmout, Nasser and Habash, Nizar. Joint diacritization, lemmatization, normalization, and fine-grained morphological tagging. *arXiv preprint arXiv:1910.02267*, 2019b.
473. Zalmout, Nasser; Erdmann, Alexander, and Habash, Nizar. Noise-robust morphological disambiguation for dialectal Arabic. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

*Technologies, Volume 1 (Long Papers)*, pages 953–964, 2018.

474. Zbib, Rabih; Malchiodi, Erika; Devlin, Jacob; Stallard, David; Matsoukas, Spyros; Schwartz, Richard; Makhoul, John; Zaidan, Omar, and Callison-Burch, Chris. Machine translation of Arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59, 2012.
475. Ziamari, Karima; Caubet, Dominique; Miller, Catherine, and Vicente, Ángeles. Matériaux d'enquêtes autour des' usages jeunes' dans quatre villes marocaines casablanca, meknès, tétouan, marrakech. In Trimaille, Cyril; Pereira, Christophe; Ziamari, Karima, and Gasquet-Cyrus, Médéric (ed.), *Sociolinguistique des pratiques langagières de jeunes*. UGA Éditions, Université Granoble Alpes, 2020.
476. Ziemski, Michał; Junczys-Dowmunt, Marcin, and Pouliquen, Bruno. The united nations parallel corpus v1. 0. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC)*, pages 3530–3534, 2016.
477. Zribi, Inès; Graja, Marwa; Khmekhem, Mariem Ellouze; Jaoua, Maher, and Belguith, Lamia Hadrich. Orthographic transcription for spoken Tunisian Arabic. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 153–163. Springer, 2013a.
478. Zribi, Inès; Khemekhem, Mariem Ellouze, and Belguith, Lamia Hadrich. Morphological analysis of Tunsian dialect. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 992–996, 2013b.
479. Zribi, Inès; Boujelbane, Rahma; Masmoudi, Abir; Ellouze, Mariem; Belguith, Lamia Hadrich, and Habash, Nizar. A Conventional Orthography for Tunisian Arabic. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, pages 2355–2361, 2014.
480. Zribi, Inès; Ellouze, Mariem; Belguith, Lamia Hadrich, and Blache, Philippe. Spoken Tunsian Arabic corpus "STAC": transcription and annotation. *Research in computing science*, 90:0 123–135, 2015.
481. Zribi, Inès; Kammoun, Inès; Ellouze, Mariem; Belguith, Lamia, and Blache, Philippe. Sentence boundary detection for transcribed Tunisian Arabic. *Bochumer Linguistische Arbeitsberichte*, 323, 2016.
482. Zribi, Inès; Ellouze, Mariem; Belguith, Lamia Hadrich, and Blache, Philippe. Morphological disambiguation of Tunsian dialect. *Journal of king Saud University-computer and information sciences*, 290 (2):0 147–155, 2017.

# LIST OF FIGURES

# LIST OF TABLES