Romano-Arabica
Nr. XXIII, 2024, p. 49-76
ISSN 1582-6953, e-ISSN 3044-8670

ELISA GUGLIOTTA[1]
Université Grenoble Alpes (LIG, LIDILEM)
ANGELAPIA MASSARO[2]
University of Siena
GIULIANO MION[3]
University of Cagliari
MARCO DINARELLI[4]
Université Grenoble Alpes (LIG),

DEFINITENESS IN TUNISIAN ARABIZI: SOME DATA FROM STATISTICAL APPROACHES[5]

Abstract. We present a statistical analysis of the realization of definiteness in Tunisian Arabic (TA) texts written in Arabizi, a hybrid system reflecting some features of TA phonetics (assimilation), but also showing orthographic features, as the use of arithmographs. In §1, we give an overview of definiteness in TA from a semantic and syntactic point of view. In §2 we outline a typology of definite articles and show that TA normally marks definiteness with articles or similar devices, but also presents zero-markings or weak definites. In §3 we discuss TA and how definiteness is instantiated in TA. In §4, we present data from the Tunisian Arabizi Corpus (TAC), a multidisciplinary work with a hybrid approach based on dialectological questions, corpus linguistics standards, and deep learning techniques. In §5 we define the behavior of TA with respect to what we observed in §1, §2 and §3, describing our TAC-based analysis.

Keywords: Arabizi, Definiteness, Corpus Analysis, Deep Learning, Tunisian Arabic

[1] E-mail: egugliotta@uniss.it

[2] E-mail: angelapia.massaro@unipi.it

[3] E-mail: giuliano.mion@unica.it

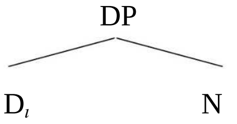
[4] E-mail: marco.dinarelli@univ-grenoble-alpes.fr

[5] All four authors collaborated on the project. For academic purposes, Massaro is responsible for §1 and §2. Mion is responsible for §3.1 and §3.2, while Gugliotta for §3.3 and §4.1. §4.2 has been jointly prepared by Dinarelli and Gugliotta, being based on Gugliotta's post-doctoral research work supervised by Dinarelli; §5.1 is under Gugliotta responsibility, while §5.2, §5.3 and *Conclusions* have been jointly elaborated by Massaro, Gugliotta and Mion.

1.Introduction

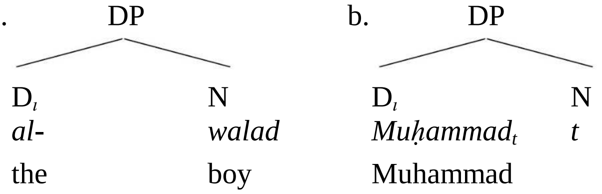
Definiteness is a semantic feature. In logical terms, a definite noun undergoes an iota (*ι*) operator, which binds it to specific referents of the same noun's property. Put simply, an iota operator selects a precise element from a set of all possible variables of the noun, shifting from property-denoting to individual-denoting elements (Longobardi 2008) as in (1), a language with pre-nominal articles.

1)



Certain nouns, like proper names, inherently possess iota semantics: they are inherently definite as they refer to unique entities. In certain languages with definite articles, proper names are non-articled, as in Standard Italian or Arabic. Syntactic theories (Longobardi 1994) suggest that in such languages, proper names occupy the position typically occupied by determiners, through a mechanism called N-to-D.

2) a.



A number of languages[6] that typically mark definiteness with articles or similar devices have in fact bare proper names. Definiteness exists in all natural languages, but its grammatical representation varies significantly. Not all languages have definite articles, yet they still express definiteness, and strategies for expressing definiteness differ across languages. This raises the question of whether definiteness is a feature specific to determiners, a syntactic position independent of determiner realization, or silent determiners (Wiltschko 2009).

[6] With the obvious exception of languages where proper names are articled, like Greek or Northern Italian varieties:

i) a. *O Yanis* /*to Yani* Modern Greek, Matushansky (2006: 286)
the Yanis /the Yani
'Yanis'

b. *La Maria* Northern Italian varieties
the Maria
'Maria'

At a level greater than simple determiner phrases, definiteness is linked to the organization of information structure. It also affects the (un)availability of certain syntactic operations which is tied to the phasehood status of DPs (Bošković 2012), (3).

3) a. *Which poem did you hear Homer's recital of last night?
(Adger 2003: 327)
b. Which poem did you go to hear a recital of last night?

In (3a), *wh-* extraction (*which poem*) is disallowed with definites, whereas (3b) demonstrates that extraction is possible with indefinites, and the impossibility of extraction with definites is connected to phases. Structure is constructed phase by phase, and once a phase is completed, its internal content becomes frozen and inaccessible to further syntactic operations (Chomsky 1998). Adger (2003), Bošković (2012), Jiménez-Fernández (2012), and others argue for the phasehood status of DPs. With regard to Semitic languages, Construct State genitives have been considered as phases (Shormani 2016), and within Romance, the same idea has been applied to genitives with definiteness agreement (Massaro 2022).

2. A Short Typology of Definite Articles

The contexts in which elements grammaticalize definiteness greatly vary across languages. Greenberg (1978) proposed four configurations, intended as diachronic stages, while also applicable synchronically. The boundaries between these stages are not clearly defined, and languages exist in between them.

Table 1
Types of languages according to the realization of definite articles, Greenberg (1978)

0	I	II	III
No definite articles. Definiteness is interpreted via other means.	A definite article emerges. Specific to definites.	Definite articles also appear with generics and nouns which are not necessarily definite.	The article is completely generalized, with no definite semantics being expressed. It functions as a nominality marker

Persian is a type 0 language. With an indefinite article, it realizes definiteness elsewhere, *e.g.* through Differential Object Marking morphology. In Mandarin, indefinite nouns are never pre-verbal, while the post-verbal position can convey definite or generic interpretations (Cheng and Sybesma 1999).

Type 1 languages can be found within Old Romance. Definite articles are a Romance innovation, in which the Latin demonstrative *ille* morphed into what we know as the definite articles of most of the Romance languages (exceptions include Sardinian, Mensching 2005, and Balearic Islands Catalan, Gaspar 2013, which developed their definite articles from *ipse*). In turn, contemporary Romance languages like Italian are type II languages. Arabic and several Arabic varieties can be considered as type II languages too (see §5.). Type III is instead represented, according to Greenberg, by languages like Gunwiggu.

In a type II language, a determiner phrase can be ambiguous between definite and generic, so interpretation depends on something more than the mere determiner phrase. Additional syntactic structure or other factors may override the definiteness feature of definite articles (5). For instance, Italian simple DPs can be ambiguous, allowing for both definite and generic interpretations.

4) *il libro* *ambiguous (either definite or generic)*
the book
'the book'

Ambiguity in the Italian DP is instead ruled out in cases as the following:

5) a. *il libro è un oggetto composto di fogli* *generic*
the book is an object made up of sheets
'a book is an object made up of sheets'

b. *il libro di mia madre* *definite*
the book of my mother
'my mother's book'

Additional structure dissolves the ambiguity that we found in the simple DP. Anticipating the discussion on TA somewhat, the following example shows that also in this language, definite articles do not always trigger a definite interpretation.

6) *awel mara nozi nilbes robe fel chté,*
/āwwəl maṛṛa nūzi nəlbəs robe f-əl šita/
First time I:dare I:wear dress in-the winter
'It is the first time I dare to wear a dress in winter'

Exactly the same happens in Italian.

7) *messo al muro*
put to.the wall
'painted into a corner'

Type II languages are particularly apt to show that, as Ramchand and Svenonius (2008) argue, the mapping from syntax to the C-I system is not trivial[7], consequently posing a challenge for NLP tasks. For Semitic Construct State genitives and Romanian genitives, it raises the question of how a definite interpretation is achieved without explicit marking. In Construct State, for instance, heads lack definiteness marking, yet the entire phrase is interpreted as definite.

8) a. *ṣəʾif* *ha-yaldá* Hebrew, Borer (1988: 48)
scarf the-girl
'the girl's scarf'

b. *kitābu* *l-binti* Arabic, Hoyt (2008: 5)
book the-girl
'the girl's book'

Borer (1988) proposes that the definiteness feature of the modifier percolates to the head, resulting in the whole phrase being definite. Hoyt (2008) demonstrates that phrases with heads similar to (8b), but with indefinite modifiers, are indeed interpreted as indefinite.

9) *kitābu* *bintin* Hoyt (2008: 6)
book girl
'a girl's book'

Romanian has two types of genitives. In one type, oblique morphology is sufficient. In the other type, a linker element appears between the head and the modifier, bearing oblique morphology. Typically, non-linker genitives are limited to definites, while indefinite contexts require a linker (Dobrovie-Sorin 2000).

10) a. *casa* *vecin-ului*
house neighbor-the
'the neighbor's house'

b. *o casa* *a* *vecin-ului*
a house LKR neighbor-the
'a house of the neighbor's

However, in some instances, non-linker genitives can also contain indefinite nouns, as in (11).

11) *confesiunile* *unui* *asasin* *economic*
confessions-the a hitman economic
'the confessions of an economic hitman'

[7] And namely a conceptual-intentional system processing linguistic information, i.e. responsible for its interpretation (Hauser et al 2002).

Dobrovie-Sorin (2000: 216), states that “the denotation of the overall nominal projection is obtained by applying the denotation of the head N to the denotation of the DP in SpecDP” (SpecDP is the position assigned to the genitive, in her work). Like for Semitic Construct State (with a difference in the direction of definiteness percolation), a definite interpretation is achieved through mechanisms like (in)definiteness spreading.

Complementizers are similar to iota operators. The variable they bind is then realized within the predication contained in the complementizer phrase.

12) *ʾaxu l-walad illi byidrus bi-ʾamērka* Palestinian Arabic,
brother the-boy that studies in-America Mohammad (1999: 32)
‘the brother of the boy who studies in America’

The predication inside the complementizer phrase serves in fact as precise individuation of the reference expressed by the noun (*l-walad*) it modifies. Higginbotham (1985: 563) suggested that modification is analogous to coordination (see also Bošković 2020).

13) *a big butterfly=that is a butterfly, and it is big (for a butterfly)*

In a similar vein, also the restrictive interpretation yielded by complementizers can be said to be similar to coordination.

14) *the brother of the boy who studies in America=he is the brother of the boy, and the boy studies in America*

Next in this paper we will try to make sense of how definiteness is realized in TA Arabizi. But first, an introduction to TA is in order.

3. Tunisian Arabic

3.1. General Overview

TA, also known by the autoglottonym *derja* (or, in scientific transcription, *dārža*; see St. Ar. *dāriǧa* ‘current language, dialect’), is one of the North African varieties of Neo-Arabic. The label generally refers to the Arabic dialects spoken in the Republic of Tunisia.[8]

According to the general classification established in the Arabic dialectology, TA is one of the varieties spoken in the Eastern Maghreb and, as a Maghrebi dialect,

[8] A TA diasporic dialect is spoken in Mazara del Vallo (Sicily, Italy), for which see D’Anna (2017).

it is typically characterized by the *n*-prefix of the imperfective, as in *nqūl* 'I say' and *nqūlu* 'we say' (whereas both Old-Arabic and the Eastern Arabic dialects have *ʾaqūl* ≠ *niqūl*).

TA is considered as particularly relevant for its crucial role in the Arabicization of North Africa. In fact, it is worth remembering that the city of Kairouan (Central Tunisia) was the first Arab settlement in Ifrīqiyā, founded in 670 A.D. by ʿUqba ibn Nāfiʿ. The Arabicization of the Maghreb had its starting point in this city.[9] Consequently, the other North African sedentary dialects would be genetically related to Kairouan to the point that they have been named *parlers kairouanais* according to the definition given by Cohen (1988).

In the eleventh century, North Africa was invaded by some Bedouin tribes of Arabian origin, the Banū Hilāl and the Banū Sulaym, who came from Egypt. This event is traditionally considered a significant watershed in the linguistic history of the region, as the arrival of these tribes is at the basis of a typological dichotomy existing until nowadays between the sedentary and the Bedouin dialects. The first ones date back to the first phase of the Arabicization, when the Arabs conquered North Africa in the seventh century, while the latter resulted from the Hilalian invasions.[10] The current dialectological situation, that is the result of these historical events, consists of several urban dialects (mainly situated in the coastal areas), some rural dialects (the best known, even if partially, are those of the Sahel region), and a great number of Bedouin dialects. These differences have not been taken into consideration in our research, as the language expressed through Arabizi often appears as a pretty koineized dialect.

Today, TA is an unofficial language, and is still used mainly as a spoken language for informal communication, and there is no fixed tradition for its practice in written domains. But nonetheless, after the so-called Jasmine Revolution of 2011, publications in TA began to touch several written domains that had previously been a prerogative of Standard Arabic, and both their amount and their quality increased considerably. These publications consist of novels, translations of foreign novels, magazines, and even some essays, and they find generally positive feedbacks in post-revolutionary Tunisia. The TA used in these publications is habitually a standardized koine based on the urban Tunis dialect, and it is written in an Arabic alphabet that tends to replicate the orthographic rules of Standard Arabic. Aside from these habits, some activists began to claim the full independence of TA (simply called by its autoglottonym *Derja*) from the Arabic phylum.[11] They proposed the adoption of two parallel writing systems: a first and more traditional system that consists in the adoption

[9] There is a large bibliography on the history of the Arabic language in North Africa, a first reference is Marçais (1961).

[10] For a general presentation of the Tunisian dialectological situation, see Marçais (1950) and, more recently, Baccouche (2009).

[11] They organized themselves in a very active association named *Derja*.

of the Arabic alphabet, and a second system based on the Latin alphabet with some modifications concerning special graphemes inspired by Maltese and IPA, that, however, is completely different from the Arabizi system used for our research. These proposals are still far from being adopted or, at least, seriously taken into consideration by Tunisians, and they have not been analyzed in our contribution.

3.2. Definiteness in Tunisian Arabic

In the field of Arabic linguistics and dialectology, several studies deal with definiteness from very different points of view (Turner 2018). Many of them concern the formal representation of definiteness and discuss forms and roles of the definite article */al-/.[12] Conversely, others analyze the emergence and the development in several Neo-Arabic varieties of an 'indefinite article', an element that is not attested in Old-Arabic (Mion 2009; Edzard 2006). So, while the situation of the definite article cross-dialectally is quite stable, instead it has been noticed that elements representing indefinite articles emerged mainly in the peripheries of the Arabic-speaking world, due to interlinguistic contacts (Mion 2009; Turner 2021). In fact, in some Neo-Arabic varieties located at the edges of the Arabic Sprachraum, an indefinite article is issued from the grammaticalization of terms related to the notion of 'singularity', typically the numeral *wāḥid* 'one' or other items referring to individuality like *e.g. fard* 'single or individual (thing/person)': from the first Moroccan Arabic derives *waḥd-əl-*, from the second Mesopotamian Arabic derives *fadd* and other variants (Leitner and Procházka 2021).

But beyond the extremely schematic introduction given so far, the situation of the strategies marking (in)definiteness among the Arabic dialects is more entangled. Recently, Turner (2021) proposed a general classification of the Arabic dialects using a semantic typology that distinguishes two main groups: 1) dialects with a strict formal distinction between true definites and indefinites, and 2) dialects with a lax formal distinction between true definites and indefinites, each group having its subgroups. Even if not expressly mentioned in Turner's work, TA can easily fit in the subgroup with no highly conventionalized marking of indefinites, which belong to the first group.

So, broadly speaking, a non-articulated noun like *ṛāžəl* is unmarked and indefinite and it means 'a man', while an articulated noun like *əṛ-ṛāžəl* is marked and definite and it means 'the man'. Anyhow, a non-articulated noun can be considered definite if it appears in certain syntactic contexts (or if it is a proper noun) and, on the contrary, as already shown in 1.2, a definite noun does not always imply a definite interpretation: the Arabizi corpus of our research includes several examples of both these conditions.

[12] See, e.g., Zaborski (2006) for a concise diachronic perspective.

Consequently, definiteness is a system more complicated than the mere morphological operation of marking or unmarking a noun with or without an article.

In TA definiteness appears to be organized hierarchically through a regular series of levels. As shown in Table 2, definiteness is delineated along a continuum that ranges from strongly marked as generic elements (++generic) to strongly marked as specific elements (++specific), passing by the intermediate levels of genericity (+generic) and specificity (+specific). The division between (+generic) and (+specific) exhibits the transition from an unmarked indefiniteness (Ø) to a marked definiteness (*/al-/) feature.

As for the strongly marked elements (++generic and ++specific), in addition to the typical features of indefiniteness or definiteness, we can find elements reinforcing definiteness: in the case of (++generic) we find the intervention of the numeral *wāḥəd*, and in the case of (++specific) the intervention of demonstrative adjectives, like *e.g. hāḏa* 'this'. Demonstratives often function as reinforcers, as in the case of Romance and Germanic languages, for instance (see Bernstein 1997; Brugè 1996).

Table 2

Definiteness continuum in TA

++ ***GENERIC***	+ ***GENERIC***	+ ***SPECIFIC***	++ ***SPECIFIC***
I look for a man	I look for a man	I look for the blond man	I look for that blond man
nlawwəž ʿla ***wāḥəd*** *ṛāžəl*	*nlawwəž ʿla ṛāžəl*	*nlawwəž ʿla* ***ər***-*ṛāžəl* ***lə***-*blond*	*nlawwəž ʿla* ***ər***-*ṛāžəl* ***lə***-*blond* ***hāḏa***

In conclusion, in the case of (++ generic) *wāḥəd* remains in the orbit of the nominal class without becoming an indefinite article, and its intervention can be reinterpreted as a sort of reduced relative clause (= 'someone who is…'). Semantically, this becomes even more evident when the element that appears after *wāḥəd* is an adjective: *nlawwəž ʿla wāḥəd rūsi* 'I look for a Russian' → 'I look for someone who is a Russian'. In the case of (++ specific) the deictic element reinforces the level of definiteness and it is worth noting that is usually postponed to the noun, according to the syntactic rules of TA, or that the noun can be inserted between two deictic elements, the first one proclitic and the latter postponed: *nlawwəž ʿla hā-r-rūsi hāḏa* 'I look for this Russian'.

3.3. Tunisian CMC and the Arabizi Encoding

The Arabizi encoding emerged in the Arabic-speaking world to bridge a technological gap following the introduction of electronic devices in the late 1990s. These devices lacked Arabic keyboards or input systems for typing in Arabic script. Arabizi, along with “Arabish”, is the most popular term today (Bianchi 2013).

The use of Latin-based encoding in languages with non-Latin alphabets is also observed in Greek (*Greeklish*) and Serbian (*Latinica*). Androutsopoulos and Schmidt (2002) and Jaffe et al. (2012) employ the term *neography* to describe Greeklish. Similarly, Arabizi approximates TA phonology while incorporating elements, like digits, to represent Arabic graphemes, as shown in the following table.

Table 3

Arabizi Code System for TA – only most common Arabizi graphemes have been reported

Arabic script	ا ى	ء	ب	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص
Tunisian Arabizi	a e	2	b p	t	th	j	7 h	5 kh	d	dh	r	z	s	ch	s
Arabic script	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ي	ة
Tunisian Arabizi	th dh	6 t	th dh	3 a	4 gh	f	9 q	k	l	m	n	8 h	ou w	y i	h a t

Albirini (2016) discusses Arabic usage on the Internet from a socio-linguistic perspective, highlighting the prevalence of multilingualism and code-switching. He observes that young social network users employ an informal register, understanding the context of their communication rather than adhering strictly to standard language rules. Caubet (2019) ascribes the widespread adoption of *dārža* in written form to grassroots movement termed ‘Do It Yourself’, describing it as a collective effort to acquire literacy in an unstandardized language (Caubet 2019: 391).

We view Arabizi from two perspectives: as a neutral technology for representing spoken language, and as a socio-linguistic phenomenon itself. In the case of TA, no standardized system exists for its representation, leading to challenges with both Arabic and Latin scripts. Alghamdi and Petraki (2018) show that Arabizi appears, to the young CMC users in Saudi Arabia, as easier, faster, flexible, and also stylish. This preference may stem from the familiarity with the Latin keyboard. While Facebook and Twitter emerged in 2006, their Arabic versions were introduced in 2009 and 2012, respectively (Alghamdi and Petraki 2018). Facebook’s impact on Tunisian society, highlighted by Salem (2017), underscores the significance of Arabizi.[13] Younes and Souissi (2014) collected a corpus of TA messages, revealing that over half were encoded in Arabizi.

[13] Tunisia is the third most active Arab country on Facebook. Twitter is only 2% widespread.

The substantial volume of linguistic data generated in Arabizi significantly impacts linguistic research, particularly in Natural Language Processing (NLP) for colloquial Arabic. Access to extensive text data is crucial for NLP, and the abundance of Arabizi content has led to increased visibility for TA in recent years within the field of Arabic NLP.

4. Tools and Data Employed

4.1. An Overview on the Tunisian Arabizi Corpus (TAC)

The rise of Dialectal Arabic (DA) data has boosted research on DA in the NLP field (Bouamor et al. 2018; El-Haj 2020). This facilitates DA tool development by adapting existing MSA tools like Penn Arabic Treebank (Maamouri et al. 2004) and creating DA corpora from web data. Specific DA tools are crucial for effective NLP on Arabic social media, where DA is prevalent (Diab et al. 2010: 66). Our research employs the Tunisian Arabizi Corpus (TAC) (Gugliotta and Dinarelli 2020), designed for web-based dialectological investigation using a hybrid approach of dialectology, corpus linguistics, and deep learning techniques. TAC addresses the challenge of the lack of standardized DA encoding by employing the Conventional Orthography for Dialectal Arabic (CODA), providing specific guidelines for dialect-based conventions (Habash et al. 2018). TAC texts were encoded into Arabic script using CODA*.

Various corpus types include parallel, mono-varietal, and annotated corpora, like LDC’s Levantine and Egyptian Arabic Treebanks (Maamouri et al. 2014), offering syntactic annotations. Fisher Levantine Arabic Conversational Telephone Speech (Maamouri et al. 2007) contains spoken text. The Levantine Dialect Corpus (Shami) covers Palestine, Jordan, Lebanon, and Syria dialects with 117,805 non-annotated tweets (Kwaik et al. 2018). Curras is a written Palestinian Arabic corpus with about 56,000 tokens, morphologically annotated using the MADAMIRA tool (Jarrar et al. 2017). MADAMIRA (Pasha et al. 2014) was also used for SUAR, a Saudi Arabic corpus with 104,079 words, where automatic annotations underwent manual review (Al-Twairesh et al. 2018). Alsarsour et al. (2018) built DART, a dataset of about 25,000 crowd-sourced annotated tweets. TAC follows a similar approach using a multi-task architecture (§4.2) for semi-automatic annotation on five levels.

- Word classification into three classes - *arabizi* (TA and MSA words), *foreign* (non-Arabic code-switching), and *emotag* (smileys or emoticons).
- Encoding in CODA* (Habash et al., 2018).
- Tokenization, words split into morphemes.
- PoS tagging, adhering to the PATB guidelines (Maamouri et al., 2009).
- Lemmatization in CODA*.

All annotation levels were produced semi-automatically, detailed in Gugliotta and Dinarelli (2020) and Section 4.2, where we explore how the multi-task architecture benefited from leveraging the MADAR corpus. MADAR, a parallel corpus, encompasses 25 Arab-city dialects, along with existing English, French, and MSA parallel sets (Bouamor et al. 2018).

Social media's advent has facilitated the corpus construction through web-data extraction. However, TA still lacks large and consistently annotated corpora to explore innovative automatic processing methods (Gugliotta and Dinarelli 2020). Research efforts has been on multi-dialects, mainly Saudi, Gulf, and Egyptian Arabic, with less emphasis on Maghrebi dialects, particularly TA (Guellil et al., 2019: 9). Although there are corpora that include or focus on TA, freely available Tunisian corpora are limited in quantity.

TAC corpus is readily available for free download.[14] It captures a snapshot of TA in Arabizi and its evolution over the past decade. The corpus selection adheres to specific criteria (Gugliotta and Dinarelli 2020):

a) Text mode: informal writing;
b) Text genres: forum, blog, social networks;
c) Domain: CMC;
d) Language: TA in Arabizi;
e) Location;
f) Publication date.

Metadata extraction recorded the publication date, user's age, gender, and provenience. TAC's creation involved a semi-automatic annotation (§4.2), aiming to achieve consistent linguistic annotation. Table 3 displays some statistics from the data collected in TAC.

The applicative corpus goals involve developing NLP tools for processing TA Arabizi, facilitated by the multi-functional annotation levels in TAC. This enables comprehensive and systematic studies of TA and its Arabizi encoding, contributing to the dialectological domain where the initial research questions were addressed.

[14] TAC corpus is available at: https://github.com/eligugliotta/tarc.

Table 4

TAC Data information

Total:	**sentences**			
	4,790	**Tokens Classification**		
Text Genres:		*arabizi*	*foreign*	*emotag*
Forum	756	6,022	5,874	13
Social Networks	3,154	11,833	3,624	598
Blog	366	5,988	674	7

4.2. Corpus Collection Incremental Semi-Automatic Procedure

To streamline corpus collection process for human annotators, deep-learning techniques were employed, implementing a semi-automatic annotation procedure (Gugliotta et al. 2020). Specifically, a multi-task sequence-to-sequence neural architecture based on LSTM Recurrent Neural Networks (RNN) (Hochreiter and Schmidhuber 1997; Sutskever et al. 2014) was utilised. This system can handle one or more input sequences, automatically adapting to the number of outputs based on the data format, making it versatile for various phases of the annotation procedure with different levels of annotation available.

Figure 1 illustrates the multi-task system, instantiated to take one input (*x* in Figure 1) and generate three different outputs (ô1, ô2, ô3). Figure 1 highlights an essential aspect of the model: learning jointly and sequentially to generate multiple outputs allows the system to factorize information between annotation levels. Training on different tasks simultaneously, the model learns from each level, leading to mutual improvements across all the generation levels. This inter-task learning enhances the overall performance and effectiveness of the annotation procedure.

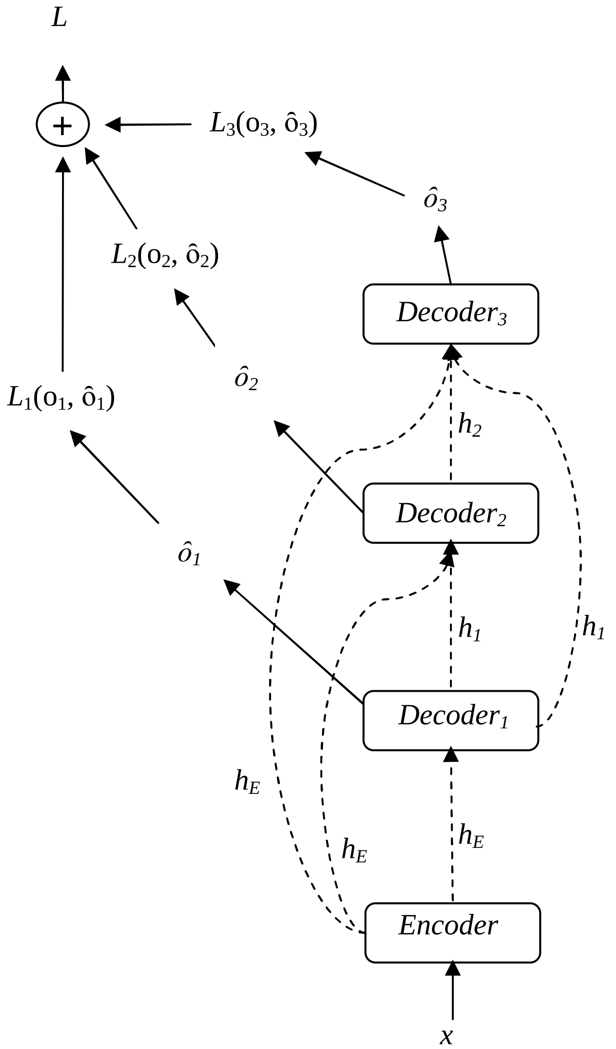


Figure 1. Multi-Task Architecture high-level schema

The iterative semi-automatic procedure for TAC (Gugliotta and Dinarelli 2020) initially lacked annotated data. We chose to manually transliterate three TAC blocks into Arabic-script. The accuracy was around 65%. To address the challenge of Arabizi data's spontaneous nature during transliteration, we introduced 2,000 sentences (not-spontaneous Arabic-encoded Tunisian data) from the MADAR corpus. This involved semi-automatic annotation with the intended levels: Classification, Tokenization, and PoS tags, before continuing with the other data blocks.

Subsequently, the semi-automatic TAC annotation procedure started. Global results, including Lemmatization, can be found in Gugliotta and Dinarelli (2022). The table below displays experiments involving the last TAC block (seventh) and mono-task experiments.

Table 5

Last step of the semi-automatic procedure used for TAC annotation compared with mono-task results

	Train. Tokens	LSTM			
Task		**Class**	**Arabic**	**Token**	**PoS**
Corpus: MADAR$_{\text{Arabizi}}$+TAC					
Step6	46,197 (33,806)	96.5%	83.3%	81.94%	81%
Step6 - Arabic only	46,197 (33,806)	92.8%	79%	-	-
Step6 - Token only	46,197 (33,806)	-	-	95.4%	-
Step6 - PoS only	46,197 (33,806)	-	-	-	86.2%

Table 5 displays the number of tokens used for training the model for each specific step (TAC corpus tokens are shown in parentheses, while the rest belongs to MADAR). According to the observations made by Gugliotta and Dinarelli (2022), the transliteration task in Arabic-script, using Arabizi as input, is the most ambiguous annotation task. To mitigate data scarcity and reduce ambiguity, the MADAR data were also annotated with an Arabizi script level. This helps improve the prediction of Arabic script from Arabizi. Consequently, ‘Step 6’ represents the last annotation stage for data block 7, achieved by concatenating MADAR and TAC data.

In Table 5, we present results for a proof of concept in a mono-task setting, beyond the Arabic-script encoding (Arabic only in Table 5) where Class information is used. Predicting Arabic-script from Arabizi (plus Class) achieved 79% accuracy, slightly worse than the multi-task setting (83.3%). However, predicting Tokenization from Arabic-script only (Token only in Table 5) resulted in 95.4% accuracy, significantly better than the multi-task setting (81.9%), indicating the impact of Arabic script encoding errors. For PoS tagging (PoS only), the accuracy reached 86.2%, more than 5 points better than the multi-task setting, considering the challenges of predicting two previous annotation levels. Overall, the system is mainly affected by the ambiguous Arabizi to Arabic-script transliteration.

5. Analyses

5.1. Background

We examined the definiteness marker in CODA* to ensure its semantic and syntactic accurate placement in Tunisian sentences encoded in Arabizi. TA, like MSA, uses a single definite article, */al-/. When preceding nouns starting with a coronal phoneme,[15] it assimilates, leading to gemination of the noun's initial, as in (15), with an original Arabizi phrase from TAC followed by our translation.

15) *Inchalla cycle ejjay wala eli ba3dou*
/nšālla cycle əž-žāy walla əlli baʿdu/
God willing time the-next or the one after:that
'God willing next time, or the time after that'.

Example (15) highlights how TA phonological characteristics are mirrored in Arabizi, presenting challenges for automatic processing. In the multi-tasking system, this complexity leads to imprecise outputs in transliteration, subsequently affecting tokenization and PoS-tagging (Section 4.2). To mitigate these inaccuracies, manual corrections are implemented iteratively and added to the training data, offering the system accurate learning examples (see §4.2).

To ensure data corrections, we analysed definiteness in TA, revealing a continuum (Table 1). However, while reviewing automatic annotations, we identified nominal phrases deviating from the prototypical categories, as shown in (16-17).

16) *ennes tamel fel fazet*
/ən-nās taʿməl fīl-fāzāt/
the-people do in:the-things
'People do things'.

Example (16) exhibits generic names with definiteness marks, posing no processing issues as the system recognises, transliterates, and morphologically annotates the text accordingly to CODA. Conversely, difficulties arise when nouns are specific or contextually defined but lack definiteness marks, as in (17).

17) a. *fi zit eldeifi*[16]
↓/fi zīt əd-dāfi/
In oil the-tepid
'In the tepid oil'.

[15] These in TA are /t/, / ṯ/, /d/, /ḏ/, /r/, /z/, /s/, /š/, /ṣ/, /ḏ̣/, /ṭ/, /l/, /n/ and /ž/.

[16] Having found other similar cases for the same user, we leave open the hypothesis of cases of the user's idiolect.

b. *fi zit eldeifi*
*/fi zīt (əl)li dāfi/
In oil that one tepid
'In the oil which is tepid'.

In Tunisian, modifiers of a definite noun are also definite, but here only the modifier *dāfi*, 'tepid', has the marking, while the noun *zīt*, 'oil', seems lacking it. Considering Arabizi's tendency to graphically represent article assimilation, we considered an alternative. The proclitic element preceding *dāfi* might be a relative pronoun, *illi*.[17] In this case, the sentence would result as in (17b).

Neither (17b) or (17a) are completely acceptable, if we consider that the head of a relative sentence in TA is generally definite. However, in TAC it is possible to encounter relative sentences apparently with generic heads (see also Mion 2014: 69; Marçais 1952: 504), as in (18).

18) *fi jarayéd elli na9raw*
↓/fi žrāyəd əlli naqrāw/
In journals that we:read
'In the journals that we read'.

However, such occurrences are rare; it is more likely that the nouns *zīt* and *žrāyəd* are definite but lack graphic traces of the definiteness marking due to assimilation. The nominal phrase is preceded by the preposition *fi*, which tends to absorb the initial phoneme of the definiteness marking */al-/, while /l-/ assimilates to the initial of the noun *zīt*, a coronal phoneme. Arabizi, a hybrid system reflecting Tunisian phonetics, incorporates orthographic features like *arithmographemes* – digits used as graphemes selected through analogical substitutions with Arabic graphemes. By observing similar cases, dedicated analyses were needed to identify potential causes of mismatch between definiteness traits and marking, to improve text transliteration and annotation. These analyses are discussed in the following sections (5.2 and 5.3).

5.2. TAC-based Analyses

In a preliminary analysis phase, we decided to examine the first 15,000 TAC tokens, containing 1,036 nouns. We categorized these nouns into generic and specific. In TA, non-articulated nouns are considered definite if they:

[17] Among the variant of *illi* there is *li*, if preceded by a word ending with vowels.

a) Present the possessive pronouns, i.e.: /žīb-u manqūb/, lit.: ‘His pocket (is) holed’.
b) Are preceded by the vocative particle (/yā/), i.e.: /yā žmāʿa/, lit.: ‘Hey group’.
c) Are proper nouns, i.e.: /tūnəs/, ‘Tunis’.
d) Are in the Construct State, i.e.: /rūḥ əl-lūz/, ‘Almond essence’ (litt. ‘The spirit of the almond’).

With the aim to identify non-prototypical NP for both generic and specific categories, we observed the following percentages: in 25% of the sentences NP is generic, but is preceded by a definite article.

19) *awel mara nozi nilbes robe fel chté*
/āwwəl maṛṛa nūzi nəlbəs robe f-əl šita/
First time I:dare I:wear dress in-the winter
‘It is the first time I dare to wear a dress in winter’.

In 74% of the sentences the NP is specific but lacks explicit marking.

20) *elli y3ichou fi bled*
/əlli yʿīšu fi blād/
REL they:live in country
‘That ones who live in (the) country’.

Observing 25% of the articulated generics category, like the sentence in (19), we found that these are primarily idioms. Non-articulated specifics present challenges for high accuracy in NLP tasks, comprising 19% of the observed data. Within this subset, 32% exhibits typical Arabizi encoding behavior, where article assimilation resulting in gemination is not always represented. The remaining 68% of non-articulated specific nouns can be attributed to other summarized cases. Idioms, as in (21).

21) *klem 3lik w ma3na 3la jarek*
/klām ʿlīk w maʿna la žārək/
Words on:you and meaning on neighbor:your
‘I speak to you but I refer to someone else’.

Definite generic phrases may appear as indefinite specific phrases, where *klem*, ‘words’, lacks the definite mark despite its specific referent in the context. Some non-articulated specific nouns results from typing inattention common in CMC writing. Additionally, several non-articulated (but specific) nouns are due to elative adjectives. In TA, a superlative structure is expressed using an elative adjective (on the *ʾafʿal* form) followed by a bare noun (22).

22) *konna a7la couple t7attét a3lina el 3ine*
/kunna āḥla *couple* tḥaṭṭīt ʿalīna əl īn/
we:were the:best couple direct at:us the envy
'We were the best couple to direct envy at'.

Particular structures connected to quantity semantics (23).

23) *9ad ka3bet ma9roud*
/qadd kaʿbāt maqrūḏ̣/
same size units maqroudh
'The same size as Maqroudh units*'*.[18]

24) *chweya 7achw*
/šwəyya ḥašu/
little bit filling
'A bit of filling'.

As seen in (23-24), in TA, nominal elements quantify nouns. We investigated specialized quantifiers for countable and uncountable elements like liquids, powders, and gases, and whether their presence leads to different structures based on specificity or genericity of the quantified nouns. In the definiteness continuum scheme (Figure 1), we observed a de-numeral element, *wāḥəd*, functioning as an indefinite pronoun or adjective, reinforcing noun genericity. In post-nominal position, it acts as a noun modifier with the original semantic trait of unicity. For example, we provide two sentences, each showcasing a different *wāḥəd* usage.

Sentence (25) illustrates the indefinite function of the pronoun *wāḥəd*, serving as the head of a reduced relative clause. In contrast, sentence (26) demonstrates *wāḥəd* employed as a numeral adjective with its original quantifier meaning.

25) *nlawwej 3la we7ed rajel*
/nlawwəž ʿla wāḥəd ṛāžəl/
I:look for someone man
'I look for someone (who is a) man'.

26) *nlawwej 3la rajel we7ed*
/nlawwəž ʿla ṛāžəl wāḥəd/
I-look for man one
'I look for (only) one man'.

[18] *Maqroudh* is a typical Tunisian sweet.

The usage of *wāḥəd* as an indefinite pronoun is more prototypical for referents with the [+human] feature. However, for [-human] nouns, the situation is more complex, as illustrated in Table 6.

Table 6

Definiteness & quantity continuum in TA

++ GENERIC	+ GENERIC	+ SPECIFIC	++ SPECIFIC
⟶			
I look for someone (who is) a man	*I look for (only one) man*		
/nlawwəž ʿla **wāḥəd** ṛāžəl/	/nlawwəž ʿla ṛāžəl **wāḥəd**/		
Every apple is good	*Every morning I eat **a unit** of apples*	*Tomorrow I will eat **a unit** (of) the apples*	*Tomorrow I will eat **a unit** (of) the apples **this***
/kull tuffāḥa bnīna/	/kull ṣbāḥ nākəl **kaʿba** tuffāḥ/	/ġudwa bāš nākəl **kaʿbət** əl-tuffāḥ/	/ġudwa bāš nākəl **kaʿbət** əl-tuffāḥ **hāḏi**/

However, TA employs the *kaʿba* quantifier to define the uniqueness of elements. For example, in *kull ṣbāḥ nākəl* ***kaʿba tuffāḥ*** it selects a unique apple from a set characterized by a [-human] feature, and functions as a first element in an appositive structure with the plural noun *tuffāḥ*.[19] In specific contexts, *kaʿba* modifies the noun to express partitivity, as seen in *ġudwa bāš nākəl* ***kaʿbət əl-tuffāḥ***.[20] Furthermore, in ++*specific* contexts, TA can also reinforce specificity with demonstrative elements like *hāḏi* (fourth sentence of Table 6). To further examine quantifier behavior, we conducted a dedicated survey outlined in the following section (5.3).

5.3. A Survey on Tunisian Quantifiers

From the previous paragraph, the structure [*kaʿba*[(DEF-)[N]]] is unacceptable in Tunisian Arabizi if the noun is [+human]. Instead, some quantifiers help select a quantity of elements in a set with similar physical features (as for collective nouns), excluding human beings. Instead, [+human] nouns can be quantified employing the universal *kull* 'all' or its opposite *ḥadd* 'nobody' (31-33); indefinite adverbs like *barša* 'a lot' (27), and numerals.

[19] The structure is [*kaʿba*+n.PL].

[20] The structure [*kaʿba*[DEF-[N-PL]]] coincides with the Construct State's one: [N[DEF-[N]]]. We express a doubt on the grammaticality of the former, as noted in Massaro (2022).

As mentioned earlier (sections 5.1, 5.2, and Table 6), definite nouns [+human] can be accompanied by reinforcers (R), like *wāḥəd* for ++*generic* nouns or demonstrative adjectives, like *hāḏa*, for ++*specific* ones, as outlined below. In affirmative sentences:

27) *barša aʿbād təsraq* *quantifier*
many people they:steal
'Many people steal'.

28) *nlawwəž ʿla wāḥəd ṭbīb* *pronoun*
I:look for one doctor
'I look for one (who is a) doctor'.

29) *wāḥəd ṭbīb yḥəbb…* *pronoun*
One doctor he:want
'One (who is a) doctor wants…'.

30) *yxaddmu ṭbīb wāḥəd* *quantifier*
They:hire doctor one
'They hire (only) one doctor'.

In negative sentences:

31) *mā famma ḥadd ṭbīb* *pronoun*
not there is nobody doctor
'There is nobody, who is a doctor'.

32) *mā fammā-š wāḥəd ṭbīb* *pronoun*
not there is-not one doctor
'There is not (someone who is) a doctor'.

33) *mā famma ḥadd* *pronoun*
not there is nobody
'There is nobody'.

34) *mā fammā-š wāḥəd kbīr*[21] *pronoun*
not there is-not one big
'There is not a big one'.

35) *mā fammā-š ṭbīb wāḥəd*[22] *quantifier*
not there is-not doctor one
'There is not (only) one doctor'.

[21] The sentence, without the second part of the circumfix negative mark /š/, is not correct.

[22] The whole sentence is /mā fammā-š ṭbīb wāḥəd famma barša/, 'There is not (only) one doctor, there are a lot'.

Wāḥəd can also be used in interrogative sentences, as a pronoun (36) or as a quantifier (37).

36) *mā fammā-š wāḥəd ṭbīb?* *pronoun*
not there is-not one doctor
'Is there not (someone who is) a doctor?'.

37) *famma ṭbīb wāḥəd?* *quantifier*
there is doctor one
'There is (only) one doctor?'.

Observing the examples, *wāḥəd* seems to function as a genericity reinforcer solely in pre-nominal position, being an indefinite pronoun meaning '(some)one'. here, it heads a reduced relative clause, as in (28-29), (32), (34) and (36), where *ṭbīb*, 'doctor' is a predicate. Instead, in post-nominal position (typical adjectival position), it functions as a numeral, conveying '(only) one', as in (30), (35), and (37). Therefore, generic reinforcement follows the [R[CPø □[N_{+human}]]] structure.

Specific DPs, reinforced by demonstrative adjectives, present a [DP[N_{+human}]][DP_R]] structure (see Brugè 1996:19), as in Table 6 and in (38):

38) /ġudwa bāš nqābəl l-aʿbād hāḏūma/
tomorrow will I:meet the-people these
'Tomorrow I will meet these people'.

Regarding quantifying [-human] nouns, different quantifiers are employed for nouns with [+countable] or [-countable] features. To examine quantifiers adhering to semantic categories and the generic-specific continuum in TA, a survey gathered additional data beyond the corpus. Sixty sentences featuring countable and uncountable nouns with quantifiers were rated by informants on a 1-5 scale (1 for 'not acceptable' and 5 'very acceptable'). Sixty-four informants participated, with fifty-three proving partial responses, totaling one hundred and seventeen informants. For instance, the first sentence of the survey and its results in Table 7 indicates that 89.69% of informants deemed it 'not acceptable' due to the absence of definiteness masking for 'doctor' and the demonstrative reinforcer 'this'.

Table 7

Sentence: ‘There is not this (Ø)doctor?’. Command translation: ‘Please choose only one among the following: /1/2/3/4/5’

ما فمّاش هاذا دكتور ؟			
من فضلك اختر واحدا فقط مما يلي:	**COUNT**	**GRAND PERCENTAGE**	**TOP 2**
①	76	64.96%	89.69%
②	11	9.40%	
③	5	4.27%	5.15%
④	1	0.85%	
⑤	4	3.42%	5.15%
VALID TOTAL	**97**		**100%**
NO ANSWER	4	3.42%	
NOT VISUALIZED	16	13.68%	
GRAND TOTAL	**117**		**100%**

Based on the survey, our initial conclusions on Tunisian quantifiers, in Table 8, classify them into three classes based on the traits of the quantified noun. The first class comprises quantifiers primarily used for uncountable nouns. For instance, *ḥafna* ‘handful’ is suitable for nouns like ‘flour’ or small countable elements, like ‘almonds’, but not for ‘tomato’. *Rašfa* ‘sip’ is exclusively used for liquids for drinking, while *kīla* ‘measure, portion’ and *kās* ‘glass’ are specific to quantifying uncountable nouns. However, not all uncountable elements can be quantified by the latter two nouns; for instance, *kīla* is unsuitable for ‘milk’, and *kās* is unsuitable for ‘soup’.

Table 8

Quantifiers classes identified through the survey. A stands for ‘acceptable’, NA stands for ‘not acceptable’

	QUANTIFIERS +	**N[-countable]**	**N[+countable]**
1	***ḥafna*** (handful)	A	A / NA
	rašfa (sip)	A	NA
	kīla (measure, portion)	A / NA	NA
	kās (glass)	A / NA	NA
2	***kaʿba*** (unity)	NA	A
	ṭuẓẓīna (dozen)	NA	A
	kamša (handful)	A / NA	A / NA
	škāra (sack)	NA	A
3	***qaṭʿa*** (piece)	A	A
	ḅākū (pack)	A	A / NA
	ṭarf (part)	A	A
	šabʿa (a lot)	A	A

The second category includes quantifiers applicable to countable nouns like eggs and apples, but not for ‘shoes’ (*ṣabbāṭ*), which has a specific quantifier *fard*, ‘pair’. For precise quantification, we already knew about *ḥāṛa* ‘4-units’, typically used for eggs, and *ṭuẓẓīna* ‘dozen’. The survey confirmed that *ṭuẓẓīna* is also used for other countable items like ‘apples’ or ‘cigarettes’. However, the survey revealed that *kaʿba*, ‘piece’, cannot be used for ‘book’ (*ktāb*).

The third includes elements usable with both types of nouns, like *qaṭʿa*, ‘piece’ or *ḅākū*, ‘packet’, *ṭarf* ‘part’ and *šabʿa* ‘a lot’. *qaṭʿa* is widely acceptable for quantifying nouns like ‘land’ or ‘cheese’, which beside being uncountable, are not collective nouns. Similarly, *ṭarf, šabʿa* and *ḅākū.* The latter is suitable for ‘milk’, commonly sold in packs, as for ‘cigarettes’, but not for ‘books’ or ‘eggs’. Instead, *šabʿa* is acceptable for ‘books’ and ‘cigarettes’, suggesting it may still be related to its lexical meaning. Similarly, *škāra* ‘sack’ (second class), is used with nouns of objects stored in sacks.

Conclusions

This article presents statistical analyses on the morphological realization of definiteness in TA encoded in Arabizi. We discussed definiteness from a semantic and syntactic perspective, focusing on TA in particular. We introduced TA data in §3 and analyzed its behavior in accordance with the observations made in §1. In §4, we described the data used for the analyses, detailing the methodology used to construct the corpus and demonstrating its value for automatic processing of TA. Our analyses in §5 were based on corpus data, and we drew conclusions from a survey that assessed the acceptability of specific sentences in TA. Further investigation is planned to explore the interconnection between definiteness and nominal quantification in TA through an additional survey.

REFERENCES

Abu-Kwaik, K., Saad, Motaz K., Chatzikyriakidis, S. & Dobnik, S. 2018. Shami: A corpus of Levantine Arabic dialects. In *Proceedings of the eleventh LREC.*

Adger, D. 2003. *Core Syntax: A minimalist approach* (Vol. 20). Oxford: Oxford University Press.

Albirini, A. 2016. *Modern Arabic sociolinguistics: Diglossia, variation, codeswitching, attitudes and identity*. Routledge.

Alghamdi, H. & Petraki, E. 2018. Arabizi in Saudi Arabia: A deviant form of language or simply a form of expression?. *Social Sciences*, *7*(9), 155.

Alsarsour, I. et al. 2018. Dart: A large dataset of dialectal Arabic tweets. In *Proceedings of the eleventh LREC*.

Al-Twairesh, N. et al. 2018. Suar: Towards building a corpus for the Saudi dialect. *Procedia computer science*, *142*. 72-82.

Androutsopoulos, J. & Schmidt, G. 2002. SMS-Kommunikation: Ethnografische Gattungsanalyse am Beispiel einer Kleingruppe. *Zeitschrift für angewandte Linguistik*, *36*. 49-80.

Baccouche, T. 2009. Tunisia. In K. Versteegh (ed.), *Encyclopedia of Arabic Language and Linguistics*, Vol. 4. 571-577.

Bernstein, J.B. 1997. Demonstratives and reinforcers in Romance and Germanic languages. *Lingua*, *102*(2-3). 87-113.

Bianchi, R.M. 2013. Arab English: The case of 3arabizi/Arabish on Mahjoob.com. *Voices in Asia Journal*, *1*(1). 82-96.

Borer, H. 1988. "On the morphological Parallelism between Compounds and Constructs." *Yearbook of Morphology 1*. 45–65.

Bošković, Ž. 2012. Phases in NPs and DPs. Phases: Developing the framework, in Ángel J. Gallego (ed.), *Phases*, 343-383. De Gruyter Mouton.

Bošković, Ž. 2020. On the coordinate structure constraint and the adjunct condition. In Bárány, T. Biberauer, J. Douglas and S. Vikner (eds.), *Syntactic architecture and its consequences II*. Berlin: Language Sciences Press.

Bouamor, H. et al. 2018. The MADAR Arabic Dialect corpus and Lexicon. In *Proceedings of the eleventh LREC.*

Brugè, L. 1996. Demonstrative movement in Spanish: A comparative approach. *Working Papers in Linguistics*, 6.1, 1996. 1-53.

Caubet, D. 2019. Vers une littératie numérique pour la darija au Maroc, une démarche collective. In *Studies on Arabic Dialectology and Sociolinguistics: Proceedings of the 12th International Conference of AIDA, Marseille, May 30th-June 2nd 2017*.

Cheng, L.L.S. & Sybesma, R. 1999. Bare and not-so-bare nouns and the structure of NP. *Linguistic inquiry*, *30*(4). 509-542.

Chomsky, N. 1998. Minimalist inquiries: the framework. *MIT occasional papers in linguistics*, *15*.

Cohen, D. 1988. Arabe. In J. Perrot (éd.), *Les langues dans le mondes ancien et moderne. Langue chamito-sémitiques*. Paris: Editions du CNRS.

D'Anna, L. 2017. *Italiano, siciliano e arabo in contatto. Profilo sociolinguistico della comunità tunisina di Mazara del Vallo*. Palermo, Centro di studi filologici e linguistici siciliani.

Diab, M. et al. 2010. COLABA: Arabic dialect annotation and processing. In *Proceedings of the seventh LREC, workshop on Semitic language processing*. 66-74.

Dobrovie-Sorin, C. 2000. (In)definiteness spread: from Romanian genitives to Hebrew construct state nominals. In Motapanyane, V. (ed.), *Comparative Studies in Romanian Syntax*, 177-226.

Edzard, L. 2006. Article, Indefinite. *Encyclopedia of Arabic Language and Linguistics, 188-191.*

El-Haj, M. 2020. Habibi-a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the eighth LREC.* 1318-1326

Gaspar, C. 2013. Ipsas kalendas/nonas-an approach to the evolutionary process of the definite article in the Iberian Peninsula. *La Variation et le Changement en Langue (Langues Romanes), 87*, 461-476.

Greenberg, J.H. 1978. How does a language acquire gender markers?. In Greenberg J.H., Ferguson Charles A. & Moravcsik E.A. (Eds.), *Universals of human language, 3*. Stanford, CA: Stanford University Press. 47–82.

Guellil, I. et al. 2019. Arabic natural language processing: an overview. *J. King Saud University – Computer and Information Sciences*. 1-3.

Gugliotta, E., & Dinarelli, M. 2020. TArC: Incrementally and semi-automatically collecting a Tunisian arabish corpus. In *LREC 2020*. pp. 6279-6286.

Gugliotta, E. et al. 2020. Multi-Task Sequence Prediction for Tunisian Arabizi Multi-Level Annotation. In *The Fifth Arabic Natural Language Processing Workshop* (WANLP). pp. 178-191).

Gugliotta, E., & Dinarelli, M. 2022. TArC: Tunisian Arabish Corpus, First complete release. In *LREC 2022*. pp. 1125-1136.

Habash, N. et al. 2018. Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the eleventh LREC*.

Hauser, M.D., Chomsky, N. & Fitch, W.T. 2002. The faculty of language: what is it, who has it, and how did it evolve? *Science*, *298*(5598). 1569-1579.

Higginbotham, J. 1985. On semantics. *Linguistic inquiry*, *16*(4). 547–593.

Hochreiter, S. & Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, *9*(8). 1735-1780.

Hoyt, F. 2008. The Arabic noun phrase. *The Encyclopedia of Arabic Language and Linguistics*. Leiden: Brill.

Jaffe, A. et al. 2012. Orthography as Social Action. *Scripts, Spelling, Identity and Power*.

Jarrar, M. et al. 2017. Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, *51*(3). 745-775.

Jiménez-Fernández, Á. L. 2012. A new look at subject islands: The phasehood of definiteness. *Anglica Wratislaviensia*, *50*. 137–168.

Leitner, B. & Procházka, S. 2021. The Polyfunctional Lexeme /fard/ in the Arabic Dialects of Iraq and Khuzestan: More than an Indefinite Article, *Brill's Journal of Afroasiatic Languages and Linguistics* 13(2). 143–186.

Longobardi, G. 1994. Reference and proper names: A theory of N-movement in syntax and logical form. *Linguistic inquiry*. 609–665.

Longobardi, G. 2008. Reference to individuals, person, and the variety of mapping parameters. *Essays on nominal determination: From morphology to discourse management*. 189-211.

Lyons, C. 1999. *Definiteness*. Cambridge University Press.

Maamouri, M. et al. 2004. The Penn Arabic Treebank: Building a Large-scale Annotated Arabic Corpus. In *NEMLAR,* Vol. 27. 466-467).

Maamouri, M. et al. 2007. Fisher Levantine Arabic Conversational Telephone Speech. *Linguistic Data Consortium, University of Pennsylvania, LDC Catalog No.: LDC2007S02.*

Maamouri, M. et al. 2009. Penn Arabic Treebank Guidelines. *Linguistic Data Consortium*.

Maamouri, M. et al. 2014. Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In *Proceedings of the ninth LREC.* 2348-2354.

Marçais, P. 1952. *Le Parler arabe de Djidjelli.* PhD dissertation, Université de Paris. Paris, Lib. d'Amérique et d'Orient-Adrien-Maisonneuve.

Marçais, W. 1950. Les parlers arabes. *Initiation à la Tunisie*. 195-219.

Marçais W. 1961. Comment l'Afrique du Nord a été arabisée. In W. Marçais, *Articles et conférences*. Paris: Adrien Maisonneuve. 171-192.

Massaro, A. 2022. Romance genitives: agreement, definiteness, and phases. Transactions of the Philological Society, 120(1), 85-102.

Matushansky, O. 2006. Why Rose is the Rose: On the use of definite articles in proper names. *Empirical issues in syntax and semantics, 6,* 285-307.

Mensching, G. 2005. Remarks on specificity and related categories in Sardinian. In Klaus von Heusinger, Georg A. Kaiser & Elisabeth Stark (eds.*), Proceedings of the Workshop: Specificity and the Evolution/Emergence of Nominal Determination Systems in Romance*. U. Konstanz, 81-106.

Mion, G. 2009 « L'indétermination nominale dans les dialectes arabes : une vue d'ensemble », in A. Arioli (ed.), *Miscellanea Arabica 2009,* Nuova Cultura, Roma, pp. 215-231.

Mion, G. 2014 Éléments de description de l'arabe parlé à Mateur (Tunisie). *Al-Andalus Magreb*, 21, 57-77.

Mohammad, M.A. 1999. Checking and licensing inside DP in Palestinian Arabic. *Amsterdam Studies in the Theory and History of Linguistic Science series 4*. 27-44.

Pasha, A. et al. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the ninth LREC.* 1094-1101.

Ramchand, G. & Svenonius, P. 2008. Mapping a parochial lexicon onto a universal semantics. *The limits of syntactic variation*. 219-245.

Salem, F. 2017. 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.*

Shormani, M.Q. 2016. Are noun phrases phases? Evidence from Semitic construct state. *International Journal of Arabic Linguistics*, *2*(2). 96–132.

Sutskever, I., Vinyals, O. & Le, Q.V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104-3112.

Takezawa, T. et al. 2007. Multilingual spoken language corpus development for communication research. In *ISCSLP 2006*. 303-324.

Turner, M.L. 2018. *Definiteness in the Arabic Dialects*. Ph.D. dissertation, The University of Texas at Austin, Texas, USA.

Turner M.L. 2021. Definiteness Systems and Dialect Classification. *Languages* 6(3). 128.

Wiltschko, M. (2009). What's in a determiner, and how did it get there?. *Determiners: Universals and variation*, 147, 25.

Younes, J. & Souissi, E. 2014. A quantitative view of Tunisian dialect electronic writing. In *5th International Conference on Arabic Language Processing*. 63-72.

Zaborski, Andrzej. 2006. Article, Definite. Ed. by Kees Versteegh, *Encyclopedia of Arabic Language and Linguistics 1*, Leiden. 187-188.