LDK 2023

Towards a Unified Digital Resource for Tunisian Arabic Lexicography

Elisa Gugliotta[1] and **Michele Mallia**[1] and **Livia Panascì**[2*]
1. Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche;
2. *Sapienza* University of Rome
1. {firstname.lastname}@ilc.cnr.it, 2. livia.panasc@outlook.it

Abstract

This paper presents our work on linking language tools for Tunisian Arabic, focusing on a lexicographic database and a corpus of informal written texts. This work on Tunisian Arabic is an ongoing pilot study, while our wider goal is to create resources for various under-resourced languages. We outline a methodology that emphasises open science principles, leveraging existing language resources and NLP tools for standardisation and annotation. Our approach ensures reproducibility and benefits other researchers. We share annotated data on a digital platform and release NLP tools on a dedicated repository. Our work aligns with FAIR principles, facilitating open and effective research on under-resourced languages.

1 Introduction

This paper describes a research methodology for the study of under-resourced languages, presenting it through the exemplification of a pilot study we are conducting on Tunisian Arabic dialect (TA). Therefore, the work is part of a wider project aiming at supporting studies on under-resourced languages using both quantitative research methods, such as statistical analysis and Deep Learning techniques, and qualitative research methods, such as Linguistics and Dialectology. The lack of computational resources, such as annotated corpora, language models, and digital lexicons, to name a few, has been a major roadblock to the processing of under-resourced languages. Usually, these languages have a poor tradition of linguistic studies: to a few ancient written sources correspond few analyses on lexicography, morphology, phonetics, etc. Moreover, it lacks communication between scientific sectors: different research areas, such as Digital Humanities and Dialectology, hardly converge and collaborate in the study of under-resourced languages. Consequently, the studies that have been carried out remain isolated and underexploited. On the contrary, only a comprehensive approach can reflect the dynamism and complexity of a language, by preserving the quality of linguistic data at all stages of data processing, from identification and selection, collection, pre-processing, processing, analysis, annotation and data fruition. For what concerns Arabic dialects, i.e. Colloquial Arabic (CA), to which TA belongs, the limited availability of data is one of the main reasons why these varieties are still defined as under-resourced.[1] At the same time, the specificity of the multilingual realities of the Arab countries, with special reference to the diglossic situation,[2] makes building corpora of CA a challenge. CA has always been a predominantly oral language, very few written texts have been recorded and texts prior to the 20th century are extremely rare.[3] There is no standardised writing system, the studies that have been conducted so far have often focused on specific aspects of the language and have almost never been connected with each other. Linguistic research that has been conducted in the past often did not respect strict methodological criteria (for example, not reporting the number of informants, their age, or geographical origin). It is for all these reasons that, although in the last decades the building of linguistic corpora for Arabic has incredibly increased (Darwish et al., 2021) and although a number of CA corpora has recently been released (see Section 3.2), these corpora cannot support wide linguistic analysis.

Therefore, our project, whose ultimate goal is to connect and make linguistic data on under-resourced languages easily available by users, has as its first step the data collection. To collect

[*] All three authors collaborated on the project. For academic purposes, E. Gugliotta is responsible for sections 2, 3.2, 4.2, 5.2, 6; M. Mallia for sections 2 (Step 3), 5.1; L. Panascì for sections 1, 3.1, 4.1, 5 (introduction).

[1] For details on the causes that lead some languages to be defined as under-resourced, see Pretorius and Soria (2017).

[2] See Ferguson (1959); Versteegh (2014); Owens (2006); Abboud-Haggar (2006); Sayahi (2014).

[3] The CA literature is really rare: see Davies (2006).

data, we exploit existing resources, i.e. ancient (dialectological sources from the 19th century to the present) and modern (corpora of authentic written TA), which, although originally created for very different purposes, come together to present more complete and detailed data possible.[4]

In Section 2 we present the main aims of our project, while in Section 3 we start reporting on the pilot study, by outlining different kinds of work and data available for TA. In Section 4, we describe the linguistic resources employed for our study (a lexicographic database TA-Italian and vice versa and a TA corpus). These were previously created for specific purposes, that we are currently normalising in terms of content and format standardisation. Such data will be released through a digital platform aimed at providing access to linguistic information and facilitating complex queries, which would undoubtedly be a milestone in this domain. At the same time, computational tools built to process these data will be made available through a dedicated repository.[5] In Section 5 we outline the project methodology stages applied to the pilot study so far. Indeed, our ultimate goal is to unify a big amount of TA data (described in Section 4), to be employed for future studies, in different fields (NLP, Digital Humanities, Linguistics and Dialectology).[6] With this aim, we devised a methodology inspired by the principles of the data economy, sustainability of research and the FAIR principles of open science.[7] Finally, in Section 6, we discuss our conclusions and future works.

2 General Project Aims

The macro-objective of this project is to develop and put into practice a hybrid methodology that could strongly contribute to the current state of research on under-resourced languages, starting from Arabic dialects. Following open science principles, the methodology aligns with transparency, collaboration, and accessibility. Such methodology is organized in three steps. In Step 1, existing linguistic resources are compiled using freely available tools, corpora, glossaries, and dictionaries from the scientific community, promoting openness. The work of Step 2 adheres to open science principles. In fact, text standardization and annotation are realised by using NLP tools. This enables work reproducibility and allows other researchers to exploit our tools and methodology. In Step 3, annotated data and NLP tools are provided, emphasizing open data. Overall, the methodology adheres to the FAIR principles (Wilkinson et al., 2016; De Jong et al., 2018), promoting Findability, Accessibility, Interoperability, and Reusability of linguistic resources and data, facilitating open and effective research on under-resourced languages.[8] Since our ultimate goal is to advance research on different under-resourced languages, at the end of Step 3 there is a recursive cycle to start the process again (Step 1) with a new under-resourced language or language variety.

Step 1. *Resource Compilation: Economizing Data*. This first work stage is based on the concept of 'data economy' rather than 'creation from scratch'. It aims to identify existing linguistic tools, corpora, glossaries, and dictionaries available among the scientific community in various formats and for different purposes. Such resources are often underutilized after their initial creation and use (Macchiarelli, 2023). This is because, once used for the purposes for which they were created, they are not maintained, extended, or adapted to standards that would allow their use by audiences other than those imagined at the time of their creation (Pretorius and Soria, 2017). We will use any available resources that we become aware of, such as resources created for other purposes, like corpora created for sentiment analysis, which perhaps do not have fine-grained grammatical annotations. We will be in charge of the annotation of these data. Our first objective is to retrieve these resources, promoting data sustainability, and standardise them into a unified format (Step 2).

Step 2. *Standardisation and Annotation: Enhancing Linguistic Insights*. This stage also includes text normalisation and the semi-automatic annotation of linguistic features is done using existing tools. Text normalisation ensures consistency and prepares the text for subsequent processing. In the analysis of under-resourced language data, we consider morpho-syntactic information crucial for disambiguating semantically challenging elements extracted from the production context (Jarrar et al., 2022; Nahli et al., 2023). For this reason, we train (and release at the end of Step 3) morphological embeddings for each language (Cotterell and Schütze,

[4] See Section 4 for linguistic resources description.

[5] At this link: https://github.com/LinguaeVerse.

[6] About cooperation, use, sustainability of language data in these fields, see Fišer and Witt (2022).

[7] See Section 2 for further details on these topics.

[8] For further information on the FAIR principles, please see https://www.go-fair.org/fair-principles/.

2015).[9] To produce morpho-syntactic annotations we can exploit existing tools, such as a Multi-Task architecture created for TA data annotation (Gugliotta et al., 2020). Such an architecture can learn linguistic insights from small, noisy data (Gugliotta and Dinarelli, 2023). Thus, it can be useful for processing multiple varieties of CA, starting with the varieties most similar to TA (the target language of our pilot study), such as the North African varieties.

Step 3. *Providing Data: Enabling Further Studies.* Finally, the last work stage focuses on providing annotated data to support further studies in this direction. The annotated data will be available through a digital platform that supports queries from researchers interested in linguistic and lexicographic studies on the collected texts. This, together with the release of annotated data and pre-trained morphological embeddings, could greatly facilitate the preservation and digital accessibility of these languages, thereby fostering cultural and linguistic diversity in the digital world.

On morphological embeddings. In this phase, we investigate the incorporation of morphological knowledge in word embeddings, to capture semantic and morphological similarities. Training such embeddings for the under-studied language would have several utilities. They would ease the annotation of additional data; they would help in lexical and ontological modeling of the language resources underlying the digital platform (see below). Finally, we could release a tool with great potential, which under-resourced languages generally lack, and which we could easily investigate from the data annotated in Step 2. After an initial phase of evaluating the available models (see Sezerer and Tekir, 2021), we will train on the already annotated data a model capable of generating embeddings combining morphemes, POS-tags and lemmas.[10]

Concerning our pilot study on TA, Yagi et al. (2022), shows that the evaluation metrics for Arabic embedding models need to take into consideration the morphological characteristics of the language. Moreover, Salama et al. (2018) emphasize the incorporation of morphological analysis in the training of word embedding models, given the morphological complexity of the Arabic language. The drive to exploit word embeddings for Arabic NLP has been matched by efforts to annotate Arabic texts with Linked Data. Bouziane et al. (2020) present a comprehensive framework for annotating Arabic texts with Linked Data. This kind of annotated data becomes a precious resource for training more sophisticated NLP models, contributing to the larger goal of making CA texts more accessible, less ambiguous, and more useful in various NLP applications, such as information retrieval, word sense disambiguation and other related areas.

On the digital platform. Such a platform is intended not only as a tool for conducting queries but also as an aggregator of information, particularly focusing on under-resourced languages. One of the salient features of the platform will be its capacity to perform complex queries through data correlation. This is essential for extracting nuanced information and recognizing patterns within the data (Alhafi et al., 2019). By enabling users to create complex queries that integrate data from multiple sources, the platform facilitates simultaneous analysis of the two data sources (querying both via the central Analysis Node, see Figure 1). This advanced capability helps researchers derive more meaningful insights by leveraging the combined power of integrated data.

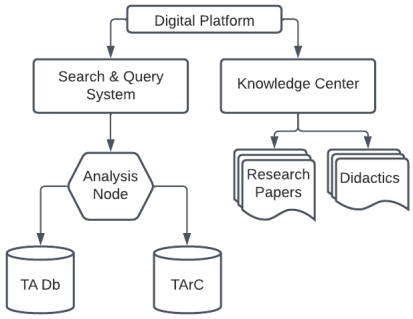

Figure 1: The digital platform general structure

To understand the type of digital platform we plan to implement, we refer to similar work on Arabic language, this is the one of Jarrar and Amayreh (2019). This lexicographic search engine is constructed atop the most extensive Arabic multilingual database, facilitating users in searching and retrieving translations, synonyms, definitions, and more.[11] Similar to this work, our platform will be developed with cutting-edge features and in alignment with the

[9] Morphological embeddings are numerical representations of morphemes or morphological units in a language, embedded in a continuous vector space (Cotterell and Schütze, 2015). For completeness, see also Bengio et al. (2003). For further information on morphological embeddings, please see below.

[10] Using morphemes for word embeddings in morphologically rich languages is useful to encode more semantic information (Romanov and Khusainova, 2019).

[11] The search engine can be accessed at https://ontology.birzeit.edu.

recommendations and best practices of the World Wide Web Consortium (W3C) for publishing data on the web. Additionally, our digital platform will serve as a comprehensive repository, aggregating diverse types of information related to the study of the under-studied language. It will encompass a wide range of resources such as recipes, travel blogs, and other existing information on the under-studied language. By incorporating this diverse information, our platform is intended to provide a holistic and rich source of data for researchers and others interested in discovering languages and cultures. Furthermore, with the texts and information collected on our platform, it will be possible to develop teaching materials based on authentic data (*Didactics* in Figure 1). Regarding the *Analysis Node*, in Figure 1, this module is understood as the one in which the matching process between the data collected in the two instruments is performed. In the case of the TA data, this process will be based on the *root* level information.[12]. Moreover, the platform will adhere to the W3C's OntoLex-Lemon RDF model,[13] emphasizing our dedication to ensuring standardisation and interoperability.

After Step 3: Milestones and Takeaways. This methodology can be applied to different languages, allowing the expansion of research and application of the results obtained. By repeating these three steps for different languages or language varieties, it is possible to extend the application of the hybrid methodology and advance research in a wide range of language contexts with scarce resources. This cycle helps to create a sustainable data ecosystem and improve linguistic knowledge for under-resourced languages.

3 Tunisian Arabic State-of-the-Art

This section presents the state-of-the-art of digital and non-digital resources available for TA, the subject of our pilot study.

[12]See the subsections 5.1 and 5.2 for more information about the root level.

[13]Resource Description Framework (RDF) is a standard model for data interchange on the web. It allows for the integration of various sources with different structures and makes it easier for machines to understand the semantics of the information. Lemon (Lexicon Model for Ontologies) is a model based on RDF and designed for representing lexical information relative to ontologies. It allows for the representation of a wide range of linguistic structures necessary for the development of NLP applications. https://www.w3.org/2016/05/ontolex/.

3.1 Available Non-digital Resources

As mentioned above, dealing with Arabic dialects means having access to a very limited number of written sources. In fact, mainly for identity reasons, Arab speakers normally have a strong hierarchical perception of the languages they speak: on the one hand, Standard and Koranic Arabic represent the high register of the language, used in written texts and in formal and non-spontaneous situations; on the other hand, dialect is perceived as a lower register, sometimes even vulgar, and it is the language of everyday life, spontaneity and orality (Boussofara-Omar, 2006). From this, it clearly follows that, over the centuries, the documents which had to be preserved and which deserved the written form, were essentially composed in the highest register of the diglottic *continuum*, i.e. in Koranic/ Standard/ Literary Arabic. However, Arabs have always had the local dialect as native language, and have always expressed themselves orally in this variety. As a consequence, there are very few written sources that report ancient dialect lexicon, linguistic traces of which are mostly found in the phenomena of loan and interference and in Middle Arabic (an intermediate variety product of the interference of the Modern Standard Arabic (MSA) and the CA[14]). In short, this means that as far as Arabic dialects are concerned, and specifically TA, it is virtually impossible to have access to primary sources prior to the 21st century. It was only in the contemporary era that Arabic dialects started to be used in digital informal communication (Caubet, 2019), providing the first appearance of sizable linguistic data of CA. However, evidences of a previous linguistic stage is found in dialectological studies, mostly performed by European researchers, starting from the 19th century. Among them, there are the works included in the lexicographic database which will be described extensively in Section 4.1. To cite some of the works that can be considered sources of TA lexicon prior to the current period, we can mention pioneering studies such as the Maghrebi (i.e. North African) Arabic dictionary by Beaussier et al. (2006), the TA grammar and glossary by Stumme (1896) and the impressive description of Takrouna's Arabic by Marçais (1961). It is also necessary to mention dictionaries and manuals

[14]Middle Arabic is described in more detail by Lentin (2008, 216) as 'the language of numerous Arabic texts distinguished by its linguistically (and therefore stylistically) mixed nature, as it combines standard and colloquial features with others of a third type, neither standard nor colloquial'.

for French students published in the early 20th century (such as, for example, the works of Nicolas ((s.d.); Jourdan (1913)). These pioneering studies represent almost the only evidence of linguistic stage that otherwise would have been forgotten. But precisely because they are forerunners, all these studies present various problems: e.g. it is sometimes not clear which linguistic variety they refer to and they do not always use accurate transcriptions of CA phonetics. For this reason, it is necessary to compare them with further sources: more recent and accurate dialectological studies (such as Behnstedt (1998, 1999); Ritt-Benmimoun (2014)), manuals for foreign students published in recent years (such as: Ben Ammar and Vacchiani (2016); Durand and Tarquini (2023)) but also, and above all, with primary sources, i.e. interviews on field and authentic exchanges in social networks.

3.2 Available Digital Resources

Concerning digital platforms for dictionaries or lexicons of TA, to the best of our knowledge, there are only the *Linguistic dynamics in the Greater Tunis Area: a corpus-based approach* (TUNICO) (Dallaji et al., 2020) and the *Tunisian Arabic Corpus* (TAC) (McNeil, 2018).[15] The first makes available through a digital platform a Tunisian dictionary and a corpus of data associated with accurate linguistic information. TUNICO data are encoded in a Latin-based transcription and can be searched using a search bar. Instead, TAC collects raw texts, encoded in not-normalised Arabic script. TAC texts can be observed by search queries based on three different systems: *Exact*, *Stem*, and *Regex*. The first two require an Arabic-encoded input, while the third one requires the users to transliterate the input by following a modified version of the Buckwalter transliteration system.[16] These tools are useful for language analysis, although they present some difficulties in their use. With regard to the processing and the study of CA in the NLP field, there is a trend in recent years to produce a multitude of CA corpora that has allowed for progress in the study of CAs. In the case of TA, among the various recently released corpora we can mention a corpus of Facebook comments, manually annotated for sentiment analysis (TSAC) (Mdhaffar et al., 2017) and a parallel corpus of TA-MSA, the TD-COM corpus, extracted from social networks (Kchaou et al., 2022).[17] Another downloadable corpus for TA is the Tunisian Arabizi Corpus (TArC), released by Gugliotta and Dinarelli (2022) and described in Section 4.2. Finally, we should mention some multi-dialectal resources that include TA among other CA varieties. One of these is PADIC (Meftouh et al., 2018), a parallel corpus of six CAs. Another one is MADAR (Bouamor et al., 2014), which consists of a parallel corpus of the CA of 25 Arab cities, including cities of Tunisia (Tunis and Sfax). The same corpus has recently been released in CODA orthography (Habash et al., 2018) by Eryani et al. (2020).

Although a number of corpora have been produced, TA is still considered an under-resourced language. It is possible that the solution to the complexity of CA (morphological and orthographic, due to the absence of standards and a situation of multilingualism, diglossia, etc.), does not lie solely in the amount of data, processed according to universally valid methodologies for all languages. As a very simple example, each of the mentioned resources was created for a specific purpose and consequently represents a portion of the linguistic reality of TA. These are indeed valuable resources, but not sufficient for a complete mapping of this language. Moreover, each resource, including TUNICO and TAC, presents its own language encoding system, based on Latin or Arabic script. Perhaps there is a need to develop a methodology suited to the case of under-resourced languages and thus aim more than ever to preserve data quality. In the next section, we will explain how our contribution attempts to investigate this possibility.

4 Linguistic Resources Description

4.1 The TA Lexicographic Database

TA is a rich and composite language, which fully reflects the history and culture of a country located in the center of southern Mediterranean cost, known since ancient times as a land of human as well as linguistic passage and exchange (Marçais, 1950; Baccouche, 2009). TA has a varied lexical composition, due to the coexistence of a main Arabic linguistic stratum (Hilali, pre-Hilali and Classical Arabic); adstrate languages (such

[15] See also: https://www.livelingua.com/arabic/courses/tunisian and https://derja.ninja/.

[16] Further information on TAC query system at page: https://www.tunisiya.org/help/. Buckwalter transliteration system at http://www.qamus.org/transliteration.htm.

[17] Other resources, released by the same Arabic NLP group, are available at https://sites.google.com/site/anlprg/corpora-corpus?authuser=0.

as Berber, Punic, Greek, and Latin) and many superstrate languages (such as Spanish, Lingua Franca[18], Turkish, Italian, French and English).[19] In addition, all these elements are combined with diglossia (with Standard Arabic) and bilingualism (with French).[20] In order to record at least a part of the lexical richness of TA and attempt linguistic analysis, it was first of all necessary to create a tool for registering the lexicon available in the TA bibliographic sources: this tool is the TA lexicographic database (Panascì, 2021), consisting of 13,800 headwords and 5,600 Arabic roots and focused on diachronic and diatopic variation in the TA lexicon. To date, the database collects all the lexical entries of ten glossaries, two papers and three dictionaries[21] representing about a century and a half of Tunisian linguistic history and various local dialects. The oldest source is in fact a grammar written in 1896 (Stumme) and the most recent one is a 2017 paper on Tunis jargon (Labidi). Moreover, the database contains dialects representative of various areas of the country, such as the dialect of the capital, Tunis (Ben Ammar and Vacchiani, 2016), that of a coastal city such as Susa (Talmoudi, 1981), or a Bedouin dialect of the South of the country, such as that of the Marazig tribe (Boris, 1958). To build the lexicographic database, all headwords have been translated into Italian and they have been marked with an abbreviation designating the reference source of the entry. The individual words referring to a specific meaning were compared with each other, adopting a criterion that highlighted the diachronic evolution of the language (that is, an insertion of the occurrences in the sources from the oldest to the most modern). To make the material more enjoyable for the reader, it has been organized in the structure of an Italian-TA dictionary, i.e. with the entries inserted in alphabetical order, as well as in the structure of a TA-Italian dictionary, i.e. according to the traditional Arabic language setting of radical letters. Finally, the database entries present additional information (when available): etymology of the word, diatopic collocation, semantic shifts, obsolescences, linguistic register, etc. Below are two examples of entries, the first one taken from the Italian-Tunisian database, the second one from the Tunisian-Italian database.

Camaleonte s.m. *omm əl-buya* AN11/ *umm əl-būyya* JJ13; *bu kɛššēš* GB58; *bu ḥremba* [dim. *bu ḥrēmba*] GB58; (Maġārba, ai confini tra Tripolitania e Cirenaica e Warfella, a Ovest dei Maġārba) *ḥerba* GB58; (Warġemma, confederazione tribale tra Gabès e Médénine, e Rbāye‘, nomadi della zona del Oued Ṣūf) *ḥerbēya* GB58; *tata* MQ2002

حنك *ḥank* [pl. *ḥnāk*] MH77 **palato;** (pan-maghrebino) *ḥanek* [pl. *aḥnāk-*; + art. *l-ḥank/ laḥnek*, pl. *laḥnāk-*; + pron. suff. *ḥanki*] JQ61/ (pan-maghrebino) *ḥnak* [pl. *aḥnāk-*; + art. *l-ḥank/ laḥnek*, pl. *laḥnāk-*; + pron. suff. *ḥanki*] JQ61/ *ḥank* [pl. *ḥnāk*] MH77 **mascella (umana); guancia;** *ḥ^{a}nɛk* [pl. *aḥnēk*; + pron. suff. I pers. sing. *ḥen^{e}ki*; + pron. suff. III pers. f. sing. *ḥ^{a}nɛkha*] GB58 **mandibola** – *daqq əl-ḥank* JQ62 **idiom. parlare di futilità; straparlare;** *ḥanka* JQ62 **esperienza di vita;** *məṭḥannek* JQ62 **esperto**

Figure 2: TA Lexicographic Database Sample

Figure 2 shows how the database works. In the first case, all the occurrences for the meaning of "chameleon" in the various sources are reported. The entries are followed by the reference abbreviation (e.g. AN11 represents (Nicolas, (s.d.)) and they are in chronological order. The diatopic variation is highlighted (e.g. the lexical variants for the term in the different tribes of southern Tunisia are specified). In the second case, instead, all the occurrences found in the sources for the Arabic root *ḥnk* are reported. The order of appearance of the terms is the traditional one of Arabic dictionaries (first the ten forms of the verb appear, then the nouns, etc.). In this case the geographical location of a term (the word for "jaw" or "cheek") is highlighted and an example of an idiomatic expression is given.

4.2 Tunisian Arabizi Corpus (TArC)

TArC gathers texts from various informal digital writing contexts, such as blogs, forums, and Facebook, including rap song lyrics shared on dedicated forums. The collection of these texts aims to investigate Arabizi, a Latin script encoding used in informal online communication. Additionally, the inclusion of rap song lyrics allows for a comparative analysis of both the Arabic and Latin script encoding systems in TA.[22] Together with the texts, were publicly available, also some metadata of the authors

[18] With Lingua Franca we refer to the Italian-based pidgin spoken in the regencies of Tunis, Tripoli and Algiers during the Ottoman rule (Cifoletti, 2004).

[19] See: Baccouche (1994).

[20] See Daoud (2007).

[21] The TA lexicographic database sources include Ben Abdelkader et al. (1977); Ben Alaya and Quitout (2010); Ben Ammar and Vacchiani (2016); Bevacqua (2008); Boris (1958); Jourdan (1913); Labidi (2017); Marçais and Hamrouni (1977); Nicolas ((s.d.); Quéméneur (1961a,b, 1962); Quitout (2002); Stumme (1896); Talmoudi (1981).

[22] TArC data are available at https://github.com/eligugliotta/tarc.

of texts were collected. These are their provenience, age-range and gender (Gugliotta, 2022).

TArC data have been semi-automatically annotated with various linguistic information at word-level, by means of a neural Multi-Task Architecture (MTA) (Gugliotta et al., 2020).[23] These annotation levels are shown in Table 2 and consist of text normalisation into CODA-*Star* orthography in Arabic script (Habash et al., 2018), sub-tokenisation, POS-tagging and lemmatisation. To avoid transliterating code-switching into Arabic script, the initial annotation level of TArC data is token classification, which, as shown in Table 1, consists of three classes: *Foreign*, *Arabizi* and *Emotag*. The *Emotag* class encompasses para-textual elements like emoticons and smileys that are not intended for transliteration. Only the tokens classified as *Arabizi* have been annotated with the linguistic information. The formalism employed for Part-of-Speech tagging is the one of the Penn Arabic Treebank (Maamouri et al., 2004), while lemmas are also encoded in CODA-*Star*. Below we report some information on TArC data.

	TArC - *Total of Lemmas: 5,063*				
Token Class	**Blogs**	**Forums**	**Facebook**	**Rap**	*Total*
Arabizi	5,978	6,026	11,833	7,680	31,517
Foreign	707	5,873	3,624	1,010	11,214
Emotag	7	10	600	1	618
Tokens	6,692	11,909	16,057	8,691	**43,349**
Sentences	366	755	3,162	515	**4,798**

Table 1: *The Tunisian Arabizi Corpus*

5 Resources Integration

The two linguistic tools described in the previous section, despite having the same variety of CA as their subject, namely TA, are very different. It is precisely in their diversity that their complementarity and the usefulness of their combination lies. In fact, the lexicographic database was created to observe the variation of TA at the diachronic and diatopic level, thus, it mainly collects lemmas through secondary sources. Instead, TArC collects authentic texts encoded in a non-standardised writing system, known as Arabizi. This is shown in Example 1, where the first line consists of the original text in Arabizi encoding; the second line is the transcription of the oral reconstruction of the same sentence; and the third line is its translation. This sentence, in TArC is provided with the annotation levels shown in Table 2, where the sentence is reported in Arabic script (normalisation in CODA-*Star*), in the first column. In the following columns, we can observe how the sentence has been processed at the sub-tokenisation, POS-tagging and lemmatisation levels.

(1) *Tdaweb zebda wtzidha lil farina*
/t-ðawːəb əz-zəbda w-t-zīd-hā l-əl fārīna/
'Melt the butter and mix it with the flour'.

CODA	**Tokeniz.**	**POS**	**Lemma**
تذوّب	تذوّب	CV2S-CV	ذوّب
الزبدة	الـبزبدة	DET+NOUN-NSUFF_FEM_SG	زبدة
وتزيدها	و+تزيد+ها	CONJ+CV2S-CV+ CVSUFF_DO:3FS	زاد
لال	لـبال	PREP+DET	ل
فارينة	فارينة	NOUN-NSUFF_FEM_SG	فارينة

Table 2: TArC Annotation Levels

The lexicographic database provides specific information about individual entries (always in the lemmatic form): diatopic and diachronic variation, etymology, semantic changes, etc. In order to give an excerpt of them, we report in the following example, the information collected at the voice */fārīna/* 'flour'.[24]

(2) **Flour** s.f. [< ita. or lingua franca *farina*] *fɛ̄rīna* HS1896/ (2) [coll. *fɛ̄rīnɛ̄*] BAR77/ *farina* AW2010/ *fɛ̄rīna* AV2016; *dqīq* AN11/ *dqīq* JJ13/ *dqīq* MH77/ *dqēq* MH77 – fine flour *degīg* GB58 – flour (probably of soft wheat) purchased already ground *fạ̄rīnɛ̄* GB58 – idiom. "add water, add flour…" (phrase to be used during an anecdotal narrative, signifying that it was a never-ending enterprise) *zīd əl-ma zīd əd-dqīq* MH77

From Example 2, it is possible to see how, from the nineteenth century to the present, the concept is mainly expressed by a loanword from Italian (Stumme, 1896) or from the Lingua Franca (Cifoletti, 2004, 234) : *fārīna*. The loanword would seem to have supplanted the Arabic *dqīq*, although we find the latter in some of the database's supplies, both with the meaning of 'flour' and 'fine flour'.

[23] This is available at https://gricad-gitlab.univ-grenoble-alpes.fr/dinarelm/tarc-multi-task-system.

[24] The abbreviations in order are: HS1896: (Stumme, 1896); BAR77: (Ben Abdelkader et al., 1977); AW2010: (Ben Alaya and Quitout, 2010); AV2016: (Ben Ammar and Vacchiani, 2016); AN11: (Nicolas, (s.d.); JJ13: (Jourdan, 1913); MH77: (Marçais and Hamrouni, 1977); GB58: (Boris, 1958).

The database thus allows hypotheses to be made: most likely the two terms must have coexisted for a long time (Stumme in the late 19th century recorded *fārīna* for Tunis; Nicolas and Jourdan in the early 20th century reported only *dqīq*), perhaps as diatopic variants or perhaps with specialization of meaning, as was the case in the 1950s in Marazig speech,[25] in which *fārīna* was merely the product of soft wheat already ground, and as reconstructed by Cifoletti (1998, 152) for Tunis, where with the entry of the loanword into common parlance, *dqīq* came to mean 'semolina'. Finally, the database (MH77: (Marçais and Hamrouni, 1977)) provides an idiomatic expression related to the concept of 'flour': *zīd əl-ma zīd əd-dqīq*.

From these examples, we can clearly see how the integration of these two resources can yield a tool that is unique in its completeness. In fact, together they can provide lexicographic, etymological, diachronic and diatopic information plus examples from real native usage occurrences and morpho-syntactic information of such sentences. In the following section, we explain how we were able to link the information of these tools.

5.1 Analysis and Conversion of Lexicographic Data structure

In the context of this research project focused on the management of under-resourced Arabic dialects, we elected to devise and implement a scraping tool specifically designed to delve into a dictionary's intricacies, extract pertinent data, and utilize this information for subsequent linguistic analyses and potential cross-referencing with other linguistic data sets. This decision stemmed from the realization of the untapped potential housed within these lexicographic structures, often layered and dense with information but largely inaccessible due to their static presentation. To accomplish this ambitious task, we deployed a carefully constructed script that meticulously parsed the dictionary, illuminating its structure on an entry-by-entry basis. The cornerstone of our process was a .docx file, the format of the lexicographic database. The document was formatted according to specific standards that allowed us to codify a system of rules for data extraction, rules contingent on the elements' location within each entry. The algorithm's cornerstone was the identification and extraction of the Italian definition within each entry, typically represented as a distinct bold string. Once this key piece of information was located, the algorithm triggered a systematic reverse sequence search designed to uncover other elements. This exploratory process, proceeding backwards from the definition, focused on locating: 1) the source reference indicating the individual or group responsible for proposing the hypothesis; 2) any enclosed morphological information presented within square brackets (see Figure 2). This could include TA variants trailed by morpho-syntactic data such as part-of-speech and further grammatical information; 3) As shown in Figure 2, the TA lemma tethered to the root, which is encoded in Arabic characters. Instead, the lemma, a central component of each entry, is rendered in italics with specific unicode characters. Furthermore, it's noteworthy that multiple variants can be linked to a single semantic interpretation within this structure. Upon extraction, the raw data underwent a transformation process designed to adapt it into a data structure capable of reflecting the inherent relationship and interlinking between disparate elements dispersed across the corpus. This was a vital aspect of the project as we frequently encountered references to other dictionary entries and cross-references that needed to be retained to maintain the richness of the dataset. Given the nature of


```
{
  "root": "شلشل",
  "definitions": [
    {
      "meaning": "casco spogliato della
maggior parte dei suoi datteri",
      "occurrences": [
        {
          "lemma": "šɛlšūl",
          "source": "GB58",
          "variations": [],
          "additional_data": [
            {
              "text": "pl. šalāšīl"
            }
          ]
        }
      ],
      "examples": [
        {
          "tun": "wągətha kunət sɛ̄reḥ ‘ala
šɛlšūl, hāk el‘ebsɛ ",
          "source": "GB58",
          "ita": "idiom. all’epoca ero pastore
per il conte di Šalšūl, quel tirchio (modo di
dire per designare un avaro);"
        }
      ]
    }
  ],
  "references": []
}
```


Figure 3: A TA dictionary entry encoded in JSON

[25] Recorded in the dictionary of Boris (1958), corresponding to the abbreviation GB58.

the source document and the complexities involved in the extraction process, it was inevitable that we would encounter a certain degree of noise within the data. This noise could manifest as characters not belonging to the target alphabet, misplaced punctuation marks, or other elements that deviated from the expected data type. To address these issues, we developed a series of rules using regular expressions, specifically designed to identify and control such anomalies, effectively cleansing the dataset.

The result of this comprehensive process was a script capable of extracting a substantial volume of data from the source dictionary. Nevertheless, we acknowledge that a completely automated process remains elusive due to the possibility of errors and irregularities inherent in the data. Consequently, a degree of manual data cleansing is still necessary. For instance, it's not uncommon to encounter text segments belonging to another lemma embedded within a definition, a complication arising from inconsistencies in formatting. While our script currently lacks the functionality to extract or classify morpho-syntactic categories or the etymological and additional information often found within dictionary entries, we view these as areas for future development rather than limitations. We are actively working on enhancements designed to incorporate these elements into the script, thereby adding another layer of richness to the extracted data. As we continue to refine and develop this tool, our focus is shifting toward addressing the broader challenges associated with data extraction for the creation of accessible and interoperable lexical resources. This ongoing endeavor aligns with our commitment to the FAIR principles. By enhancing our capacity to extract and utilize the rich data contained within lexicographic resources, we believe we can significantly contribute to the field of under-resourced language studies.

5.2 Corpus Annotation extension

Considering the different encoding employed for the level of lemmatisation of the two tools (scientific transliteration for the lexicographic database and normalisation in CODA-Star for TArC), we discarded lemmas as a common key between the two tools to be put into communication. Since, on the other hand, the lexicographic database is provided with an annotation level of the root from which the recorded lemma is derived, it decided to use the root as the first key element for joining the linguistic tools. To produce this additional annotation layer, we investigated the functionality of the CAMeL Tools (Obeid et al., 2020).[26] This is a suite of Arabic NLP tools, such as lemmatisers, tokenisers and POS-tagger, and provides also roots. However, among the databases provided with CAMeL Tools (MSA, Egyptian Arabic and Gulf Arabic databases), only the database for the MSA, according to our tests, provides roots. Annotating the Tunisian Arabizi data, collected in TArC, with an MSA database, clearly assumes difficulties in identifying tokens. However, as shown in Table 3, the results were not unsatisfactory, in terms of quality. This is mainly because TArC has been normalised to CODA-Star, an Arabic character encoding, MSA-like. In fact, as input to the Camel morphology analyser, we provided the lemma annotation level of each TArC token, by excluding the tokens classified as *foreign* and *emotag*, and the tokens POS-tagged as punctuation (*PUNC*), numerals (*NOUN_NUM*) or proper nouns (*NOUN_PROP*). The excluded tokens amount to 9,363 tokens, thus, the total of lemmas provided to the Camel analyser was 33,986.[27] In Table 3 we report the results of Camel Tools on TArC data.

Total of TArC token provided: 33,986		
Not Found	**Wrong Annotation**	**Correct Annotation**
4,017	6,056	23,913

Table 3: *Results of CAMeL Tools on TArC data*

The table shows that 4,017 tokens were not recognised at all by the analyser (*Not Found* in Table 3). In some other cases (**Wrong Annotation**), the morphological analyser provided a root instead, based on MSA, but this was incorrect in the case of TA, as shown in Example 2. These cases amount to 6,056. The cases of **Correct Annotation**, on the other hand, amount to 23,913.

(3) *al boulis*
āl- būlīs
ال بيلس fr:*police* [Camel root]
ال بلس fr:*police* [Correct root]
'The policeman'.

Considering the linking functionality envisaged for this level of TArC annotation, while manually

[26] These are available at https://github.com/CAMeL-Lab/camel_tools.

[27] As shown in Table 1, the total tokens of TArC are 43,349. These correspond to an amount of 5,063 unique, non-repeated, lemmas.

validating the roots automatically generated, we took some decisions based on the lexicographic database characteristics. When a lemma results from the combination of different words (as in the case of *blāš*, 'without', which is the fusion of *b-*, *lā* and *šy?*), the database records the TA lemma both as it is (*blāš*) and pointing to its components. Therefore, by validating TArC roots, we left these tokens as they are, instead of reducing them to their etymological components.

Finally, after the manual correction and integration of the *Not Found* and **Wrong Annotation** occurrences, respectively, the number of unique roots in TArC amounts to 1356.[28] The 76.3% of these (1034 unique roots) are matching with the lexicographic database roots.

6 Conclusions and Future Work

In this paper, we described our work on linking two linguistic tools previously created for different purposes. This work concerns Tunisian Arabic, and the resources we are working on are a large lexicographic database and a corpus of informal written texts from digital contexts. We explained the characteristics of these linguistic tools and how we managed to link them by enhancing their content. The work described is an ongoing pilot study, part of a larger project involving the development of resources for under-resourced languages. We described the methodology we developed for these types of languages. We outlined how this methodology adheres to the principles of open science, emphasizing transparency, interoperability and accessibility of data. Our project involves the use of existing language resources using tools, corpora, glossaries and dictionaries freely available among the scientific community. We deal with standardisation and morpho-syntactic annotation of texts with NLP tools. These ensure the reproducibility of our methodology. By sharing both the annotated data and the tools we create, other researchers will benefit from our work. The annotated data will be made available through a freely accessible digital platform. The NLP tools will be released on a repository dedicated to the project. Overall, the work described is in line with the FAIR principles, facilitating open and effective research on under-resourced languages.

[28] For *unique root*, we mean the roots counted only once.

References

Soha Abboud-Haggar. 2006. Dialects: Genesis. In *Encyclopedia of Arabic Language and Linguistics*, volume I, pages 613–622. Brill, Leiden – Boston.

Diana Alhafi, Anton Deik, Elhadj Benkhelifa, and Mustafa Jarrar. 2019. Usability Evaluation of Lexicographic e-services. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*. IEEE.

Taïeb Baccouche. 1994. *L'emprunt en arabe moderne*. Académie tunisienne des sciences des lettres et des arts, Beït al-Hikma.

Taïeb Baccouche. 2009. Tunisia. In *Encyclopedia of Arabic Language and Linguistics*, volume IV, pages 571–577. Brill, Leiden – Boston.

Marcelin Beaussier, Mohamed Ben Cheneb, and Albert Lentin. 2006. *Dictionnaire Pratique Arabe-Français (Arabe Maghrébin)*. Ibis Press, Paris.

Peter Behnstedt. 1998. Zum Arabischen von Djerba (Tunesien) I. *Zeitschrift für Arabische Linguistik, 35*, pages 52–83.

Peter Behnstedt. 1999. Zum Arabischen von Djerba (Tunesien) II: Texte. *Zeitschrift für Arabische Linguistik, 36*, pages 32–65.

Rached Ben Abdelkader et al. 1977. *Peace Corps English-Tunisian Arabic Dictionary*. Peace Corps, Washington D.C.

Wahid Ben Alaya and Michel Quitout. 2010. *L'Arabe tunisien de poche – Guide de conversation*. Assimil France, Chennevières sur Marne Cedex.

Hager Ben Ammar and Valérie Vacchiani. 2016. *Parler tunisien fissa! Une méthode originale pour apprendre l'arabe tunisien en 6 mois*. Editions Arabesques, Tunis.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*, 3:1137–1155.

Massimo Bevacqua. 2008. Osservazioni sul linguaggio dei giovani tunisini. *Il filo di seta – Studi arabo-islamici in onore di Wasim Dahmash*, pages 11–24.

Gilbert Boris. 1958. *Lexique du parler arabe des Marazig*. Imprimerie nationale – Librairie C. Klincksieck, Paris.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A Multidialectal Parallel Corpus of Arabic. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, pages 1240–1245.

Naima Boussofara-Omar. 2006. Diglossia. In *Encyclopedia of Arabic Language and Linguistics*, volume I, pages 629–637. Brill, Leiden – Boston.

Abdelghani Bouziane, Djelloul Bouchiha, and Noureddine Doumi. 2020. Annotating Arabic texts with linked data. In *2020 4th International Symposium on Informatics and its Applications (ISIA)*, pages 1–5. IEEE.

Dominique Caubet. 2019. Vers une littératie numérique pour la darija au maroc, une démarche collective. In Catherine Miller, Alexandrine Barontini, Marie-Aimée Germanos, Jairo Guerrero Guerrero, and Christophe Pereira, editors, *Studies on Arabic Dialectology and Sociolinguistics. Proceedings of the 12th International Conference of AIDA*. Livres de l'IREMAM.

Guido Cifoletti. 1998. Osservazioni sugli italianismi nel dialetto di Tunisi. *Incontri linguistici*, 21:137–153.

Guido Cifoletti. 2004. *La lingua franca barbaresca*. Il Calamo.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological Word-Embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292, Denver, Colorado. Association for Computational Linguistics.

Ines Dallaji, Ines Gabsi, Stephan Procházka, and Karlheinz Mörth. 2020. A Digital Dictionary of Tunis Arabic-TUNICO (ELEXIS). *Slovenian language resource repository CLARIN.SI*.

Mohamed Daoud. 2007. The Language Situation in Tunisia. *Language planning and policy in Africa, Vol. II: Algeria, Côte d'Ivoire, Nigeria and Tunisia*, pages 256–308.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R El-Beltagy, Wassim El-Hajj, et al. 2021. A panoramic survey of natural language processing in the arab world. *Communications of the ACM*, 64(4):72–81.

Humphrey Davies. 2006. Dialect Literature. In *Encyclopedia of Arabic Language and Linguistics*, volume I, pages 597–604. Brill, Leiden – Boston.

FMG De Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer, Dieter Van Uytvanck, et al. 2018. Clarin: Towards FAIR and responsible data science using language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3259–3264.

Olivier Durand and Maura Tarquini. 2023. *Corso di arabo tunisino. Manuale di comunicazione con grammatica ed esercizi. Livelli A1-B2 del Quadro Comune Europeo di Riferimento per le Lingue*. Ulrico Hoepli Editore S.p.A., Milano.

Fadhl Eryani, Nizar Habash, Houda Bouamor, and Salam Khalifa. 2020. A spelling correction corpus for multiple Arabic dialects. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4130–4138.

Charles A. Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.

Darja Fišer and Andreas Witt. 2022. *CLARIN: The Infrastructure for Language Resources*, volume 1. Walter de Gruyter GmbH & Co KG.

Elisa Gugliotta. 2022. *Realization of a Tunisian Arabish Corpus with use within the scope of NLP-Natural Language Processing*. Ph.D. thesis, Sapienza University of Rome and Université Grenoble Alpes.

Elisa Gugliotta and Marco Dinarelli. 2022. Tarc: Tunisian arabish corpus first complete release. In *13th Conference on Language Resources and Evaluation (LREC 2022)*.

Elisa Gugliotta and Marco Dinarelli. 2023. An empirical analysis of task relations in the multi-task annotation of an arabizi corpus. *Accepted paper for the 4th Conference on Language, Data and Knowledge*.

Elisa Gugliotta, Marco Dinarelli, and Olivier Kraif. 2020. Multi-task sequence prediction for Tunsian arabizi multi-level annotation. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 178–191.

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, et al. 2018. Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*.

Mustafa Jarrar and Hamzeh Amayreh. 2019. An Arabic-Multilingual Database with a Lexicographic Search Engine. In *Natural Language Processing and Information Systems*, pages 234–246. Springer International Publishing.

Mustafa Jarrar, Fadi A. Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlisch. 2022. Lisan: Yemeni, Iraqi, Libyan, and Sudanese Arabic dialect corpora with morphological annotations. *arXiv preprint arXiv:2212.06468*.

J. Jourdan. 1913. *Cours normal et pratique d'Arabe Parlé – Vocabulaire – Historiettes – Proverbes – Chants – Dialecte Tunisien, 4e édition*. Éditions Bouslama, Tunis.

Saméh Kchaou, Rahma Boujelbane, and Lamia Hadrich Belguith. 2022. Hybrid pipeline for building Arabic Tunisian dialect-standard Arabic neural machine translation model from scratch. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Emna Labidi. 2017. L'artisanat traditionnel à Tunis – Sa terminologie et son lexique. *Tunisian and Libyan Arabic Dialects – Common Trends – Recent Developments – Diachronic Aspects*, pages 147–160.

Jérôme Lentin. 2008. Middle Arabic. In *Encyclopedia of Arabic Language and Linguistics*, volume III, pages 215–224. Brill, Leiden – Boston.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. 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.

Agnese Macchiarelli. 2023. Sinergie fra vedph e cnr-ilc in termini di condivisione della conoscenza e sostenibilità dei progetti digitali. In *DH. 22–Digital Humanities. Per un confronto interdisciplinare tra saperi umanistici a 30 anni dalla nascita del World Wide Web*. L'Erma di Bretschneider.

Abrerrahmân Marçais, William Guîga. 1961. *Textes arabes de Takroûna par William Marçais et Abderrahmân Guîga – II – Glossaire – Contribution à l'étude du vocabulaire arabe, Tome I – VIII*. Imprimerie Nationale – Centre National de la Recherche Scientifique – Librairie Orientaliste Paul Geuthner, Paris.

Philippe Marçais and M.-S. Hamrouni. 1977. *Textes d'arabe maghrébin*. Librairie d'Amérique et d'Orient, Adrien Maisonneuve – J. Maisonneuve, succ., Paris.

William Marçais. 1950. Les parlers arabes. *Initiation à la Tunisie*, pages 195–219.

Karen McNeil. 2018. Tunisian Arabic corpus: Creating a written corpus of an 'unwritten' language. *Arabic corpus linguistics*, 30.

Salima Mdhaffar, Fethi Bougares, Yannick Esteve, and Lamia Hadrich-Belguith. 2017. Sentiment analysis of Tunisian dialects: Linguistic ressources and experiments. In *Third Arabic Natural Language Processing Workshop (WANLP)*, pages 55–61.

Karima Meftouh, Salima Harrat, and Kamel Smaïli. 2018. Padic: Extension and new experiments. In *7th International Conference on Advanced Technologies ICAT*.

Ouafae Nahli, Elisa Gugliotta, Nadia Khlif, and Giulia Benotto. 2023. Advancing dialectal analysis: Annotating corpora and building lexical resources for Arabic dialects. *Forthcoming*.

Alfred Nicolas. (s.d.) [1911]. *Dictionnaire Français-Arabe, Idiome Tunisien*. Imprimeur - Éditeur Frédéric Weber, Tunis.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. 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, Marseille, France. European Language Resources Association.

Jonathan Owens. 2006. *A Linguistic History of Arabic*. Oxford: Oxford University Press.

Livia Panascì. 2021. *Studi lessicali sull'arabo di Tunisia*, volume I, II, III, IV. PhD Thesis in Civilizations of Asia and Africa, Sapienza University of Rome, Rome.

Laurette Pretorius and Claudia Soria. 2017. Introduction to the Special Issue. *Language resources and evaluation*, 51:891–895.

Michel Quitout. 2002. *Parlons l'arabe tunisien: Langue & culture*. L'Harmattan, Paris - Budapest - Torino.

Jean Quéméneur. 1961a. Notes sur quelques vocables du parler tunisien figurant au Supplement de A. Lentin - 1ère partie. *Revue de l'I.B.L.A., Vol. 24/93*, pages 1–22.

Jean Quéméneur. 1961b. Notes sur quelques vocables du parler tunisien figurant au Supplement de A. Lentin - 2ème partie. *Revue de l'I.B.L.A., Vol. 24/94*, pages 167–181.

Jean Quéméneur. 1962. Glossaire de dialectal 1942-1962. *Revue de l'I.B.L.A., Vol. 25/100*, pages 325–367.

Veronika Ritt-Benmimoun. 2014. *Grammatik des arabischen Beduinendialekts der Region Douz (Südtunesien)*. Harrassowitz Verlag, Wiesbaden.

Vitaly Romanov and Albina Khusainova. 2019. Evaluation of Morphological Embeddings for the Russian Language. In *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, pages 144–148.

Rana Aref Salama, Abdou Youssef, and Aly Fahmy. 2018. Morphological word embedding for Arabic. *Procedia computer science*, 142:83–93.

Lotfi Sayahi. 2014. *Diglossia and language contact: Language variation and change in North Africa*. Cambridge University Press.

Erhan Sezerer and Selma Tekir. 2021. A Survey on Neural Word Embeddings. *arXiv preprint arXiv:2110.01804*.

Hans Stumme. 1896. *Grammatik des tunisischen arabisch: nebst Glossar*. JC Hinrichs.

Fathi Talmoudi. 1981. *Texts in the Arabic Dialect of Sūsa (Tunisia): Transcription, Translation, Notes and Glossary*. Acta Universitatis Gothoburgensis, Göteborg.

Kees Versteegh. 2014. *Arabic language*. Edinburgh University Press.

Mark Wilkinson, Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific data*, 3(1):1–9.

Sane Yagi, Ashraf Elnagar, and Shehdeh Fareh. 2022. A Benchmark for Evaluating Arabic Word Embedding Models. *Natural Language Engineering*, pages 1–26.