Challenges and Progress in Constructing Arabic Dialect Corpora and Linguistic tools: A Focus on Moroccan and Tunisian Dialects

Ouafae Nahli [a], Elisa Gugliotta [a], Nadia Khlif [a b], Benotto Giulia [a]
[a] Istituto di Linguistica Computazionale, Consiglio Nazionale delle Ricerche;
Via G. Moruzzi, 1, 56124 Pisa - Italy;
{firstname.lastname}@ilc.cnr.it
[b] Laboratoire des recherches informatiques university Mohammed First Oujda
nadia.khlif@ump.ac.ma

2023 7th IEEE Congress on Information Science and Technology (CiSt) | 978-1-6654-6133-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/CiSt56084.2023.10410009

***Abstract*—Given the lack of resources for Arabic dialects, the construction of corpora, lexical resources, and tools is a non-trivial challenge. The focus of the article is to describe our in-progress work to address these deficiencies. We start with Moroccan and Tunisian dialects to provide annotated corpora and corpus-based lexical resources. We also aim to extend an existing morphological engine with linguistic resources built *ad hoc* for each dialect. In addition, we develop an integrated component in the morphological engine to better address linguistic and sociolinguistic characteristics while preserving the integrity of dialectal texts.**

***Index Terms*—Arabic dialects, Moroccan dialect, Tunisian dialect, corpora, lexical resources, Aramorph.**

I. Introduction

Arabic Dialects (ADs) have always been relegated to an oral status, however with the emergence of social networks, people started to communicate through the linguistic form they know best, i.e. their dialect. Thus, Arabic dialects recently acquired a spontaneously written status. Consequently, we are faced with texts written in a spontaneous Arabic language used for informal communication on social networks. We define Contemporary Written Arabic (CWA) this informal and spontaneous Arabic, which reflects the linguistic reality of native speakers. With the increased accessibility of CWA data, interests in applying NLP approaches to the building of CWA corpora have risen [9], [17]. However, the NLP tools adopted for Modern Standard Arabic (MSA) do not perform well in the CWA processing (see [8], [16]). In fact, Arabic dialects are not standard languages and CWA texts have orthographic, linguistic, and sociolinguistic features that have never been addressed in texts written in MSA. In addition, freely available CWA resources, such as linguistic annotated corpora and tools, are still scarce [11]. Consequently, the biggest challenge is to create resources and tools specifically designed to study CWA corpora [10]. For this reason, within the scope of the project *A Lexical corpus-based Model of Contemporary Written Arabic* (CWALM)[1], we aim to build CWA resources, such as annotated corpora, lexical resources, and tools. CWA corpora are constructed by devising a methodology inspired by the FAIR principles of open science. [2] So, in the first phase, we studied the existing resources that could be used to answer the needs of our project. For the morpho-syntactic analysis of CWA texts, we are adapting the rule-based morphological analyser, *Buckwalter Arabic Morphological Analyser* (BAMA), also known as *Aramorph* [5], [6].

This paper presents our in-progress work which started with the building of an annotated corpus and a morphological analyser for the Moroccan and Tunisian dialects. Section II provides an overview of the availability of corpora and tools concerning the colloquial Arabic varieties of North Africa (or the Maghrebi Arabic), in particular Tunisian and Moroccan. Section III discusses linguistic, sociolinguistic, and morphological characteristics of both Modern Standard and dialectal Arabic. It also describes the general characteristics of Arabic script and features of informal texts. Our ongoing efforts are detailed in Section IV, where we provide a concise overview of the CWALM project at IV-A. The groundwork commences with data collection, discussed in IV-B. Subsequently, we delve into the Aramorph adaptation, outlining its stages and challenges, beginning with Moroccan and Tunisian Arabic in IV-C. Finally, in Section V, we summarize and analyze the completed work, while also proposing potential extensions and future endeavors.

II. State of the Art

Concerning resources on the Moroccan dialect (henceforth MD), we should mention the Darija Open Dataset (*DoDa*), which is probably the largest open-source dataset for Moroccan, provided with an English translation and built for NLP purposes [25]. The Moroccan Arabic Corpus (MAC) is a large Moroccan corpus for sentiment analysis, released in 2022. Consisting of 18,000 manually tagged tweets resulting in a lexicon dictionary of 30,000 words tagged as positive, negative, and neutral. This corpus has two types of tweets, MSA tweets (106,377 tokens) and Moroccan tweets (44,942 tokens) [24]. The Moroccan Dialect Electronic Dictionary (MDED)

[1] https://cwalm.ilc.cnr.it/

[2] The FAIR principles are the guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets. These principles emphasise machine-actionability i.e., the capacity of computational systems to find, access, interoperate, and reuse data with none or minimal human intervention [32].

is an electronic lexicon containing almost 15,000 Moroccan entries written in Arabic script and translated into MSA. In addition, MDED entries are annotated with useful metadata such as POS, etymologies, and roots [28]. MORALEX is a lexicon that includes 402 Moroccan Arabic affixes and clitics [27].

Regarding Tunisian Arabic, *TSAC* is a corpus of Facebook comments, built for sentiment analysis by [22]. The *TD-COM corpus* is a parallel corpus Tunisian-MSA, extracted from social networks by [19]. The *Tunisian Arabizi Corpus* (TArC), released by [12], is an available corpus of Tunisian *Arabizi* normalized in CODA, [3] and gathers Arabizi texts from blogs, forums, and Facebook. TArC is provided with different morpho-syntactic annotations, including POS tagging and lemmatization.[4]. Finally, there is also a wordnet for Tunisian (the *AebWordNet*) collected by [18].

We also mention some multi-dialectal resources, that include the Maghrebi Arabic. *PADIC* is a parallel corpus that includes Tunisian, Algerian, and Moroccan [23]. The *ADI17* corpus collects Algerian and Moroccan Arabic among 17 dialects [2]. The dataset of tweets, *DART*, includes Maghrebi, Egyptian, Levantine, Gulf, and Iraqi Arabic [3]. The *ArabicWeb16* corpus gathers MSA and Maghrebi, Egyptian, Gulf, and Levantine dialects. This corpus consists of a Web crawl of roughly 150M Arabic Web pages [30]. The parallel corpus *The Multi Arabic Dialect Applications and Resources* (MADAR) collects the dialects of 25 Arab cities, including cities of Tunisia (Tunis and Sfax) and Morocco (Rabat and Fes) [4].

In sum, each linguistic resource, even if created for a specific purpose, represents a portion of the linguistic reality (written or oral). These are indeed valuable resources and, for reasons that could be summarised as a strategy of "ecology of the linguistic datum", we decided to exploit some of these resources for our project (see section IV-B).

Regarding morpho-syntactic annotation tools, some of the few systems that provide annotations (morpho-syntactic and lexical) of an Arabic dialect are ADAM [29], MADAMIRA [26] and CALIMA [16], for the Egyptian Arabic. Concerning Tunisian Arabic there is the Multi-Task Architecture (MTA), provided by [14], which can be trained to produce various morpho-syntactic information. With regards to Levantine, Maghrebi, and Gulf Arabic, [8] proposed a POS tagger based on a Conditional Random Fields (CRF) sequence labeler. Concerning the Gulf Arabic, [20] built a morphological analyser covering over 2600 verbs, while [21] proposed a full morphological analysis and disambiguation system. However, as noted by [31], the existing morpho-syntactic analysers present some disadvantages, such as the low accuracy results, the low coverage of Arabic dialect, and the missing textual contexts. Consequently, we considered that adapting a tool is less time-consuming than starting our work from scratch. For this reason, we have decided to adapt Aramorph to process dialectal texts. However, the adaptation should be based on specific dialectal linguistic features. The first step must be a linguistic study and as for our case we started with the study of Moroccan and Tunisian linguistic features (see Section IV-C).

[3]Arabizi is a spontaneous orthography based on the Latin alphabet and Arabic numerals, employed by Arabic users of digital and informal writing [33]. CODA is the *Conventional Orthography for Dialectal Arabic* [15].

[4]Available at: https://github.com/eligugliotta/tarc

III. Arabic Linguistic Characteristics

1) Modern Standard Arabic & Arabic Dialects. Modern Standard Arabic (MSA) is the written language recognised in all Arab countries with grammatical rules and dictionaries decreed by academies or authorities.

Arabic dialects typically remain spoken and confined to specific geographical areas, lacking official status, grammar rules, or standardized dictionaries. However, a noticeable shift is observed as people increasingly communicate through social networks, utilizing their native dialects, which subsequently acquire a written form. As a result, written texts now reflect the linguistic reality and practices of native speakers, capturing the essence of their spoken language.

The demarcation between Modern Standard Arabic (MSA) and Arabic dialects is not precisely defined; instead, it forms a linguistic continuum where various inter-comprehensible dialects coexist. This means that despite the distinct characteristics of their respective dialects, people can generally understand one another. However, it's essential to note that inter-comprehension may not be as evident when dealing with two geographically distant dialects, such as Moroccan and Syrian Arabic. In such cases, communication might face greater challenges due to more pronounced linguistic differences between these distant varieties.

Furthermore, many Arabic dialects reflect the phenomenon of multilingualism, due to linguistic contact between Arabic and foreign languages. People master one or more foreign languages in addition to the Arabic language and can switch between them depending on the contexts in which they communicate.

2) General Linguistic Features of Arabic. In the Arabic language, different information is encoded through morphology, which has a multi-level structure. We distinguish three levels, as represented in Figure 1:

- The derivation layer is the deepest one. At this level, the root combines with the vowels, according to determined patterns, to produce a verbal or a nominal lemma.
- The inflectional layer is the one where the lemma combines with inflectional affixes to give inflectional forms.
- The morpho-syntactic layer combines the inflected form with clitics (prepositions, conjunctions, definite articles, etc.) to shape a rich and complex surface form.

Written tokens correspond to either a "minimal word form" which results in an inflected nominal or verbal form, or a morphologically more complex token resulting from a concatenation of a minimal inflected form with clitics.

Example 1, */wataktubuhu/*, demonstrates an example of maximal word form in MSA.

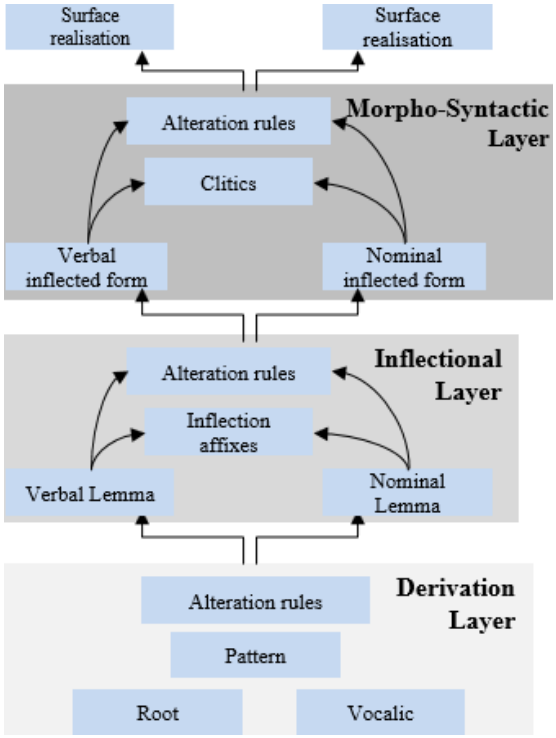

Fig. 1: Multi-tier Structure of the Arabic Morphology

Example 2, */makatktabhāš/*, illustrates a complex maximal form in MD, featuring the circumfix negation */ma- -š/*. [5]

Example 1

wa=**ta-ktub-u**=hu

and=**2MS-write.IPFV-PRS.IND**=it

'and you write it'

Example 2 /makatktabhāš/

ma=ka=**t-aktab**=hā=š

NEG=PROG=**2MS-write.IPFV-PRS.IND**=it=NEG

'you are not writing it'

In the two examples, the inflected form is surrounded by clitics and the morphological structure is:

proclitics=**prefix-stem-suffixes**=enclitics.

By removing clitics, the remaining word form is a minimally autonomous inflected form whose structure consists of:

prefix-stem-suffixes.

Due to these levels of Arabic morphological embedding, word tokenisation must be followed by a sub-tokenisation phase marking the boundaries between proclitics, the minimal word, and enclitics.

3) General Challenges of Arabic Script Processing In Arabic written texts, vowels, gemination, and other signs are written as diacritics added above or below consonant letters. Their marking, however, is not systematic. For instance, the word كَتَبَ /kataba/ 'he wrote'[6] can be written in any of the following variants: *ktb*, *katb*, *katab*, *ktaba*, *katba*, etc. Furthermore, by vocalising *ktb* differently, one can obtain other words, such as: *kutub* (books), *katb* (writing), *kattaba* (dictate; make write). As a result, the omission of diacritics in written texts causes extensive homography and ambiguity in Arabic.

The use of social media has created new challenges. We are dealing with spontaneous texts that reflect dynamic aspects of dialects, such as code-switching (inter- or intra-sentential language shift), or script-mixing.

With the rise of technologies, the lack of Arabic keyboard led people to use, in digital and informal writing contexts, the Arabizi script which is a spontaneous orthography based on the Latin alphabet and Arabic numerals [33].

On the other, some phonetic realisations of dialects are not represented in the Arabic alphabet. We note a substitution that is, using numbers and Latin letters to represent phonemes that have no equivalent in the Arabic alphabet. This is the case, for Maghrebi dialects, of the phone /g/, a phonetic variant of the phoneme /q/, encoded in MSA with the grapheme ق. In fact, /g/ has no representation in the MSA alphabet. This can be encoded with the additional letter ڤ, leveled on the MSA alphabet (ق), or represented with the Latin letter *g*.

Thus, the phonetic word /bagra/, 'cow' (بقرة in MSA) in Tunisian or Algerian Arabic can be written as بڤرة, or *bagra*, while in Moroccan Arabic can be written as بقرة or بgرة.[7]

In Table I we show some Arabizi examples, however, we note that dialectal encoding (Arabizi or Arabic script) may differ from region to region. This orthographic variety introduces consistent ambiguities in the processing of informal and dialectal texts.

IPA	*Arabic*	*Arabizi*	*IPA*	*Arabic*	*Arabizi*	*IPA*	*Arabic*	*Arabizi*
[x]	خ	5, kh	[ʕ]	ع	3, a	[h]	ه	8, h
[ʒ]	ج	j	[ɣ]	غ	4, gh	[ʃ]	ش	ch, (sh)
[ħ]	ح	7, h	[tˤ]	ط	6, t	[q]	ق	9, q

TABLE I: Some examples of the Arabizi script

In general, the main elements, typically found in social media texts, that could be considered noise, are the para-textual elements (symbols used to express an idea, mood, or feeling, e.g. emoticons and emoji), or word elongation (elements repetition to express emphasis), e.g., مبروووك *mbrwwwwk* 'congraaaats'.

[5]The transcription is phonetic and does not reflect the orthographic peculiarities in dialect.

Interlinear glosses follow the standard set of parsing conventions and grammatical abbreviations explained in: "The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses" February 2008. Hyphen marks segmentable morphemes and an equal sign marks clitic boundaries, both in transliterations and in the interlinear gloss.

[6]Concerning the examples, in this paper, we report Arabic transliterated in Latin script.

[7]The numeral *9* can be used as a substitution based on the graphical similarity between the numeral and the Arabic grapheme ق, e.g., *ba9ra*, for /baqra/.

IV. In-progress Work

A. CWALM Project Aims

Our work is part of the project *A Lexical corpus-based Model of Contemporary Written Arabic* (CWALM). CWALM aims at creating a lexical resource for Contemporary Written Arabic (CWA). Indeed, it involves vocabulary extraction from corpora collecting real texts in different dialect varieties. The new theoretical approach in CWALM aims to interpret Arabic as a single linguistic system, that lies along a linguistic continuum from colloquial Arabic to MSA. Therefore, the corpus-based lexical resource of CWA will provide objective and substantive data to test competing theories on the linguistic status of the Arabic language. Thus, we designed a procedure organised in the following steps:

- **Step 1.** *Construction of a Representative Corpus of the Colloquial Arabic Varieties.* To realise a truly representative corpus, we will include contemporary written texts from extremely different sources including different genres, regions, and sociolinguistic backgrounds.
- **Step 2.** *Tools Adaptation.* To provide our corpora with linguistic annotations, a study is underway to adapt the morphological engine Aramorph to process written dialectal words.
- **Step 3.** *Construction of a Structured Lexicon.* With this aim, we extract data from corpus annotations and analyses. Lexical items will be enriched with key information about morphology and lexical refinement.
- **Step 4.** *Construction of a Lexical Model.* The model will allow a systematic connection between corpus data, lexical entries, and lexical information from existing lexicographical sources.

B. Data Construction

At this initial stage, we relied on other existing resources, such as corpora or lexical databases.

1) Resources Exploited to Build the Tunisian Corpus. As shown in Table II, we include the pre-existing *TArC* corpora which have been annotated with various linguistic information by [12]. Additionally, we have developed a new written corpus that comprises a collection of texts extracted from Facebook and blog.

Type of the corpus	**Number of token**	
TarC		43349
new corpus	Blog	9961
	Facebook	8465
Total		**61775**

TABLE II: Statistics of the Tunisian Corpus

2) Resources Exploited to Build the Moroccan Corpus. As shown in Table III, we included the two pre-existing MADAR-MA (the Moroccan texts of MADAR) and the MAC corpora. Additionally, we have developed a new written corpus for the Moroccan Dialect, named MODIC. This corpus comprises a collection of texts extracted from Facebook and Twitter, using specific keywords characteristic of MD, for example, مبغيتش /mā=bġī=t=š/ and بزاف /bzzāf/.

Type of the corpus	**Number of token**	
MAC		44,942
MADAR-MA	Fes	11,691
	Rabat	72,120
MODIC	Facebook	101,548
	Twitter	787,581
Total		**1,017,882**

TABLE III: Statistics of the Moroccan Corpus

C. Tools Adaptation

Aramorph has been created for the morpho-syntactic annotation of MSA texts.[8] It translates Arabic text using Buckwalter's transliteration system. It analyses and produces an analytical report using a morphological analysis and Part-of-Speech (POS) tagging algorithm that includes tokenisation, word segmentation, dictionary lookup, and compatibility checks. The components of Aramorph are essentially two: the rules engine for morphological analysis and a repository of linguistic resources composed of three lexicons, i) the DictPrefix lexicon, which consists of sequences of proclitic and inflectional prefixes, ii) the DictStem lexicon, which contains 38,600 lemmas and iii) the DictSuffix lexicon, which consists of sequences of inflectional and enclitic suffixes. These lexicons come with three compatibility tables used to verify the combinations of (proclitics + prefixes), stems, and (suffixes + enclitics).

1) Methodology. In the initial phase, our primary focus was on adapting Aramorph to handle Moroccan and Tunisian Arabic. For each dialect, our initial efforts centered around modeling the linguistic resources (DictStem, DictPref, and DictSuff) using pre-existing terminological sources. To study and manage the compatibility between {proclitics + inflectional prefixes}, {roots} and {inflectional suffixes + proclitics}, we classified the lemmas into verbal, nominal, and adjectival categories. Furthermore, each category is classified into groups according to the inflectional and syntactic behavior.

Verbal stems are classified into stems of Perfect Verb (PV), Imperfect Verb (IV), and Command Verb (CV). Thus, they are classified in:

- Regular verbs, which constituted strong radical consonants and had a single stem for both PV and IV;
- Irregular verbs, having different stems, depending on the inflectional affixes. They are doubled verbs, middle and last weak verbs.

Nominal and adjectival plural forms are extracted from external resources and expressed in the DictStem. Instead, the singular forms are classified according to gender:

- Masculine nominal lemmas, having a stem that doesn't change with the addition of possessive pronouns. For example باب /bāb/ 'door'.

[8]Aramorph is downloadable at: http://www.nongnu.org/aramorph

- Feminine nominal lemmas, having a stem that changes with the addition of possessive pronouns. For example طَبْلَة /ṭablah/ ‘table’.
- Nominal and adjectival lemmas, which decline in forms that may express gender. For example زْوِين /zwiyn/ ‘beautiful’.

2) Linguistic Resources Building. For Tunisian Arabic, to build DictStem we utilized the terminological resources provided by TUNICO [7]. To build the DictSuffix and the DictPrefix, we extracted all combinations of clitics and affixes from the corpus MADAR-TUN, annotated by [13].

For the Moroccan dialect, to build DictStem we referred to the book ‘MOROCCAN ARABIC Textbook” [1], the terminological lists of MADED [28] and *DoDa* [25]. To build the DictPrefix and the DictSuffix, we used and extended MORALEX [27].

Table (IV) reports the number of lemmas (classified in *Nouns*, *Adjectives* and *Verbs*) and the number of combinations (between {proclitics+prefixes} and {suffixes+enclitics}), that we have encoded so far in the linguistic resources required by Aramorph to process Moroccan and Tunisian texts.

Lemma Classes:	*Nouns*	*Adjectives*	*Verbs*	*Prefixes*	*Suffixes*
Moroccan	1,000	900	524	216	455
Tunisian	300	65	300	93	190

TABLE IV: Aramorph Moroccan and Tunisian Linguistic Resources

3) Pre-processing of Dialectal Text. The normalization used by AraMorph is not suitable for informal dialectal texts because it removes diacritics, punctuation, numbers, and non-Arabic characters. To solve this problem, we have developed a component that performs preprocessing, tokenization, and normalization while retaining dialect characteristics. Aramorph provides only analysis using updated linguistic resources. Here are some examples of processed dialect specifics:

- When a number or a Latin letter is within the Arabic token, it is substituted with the correspondent Arabic letter:
  فت7لي = فتحلي */ftaḥliy/*
  بgرة = بقرة */bgrah/*
- Lengthened letters are replaced with a unique letter: مبروووك= مبروك */mabrūk/*.
- The interjection : هههه *hhhh*.

In addition, we marked numbers and Latin words, if they consist of separated tokens, as NUMB and FOREIGN, respectively. As shown in figure 2, the preprocessing module and Aramorph are combined in a unique application which we are called DiMorph.

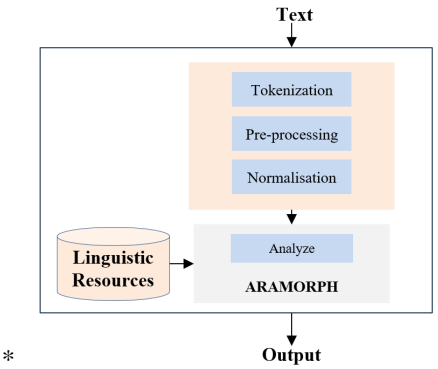

Fig. 2: Structure of DiMorph

We processed a Moroccan text (11,419 tokens) with the original Aramorph using the MSA linguistic resources, the Moroccan Aramorph that is the original Aramorph using the MD linguistic resources, and the DiMorph. In the DiMorp, the number of tokens is 12,219, considering that it also processes the punctuation, the numbers (as unique tokens), and the paratextual elements (such as INTERJ and emoticons). Table V shows the obtained results.

	Found	*Not Found*	*Total Tokens*
Original Aramorph	56,28%	43,72%	11,419
Moroccan Aramorph	76%	24%	11,419
DiMorph	87,86%	12,13%	12,219

TABLE V: First Results of DiMorph

The tokens recognized by the original Aramporph (56,28%) are MSA tokens, instead, the Moroccan-specific tokens are not-found (43,72%). After creating the linguistic resources (DictPrefix, DictSuffix, and DictStem) and the adaptation rules for Moroccan Arabic, by using these resources, Moroccan Aramorph recognizes 76% of tokens. Concerning the DiMorph, the recognized tokens are 87,86%. Part of the 12,13% not recognized tokens comes back to lemmas not yet inserted in the linguistic resources. Another part is made of orthographic errors.

V. Conclusion and Future Works

We presented an in-progress work aimed at building a corpus of Moroccan and Tunisian colloquial Arabic. The corpus was extracted from social media platforms and enriched by pre-existing resources. We are adapting the Aramorph engine to process dialectal texts and to provide them with linguistic annotations. The work has been on two levels:

- the development of a component that deals with text normalization, according to the dialectal characteristics
- the construction of dialectal linguistic resources.

For future works, we plan to:

- Enrich the Moroccan and Tunisian Aramorph linguistic resources;
- Annotate the Moroccan and the Tunisian corpus;
- Build a structured lexicon;
- Build a lexical model in a standard format.

Finally, to complete our study on North African dialects, we leave open the possibility of including Algerian Arabic in our work.

REFERENCES

[1] *Peace Corps / USA (Ed.). (2016). Moroccan Arabic textbook: student edition (N.A.). Rabat: Peace Corps (U.S.) Morocco.* http://friendsofmorocco.org/Docs/MorrocanArabicSept2016.pdf, 2016.

[2] Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019.

[3] Israa Alsarsour, Esraa Mohamed, Reem Suwaileh, and Tamer Elsayed. Dart: A large dataset of dialectal arabic tweets. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, 2018.

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

[5] T Buckwalter. Buckwalter Arabic morphological analyzer (bama) version 2.0. linguistic data consortium (ldc) catalogue number ldc2004l02. Technical report, ISBN1-58563-324-0, 2004.

[6] Tim Buckwalter. Buckwalter Arabic morphological analyzer version 1.0. linguistic data consortium. *University of Pennsylvania, LDC Catalog No.: LDC2002L49*, 2002.

[7] Ines Dallaji, Ines Gabsi, Stephan Procházka, and Karlheinz Mörth. A digital dictionary of tunis arabic-tunico (elexis). 2020.

[8] Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesouki, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy, and Laura Kallmeyer. Multi-dialect arabic pos tagging: A crf approach. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, 2018.

[9] Mahmoud El-Haj. Habibi - a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1318–1326, Marseille, France, May 2020. European Language Resources Association.

[10] Amany Fashwan and Alansary Sameh. Developing a tag-set and extracting the morphological lexicons to build a morphological analyzer for egyptian arabic. *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP).*, 2022.

[11] Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5):497–507, 2021.

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

[13] Elisa Gugliotta and Marco Dinarelli. 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 (LDK 2023).

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

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

[16] Nizar Habash, Ramy Eskander, and Abdelati Hawwari. A morphological analyzer for egyptian Arabic. In *Proceedings of the twelfth meeting of the special interest group on computational morphology and phonology*, pages 1–9, 2012.

[17] Mustafa Jarrar, Fadi A Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlisch. Lisan: Yemenu, irqi, libyan, and sudanese arabic dialect copora with morphological annotations. *arXiv preprint arXiv:2212.06468*, 2022.

[18] Ben Moussa Nadia Karmani and Adel M. Alimi. Construction d'un wordnet standard pour l'arabe tunisien. In *Proceedings of Colloque pour les Étudiants Chercheurs en Traitement Automatique du Langage naturel et ses applications, Sousse, Tunisia*, 2015.

[19] Saméh Kchaou, Rahma Boujelbane, and Lamia Hadrich Belguith. 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*, 2022.

[20] Salam Khalifa, Sara Hassan, and Nizar Habash. A morphological analyzer for gulf Arabic verbs. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 35–45, 2017.

[21] Salam Khalifa, Nasser Zalmout, and Nizar Habash. Morphological analysis and disambiguation for gulf Arabic: The interplay between resources and methods. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3895–3904, 2020.

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

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

[24] Jamal Kharroubi Moncef Garouani. An open and free moroccan arabic corpus for sentiment analysis. *The Proceedings of the International Conference on Smart City Applications*, 2022.

[25] Aissam Outchakoucht and Hamza Es-Samaali. Moroccan dialect-darija-open dataset. *arXiv e-prints*, pages arXiv–2103, 2021.

[26] Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, volume 14, pages 1094–1101. Citeseer, 2014.

[27] Tachicart Ridouane and Karim Bouzoubaa. Towards automatic normalization of the moroccan dialectal arabic user generated tex. *7th International Conference on Arabic Language Processing ICALP'19, Nancy, France.*, 2019.

[28] Tachicart Ridouane, Bouzoubaa Karim, and Jaafar Hamid. Building a moroccan dialect electronic dictionary (mded). *5th International Conference on Arabic Language Processing.*, 2014.

[29] Wael Salloum and Nizar Habash. Adam: Analyzer for dialectal arabic morphology. *Journal of King Saud University-Computer and Information Sciences*, 26(4):372–378, 2014.

[30] Reem Suwaileh, Mucahid Kutlu, Nihal Fathima, Tamer Elsayed, and Matthew Lease. Arabicweb16: A new crawl for today's arabic web. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 673–676, 2016.

[31] Ridouane Tachicart, Karim Bouzoubaa, Salima Harrat, and Kamel Smaïli. Morphological analyzers of arabic dialects: A survey. *Studies in Computational Intelligence*, 1061, 2022.

[32] Dumontier M. Aalbersberg I. et al. Wilkinson, M. The fair guiding principles for scientific data management and stewardship. *Sci Data 3, 160018.*, 2016.

[33] Mohammad Ali Yaghan. "Arabizi": A contemporary style of Arabic slang. *Design issues*, 24(2):39–52, 2008.