Lîsanī: Yemeni, Iraqi, Libyan, and Sudanese Arabic Dialect Corpora with Morphological Annotations

Mustafa Jarrar
Birzeit University
Birzeit, Palestine
mjarrar@birzeit.edu

Fadi A Zaraket
American University of Beirut
Beirut, Lebanon
fz11@aub.edu.lb

Tymaa Hammouda
Birzeit University
Birzeit, Palestine
1171779@student.birzeit.edu

Daanish Masood Alavi
UN Department of Peace-building and Political Affairs
New York, USA
masoodd@un.org

Martin Wählisch
UN Department of Peace-building and Political Affairs
New York, USA
waehlisch@un.org

***Abstract*—This article presents morphologically-annotated Yemeni, Sudanese, Iraqi, and Libyan Arabic dialects (Lîsanī) corpora. Lîsanī features around 1.2 million tokens. We collected the content of the corpora from several social media platforms. The Yemeni corpus (1.05M tokens) was collected automatically from Twitter. The corpora of the other three dialects (50K tokens each) was manually collected from Facebook and YouTube posts and comments. Thirty-five (35) annotators who are native speakers of the target dialects carried out the annotations. The annotators segmented all words in the four corpora into *prefixes*, *stems* and *suffixes* and labeled each with different morphological features such as *part of speech*, *lemma*, and a *gloss* in English. We developed the Arabic Dialect Annotation Toolkit (ADAT) to assist the annotators and to ensure compatibility with SAMA and Curras tagsets. We trained annotators on a set of guidelines and on how to use ADAT. ADAT is open source, and the four corpora are available at https://sina.birzeit.edu/currasat.**

I. Introduction

Around 300 million people in 23 countries speak and use the Arabic language in their daily lives. Classical Arabic (CA) is the old form of Arabic that is used in historical texts. Modern Standard Arabic (MSA) is used in formal communications including newspapers, media outlets, educational material and most of the televised content. Dialectal Arabic (DA) appears in colloquial and informal day-to-day communications. DA volume is massively increasing on social media. Processing and understanding such content in natural language processing (NLP) tasks is challenging [1]. This is mostly because Arabic dialects are under-resourced and have no standard orthography.

DA and MSA differ as follows.

(i) **Phonology**: People pronounce words differently with varied intonation and stress, and write them as pronounced. For example, the letter (ق */q*) is pronounced as 'g' (SAMPA phonetic notation) in Libyan and Sudanese and more of a 'k' in Yemeni as in the word قال*/kal* (say). Iraqi switches between 'q' and 'k' as in قال */qāl* and وقت*/wakt* (time). The letter (ث) typically denoting the sound 'T' in MSA, is pronounced as 's' in Sudanese and more of a 't' in Libyan as in the word ثياب */tyāb* (cloth). The letter (ک) typically denoting a 'k' sound in MSA, is pronounced *tš* 'tS' in Iraqi as in the word كلب*/tšlb* (dog).

(ii) **Morphology**: Arabic dialects are similar to MSA in inheriting templatic morphology where affixes play an important role. However, major differences exist between dialects and MSA and among dialects themselves. For example, negating the verb أكلت */ʔaklt* (I ate) in MSA precedes it with the particle لَمْ */lam* and inflects the verb itself to produce become لَمْ آكل */lm ʔākl* . The Lybian dialect uses the prefix م */m* and the suffix تش */tš* for negation, as in مأكلتش */mʔaklš* (I did not eat); while the prefix ما */mā* is enough in Sudanese. In Yemeni, the prefix (ل */l*) is used to negate the imperative as in لتخافون */ltẖāfwn* (Do not be afraid), which is a short of the لا */lā* particle in MSA, as in لا تخافوا */lā tẖāfwā* . The مو */mw* and مش */mš* particles are also common replacements for the MSA negation particles such as ليس */lys* . As will be discussed in the section IV-B, many of the MSA particles are used as affixes in Arabic dialects.

(iii) **Lexicon**: Each dialect has its own unique lexicon entries that are not used in MSA or other dialects. The word عنجاص */ʕnǧāṣ* (plum) is an Iraqi variant of the MSA إجّاص */ʔiǧāṣ* (pear), while Iraqi uses عرموط */ʕrmwṭ* to denote pears. The Sudanese use زول */zwl* to denote رجل */rǧl* (man). These variations can prove embarrassing as تينة */tynh* (tree of figs) in MSA denotes the human body's bottom in Libyan. Yemeni uses عشار */šār* and قطيب */qṭyb* for مخلل */mẖll* (pickles) and زبادي */zbādy* (thick yogurt), respectively. It shares with Iraqi ميز */myz* and بنكة */bnkh* for MSA's طاولة */ṭāwlh* (table) and مروحة */mrwḥh* (fan), respectively.

Contributions: This paper contributes to addressing the problem of under-resourced Arabic dialects and presents Lîsanī (لِسَان), which consists of four morphologically annotated corpora of the Yemeni, Iraqi, Sudanese, and Libyan Arabic dialects. We collected the text of the corpora from Facebook, Twitter, and YouTube. We then tokenized and manually annotated the text with morphological attributes. The annotation methodology we followed is similar to that advised for annotating Palestinian

and Lebanese dialects [2], [3], and based on SAMA tagsets [4]. To support and streamline the annotation process, we developed the Arabic Dialect Annotation Toolkit (ADAT) which we also provide as an open source online contribution.

Lîsañ consists of about 1.2 Million fully-annotated tokens: Yemeni (1.1M), Iraqi (46K), Libyan (52K), and Sudanese (52K). Thirty-five annotators helped annotate the corpora. They are graduating or senior univeristy students from different academic backgrounds and each is a native speaker of the dialect they annotated. Lîsañ and ADAT are available freely at (https://sina.birzeit.edu/currasat/) for academic research purposes.

The rest of the paper is organized as follows. Section II overviews the related work, Section III describes the corpora collection process, and Section IV presents the annotation methodology. We evaluate the corpora and the annotation process in Section VI. We conclude and discuss future directions in Section VII.

II. Related Work

This section reviews efforts to create annotated corpora for Arabic dialects.

An early Treebank [5] was created for the Jordanian dialect. A Palestinian dialectal corpus (called Curras [2] [6]) consists of 56K tokens collected from Facebook and scripts of the Palestinian series "Watan Aa Watar". Each word in the corpus was then manually annotated with a set of morphological attributes. Curras was recently revised (*Curras 2*) and extended with a Lebanese corpus (10k Tokens) to form a more Levantine corpus [3].

The Egyptian Arabic corpus CALLHOME [7] consisted of transliterations of telephone conversations in Egyptian. ECAL [8] was built on CALLHOME to provide morphological analysis of the Egyptian dialect. An extension of ECAL was CALIMA [9]. The COLABA project [10] gathered resources in dialectal Arabic from online blogs. This combination of projects gradually led to constructing the Egyptian TreeBank (ARZATB) [11].

Other efforts to create morphologically annotated corpora follow. [12] presented a 200K tokens corpus for seven different Arabic dialects including Taizi (Yemen), Sanaani (Yemen), Najdi (KSA), Jordanian, Syrian, Iraqi, and Moroccan. MADAR [13] is an ongoing multi-dialect corpora covering 26 different cities and their corresponding dialects. The work in [14] presented the first release of an Arabizi Tunisian corpus (42K tokens). The GUMAR Emirati dialect corpus consists of about 200K tokens collected from Emirati novels [15].

Two NLP competition tasks on *nuanced Arabic dialect identification* (NADI) in 2021 [16] and 2019 [17] provided researchers with Arabic dialect data from 21 countries. NADI targeted the identification of 100 different province-level dialects in 21 Arab countries. They provided competitors with 21,000 tweets labeled with a province-level dialect in addition to a 10 million tweet dataset with no labels.

NADI followed the *fine-grained Arabic dialect identification* task [18] that targeted identifying up to 25 city-level dialect variations in addition to MSA. The task provided two corpora: (i) The first is composed of 10,000 *basic travel expression corpus* (BTEC) sentences [19] translated to the dialects of five main cities. (ii) A separate set of 2,000 BTEC sentences translated into 25 city dialects.

TABLE I
Number of documents and tokens per corpus

Corpus	Tokens	Documents
Yemeni	1,098,222	38,819 Tweets
Iraqi	45,881	3,326 Threads
Libyan	51,686	3,053 Threads
Sudanese	52,616	3,000 Threads
Total	**1,248,405**	**48,198**

III. Corpus Collection

We collected Lîsañ from social media networks, mainly from Twitter, Facebook, and YouTube. Three corpora (Iraqi, Libyan, and Sudanese) were collected manually from Facebook and YouTube, while the Yemeni corpus was collected automatically. The manual collection was carried out by native speakers who carefully selected public posts and comments discussing politics and general affairs. We required the selected comments and posts to be at least 10 words and not larger than 30 words. We also required them to have at least one colloquial word belonging to the target dialect.

The Yemeni corpus was collected through the Twitter API using keywords related to the current political situation in Yemen. To ensure that each of the collected tweets contained at least one colloquial Yemeni word, we filtered the tweets using a list of typical and distinctive colloquial Yemeni words. No specific sub-dialect was preferred in any of our four dialect corpora as we aimed to develop a general corpus for each dialect.

Table I provides general statistics about each corpus.

IV. Corpus Annotation Methodology

This section presents the approach we used to annotate our four corpora. First, we define the tags we used in the annotation; then we describe the tool and the methodology used to annotate each word in context. Figure 1 shows a Sudanese phrase with the word عيموتو /*ymwtw* (they will die) and its POS, Prefix, Suffix, Stem, Lemma, and Gloss annotations.

A. Annotation Framework

Based on the annotation framework used to annotate the Curras and Baladi corpora ([2], [3]), as well as the framework found in [4], we define our annotation framework of a token as a tuple:

⟨ *rawToken*, *Token*, *Prefixes*, *Stem*, *Suffixes*, *POS*, *Lemma*, *Gloss* ⟩.

Where *rawToken* is the raw word as it appears in the corpus; *Token* is a normalized version of *rawToken*; and *Prefixes*, *Stem*, and *Suffixes* form the segmentation of the *Token*. *POS* is the part-of-speech and *Lemma* is the lexicon conical form of the

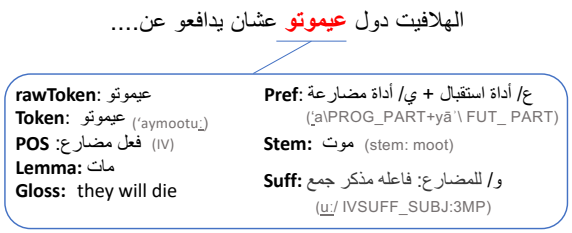

Fig. 1. Example of an annotated token in context

Token. The gloss is the meaning of the *Token* in English. The *Prefixes* and *Suffixes* use the '+' separator to separate parts from *Token* and the '/' separator to separate the prefix and suffix POS labels, respectively.

To ensure maximum compatibility with other MSA and dialectal corpora, we adopted the morphology tags used in LDC's BAMA and SAMA databases [4], which are commonly used for annotating Arabic corpora such as the Palestinian Curras [2], the Lebanese Baladi, and the Emirati Gumar [15]

B. Annotation Guidelines

Tokenization: the text was tokenized into sentences, then into raw tokens. A raw token can be a word, letter, symbol, punctuation mark, or emoji. Each token is given a unique identifier.

Token and Spelling Guidelines: A token is the normalized version of the *rawToken*. Because there are no standard orthographic spelling rules for dialects, people typically write words as they pronounce them. The same word can be written in many different ways, such as اللي */ālly* (the one) and الي */āly* , هذول */hḏwl* (those ones) and هذولا */hḏwlā* , مكو */mkw* (there is nothing) and ماكو */mākw* or شوية */šwyh* (a few) and شويا */šwyā* .

Also, people sometimes stress certain letters by repeating them, such as يسسسسس */yssssss* (yes), هههههه */hhhhhhh* (lol), and شووووو */šwwwww* (what).

In addition, unintentional typos and spelling mistakes are more likely to occur in social media content. More importantly, we noticed that people tend to concatenate some functional words (e.g., prepositions, pronouns, negation particles) with words, such as لتخافون */lt̲hāfwn* (do not be afraid originally لا تخافون */lā t̲hāfwn*).

The lack of such standard orthography makes the annotation process challenging. One solution is to develop a set of orthographic rules for each dialect, and rewrite the *rawToken* according to these rules. This solution (called CODA) was used in the annotation of the Palestinian, Lebanese, and Emarati corpora ([2], [3], [15]).

We tried to apply this solution to annotate our four corpora. However, we found it unteachable and did not scale to large and diverse corpora. Since we have many annotators participating in the annotation process, it was difficult to teach them CODA rules to maintain the consistency of their annotations. Instead of using CODA rules, we used the following simple normalization rules to produce *Token*.

1) Unintentional typos and spelling mistakes are corrected,
2) Letters repeated more than two times are removed,
3) Odd and rare spellings are corrected, if necessary, to follow more common spellings.

With these necessary changes, *Token* becomes a normalization, rather than a re-spelling, of the *rawToken*.

POS Guidelines: We used the exact SAMA POS tagset to specify part-of-speech of the stem of the *Token*.

Segmentation and Affixes Guidelines: Each token is segmented into prefix(es), stem, and suffix(es). Each prefix (and suffix) is tagged with its POS. Multiple prefixes and suffixes are combined with "+" (See Figure 1). To maintain compatibility with the tagset of SAMA affixes and related corpora, we used the SAMA POS affixes, in addition to some dialect-specific categories that we discovered and introduced during the annotation process.

As noted earlier, unlike MSA, some functional words (e.g., prepositions, pronouns, negation particles) are concatenated with words in the dialectal text. For example, in the word مايعرف */māyʕrf* (he does not know), the ما */mā* prefix plays the role of a negation particle, which we annotated as a prefix. Concatenating such different functional words as prefixes and suffixes yield a large number of prefixes-stem-suffixes combinations indeed.

Lemma Guidelines: Every token is linked with an MSA lemma. We used MSA lemmas from the SAMA database. In case a lemma is not found in SAMA, we used lemmas found in Birzeit's Lexicographic database (see [20] [21]) and Arabic Ontology [22]; otherwise, we introduced a new lemma. We note here that for dialectal lemmas, we added the following:

- Glosses (i.e., senses in Arabic), which is important for lexical semantics tasks such as word sense disambiguation ([23] [24]) and Word-in-Context WiC; and
- Equivalent lemmas in MSA, which is important to link different lemmas as synonyms, as done in [25] and [26].

Searching for an equivalent lemma in our lexicographic database is not straightforward; thus, we used a sophisticated algorithm that supports diacritic-based matching of Arabic lemmas [27].

Gloss Guidelines: This is an informal semantic annotation in English. By default, we used the glosses of the SAMA Lemmas, and edited them if needed. It is worth noting that the gloss formulation guidelines presented in this article for morphological annotations purposes are not the same guidelines we presented in [28] or [22] for semantic and ontology engineering purposes.

C. Annotation Methodology

Each token was annotated manually by native speakers using ADAT. ADAT supports the annotation guidelines described earlier. We recruited 35 native speakers and trained them on the annotation process and on the use of ADAT.

A separate team was established for each dialect, led by an expert native speaker. The training phase was divided into two steps.

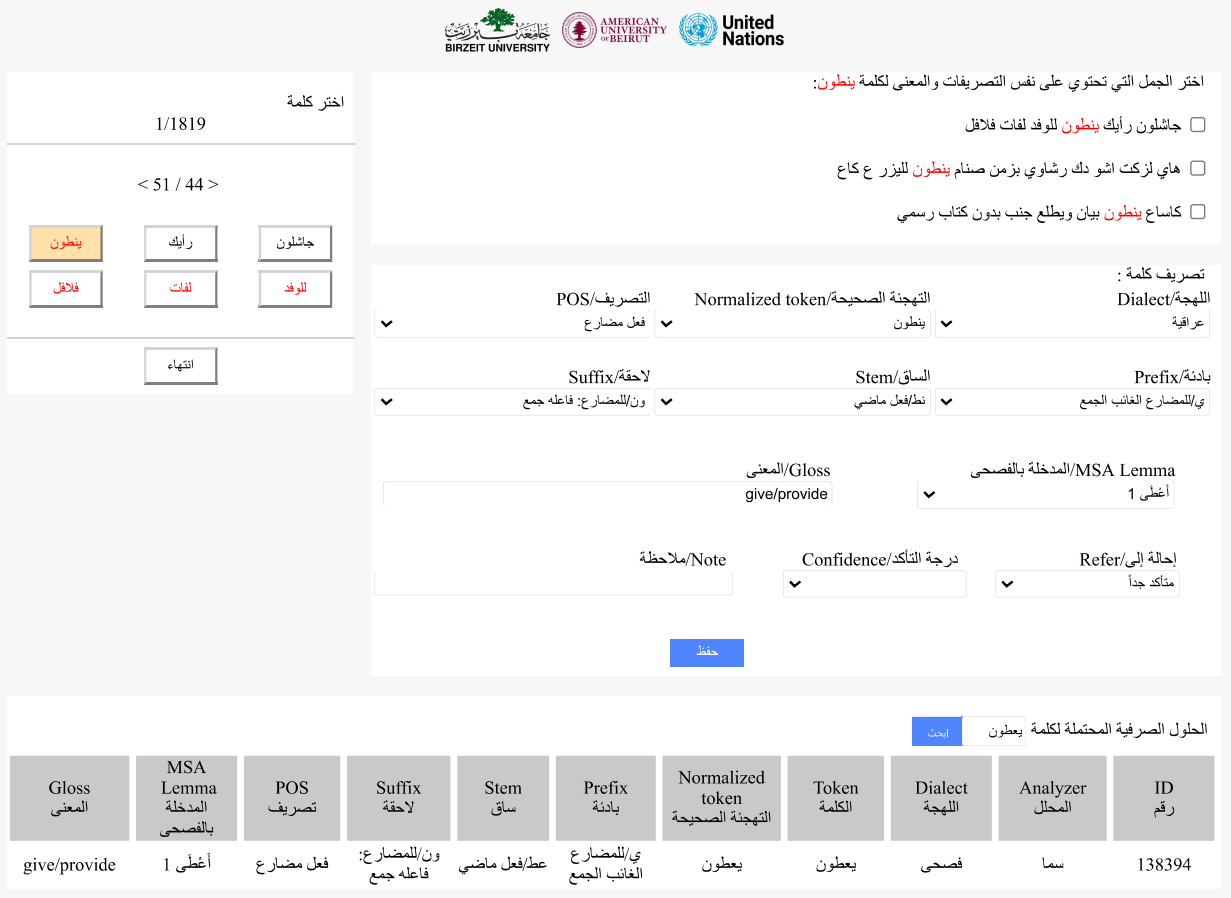

Fig. 2. Snapshot of ADAT in action

First, we conducted an online workshop for about 15 hours. The workshop explained the annotation process and the tagsets. It included an assisted annotation of a text with 200 tokens.

Second, each annotator was given a corpus of $1,000$ word tokens to annotate. This served as a quiz set. We evaluated the quality of the annotations and only the annotators with good quality were recruited. The 35 recruited annotators were paid a fair rate per hour based on their living locality (between 5$ and 10$ per hour). A discussion channel was also created to enable the annotators to discuss, post questions, and consult the leader on specific issues.

Each of the four corpora was uploaded to the tool and divided into tasks. Each task was assigned to an annotator to carry out, as described in Section IV-D. The annotation of the four dialects spanned over two years.

D. The Annotation Tool

To guide and speed up the annotation process, we developed ADAT (see a screenshot in Figure 2), which we designed to support collaborative annotations.

At the start, ADAT shows the task. Each task is a set of words that need to be annotated and that belong to a specific context (same tweet, post, comment). When the annotator clicks on a word w, ADAT retrieves the contexts (sentences) containing the w. The annotator selects those contexts where w appears with the same morphological features as in the original context, then annotate w, and apply this annotation to all the selected contexts.

ADAT also displays a set of possible morphological solutions for w, shown at the bottom of the annotation panel. ADAT retrieves these solutions from the following resources: (i) SAMA Database, (ii) Curras annotations, and (iii) previous annotations. The annotator can tick one or more sentences and select the appropriate solution for w. In case no acceptable solutions were retrieved by ADAT, fully or partially, the annotator can add or edit the annotations.

After completing the annotation of w, the annotator must select his/her degree of confidence (High, Normal, or Low). In case of hesitation about the annotation of a certain word, ADAT allows the annotators to "refer" the solution to another more experienced annotator for review.

The idea of offering the annotators a list of suggested annotations helps in speeding up the annotation process. More importantly, it is critical for minimizing errors and maintaining consistency. As will be discussed in Section V, we noticed several types of mistakes that the annotators made when typing manually, which we corrected afterward.

TABLE II
STATISTICS ABOUT THE CORPUS

Corpus name	Yemeni	Iraqi	Sudanese	Libyan
Tweets	38,819	3,326	3,000	3,053
Token	1,098,222	45,881	52,616	51,686
Unique Tokens	136,801	17,812	18,242	18,556
Unique Lemma	43,320	9,086	10,251	9,924
Nouns	627,907	26,550	28,557	27,761
Verbs	178,381	8,371	9,249	9,827
Functional words	260,655	10,097	13,347	12,954
Digit	3,962	128	7	177
Others (e.g., Foreign words)	27,317	735	1,456	967

E. Corpus Statistics

Lîsañ contains more than 1.2 million fully annotated tokens, represented by 48K documents (tweets, posts, and comments) collected from Yemeni, Iraqi, Sudanese, and Libyan dialects. Table II details the number of documents, tokens, and lemmas in each of the corpora. It also contains a number of unique tokens, lemmas, nouns, verbs, digits, functionals, and other tokens in Lîsañ.

V. ANNOTATION CORRECTIONS AND NORMALIZATION

Different types of human mistakes occurred while typing annotations, which need to be reviewed and corrected. Most of these mistakes are typos, POS variations, or syntactic issues, such as the following:

- Typing errors as in when the annotator entered مصارع instead of مضارع in a POS field
- Segmentation separator differences as some annotators used characters other than '+' to denote splitting. For example, one annotator would use و/عطف، ال/أداة تعريف instead of و/عطف + ال/أداة تعريف.
- Orderings of the group, voice, number, and gender indicators in POS tags can differ. For example, one annotator used (مضارع مفرد مذكر غائب) instead of the standard tag (للمضارع الغائب المذكر المفرد).

To correct such issues, we developed a suite of validation tools that helped through the process. The tools are used to group the annotations, by tag values, and show their frequencies. We prioritized normalizing the tags with higher frequencies. The tools also grouped annotation variants for tokens with similar POS or lemma annotations.

In correct annotations are then given to linguists to correct and normalize using automated replacements with regular expressions, and manual editing sometimes.

VI. CORPUS EVALUATION

To evaluate the quality of the annotations in Lîsañ, we first normalized the annotations as per Section V, and then computed inter-annotation agreement across the annotated features. For that, we ensured overlap in annotations and assigned overlapping annotation tasks such that more than 5% of all contexts are by design annotated by more than one annotator.

Moreover, Lîsañ benefited from the following quality methodological measures:

- We ensured that the collected texts belong to the dialect by asking the annotators to assign a dialect flag to the text. The annotators also provided a dialect-specific lemma if they could not find a suitable MSA lemma. This is to be differentiated from the lemma dialect field reported in the dataset which reports the name of the dialect as found by the annotator.
- ADAT's annotation methodology allows annotations of tokens across contexts. Each annotator views available annotations by other annotators for a specific token t_1 as follows. She selects a token t_1 in a context c_1 corresponding to a document. She then views other occurrences $t_2, \ldots, t_n$ of t_1 in other contexts $c_2, \ldots, c_n$ and also view their existing annotations. These annotations may also be the results of MSA morphological analysis of t_1.
  - The annotator has the ability to select one of the existing annotations or some of its feature annotations as annotations for t_1. This speeds up the annotation process and also provides a better opportunity for annotation overlap and review.
  - The annotator selects contexts $D_s \subseteq \{d_1, d_2, \ldots, d_n\}$ with occurrences of T_s that match the semantics of t_1 in c_1. This assigns the annotation of t_1 to its occurrences in D_s. This again improves opportunities for overlap and review.
- When an annotator is in doubt of the annotations, they can request a review from another annotator using the referral functionality.

A. Quantitative Evaluation

Given the overlap in the annotations, we performed feature and dialect-based inter-annotation agreement evaluations of Lîsañ. We applied the transformations from the normalization process in Section V to the second latest annotated overlap. We judged agreement after the normalization transformation by matching the annotation strings and by manual inspection of the annotation tags when the annotations slightly differed.

We used the Kappa-Cohen [29] metric as implemented in the Scientific Kit Learn (scikit-learn) python libraries [30]

TABLE III
LÎSAÑ INTER-ANNOTATION AGREEMENT RESULTS PER DIALECT

	Iraqi			Libyan			Sudanese			Yemeni		
Feature	IAA	UNQ	OVP	IAA	UNQ	OVP	IAA	UNQ	OVP	IAA	UNQ	OVP
Stem	.972	6,764	61,829	.975	7,072	65,670	.989	6,914	64,307	.981	25,237	1,366,425
Lemma	.933	8529	61,828	.930	9,194	65,670	.944	9,562	64,307	.948	35,503	1,366,482
LemmaD	.894	7	60,120	.904	7	65,282	.926	7	63,818	.899	18	1,335,496
POS	.950	147	61,610	.953	165	65,468	.970	117	64,222	.956	447	1,362,675
Prefix	.975	188	128,233	.976	280	135,725	.981	133	133,048	.982	788	2,827,159
Suffix	.920	265	128,257	.941	338	136,394	.938	189	132,558	.921	1,397	2,839,170
POS-P	.802	496	61,609	.785	648	65,521	.795	620	64,174	.813	3,063	1363800
POS-X	.874	806	61,090	.871	1,041	65,128	.877	881	63,689	.870	3,973	1,350,781
Voc	.978	15,400	61,796	.984	15,940	65,668	.989	15,777	64,306	.990	106,878	1,366,469

to compute the inter-annotation agreement where the number of unique categories was below 10,000. We used our own implementation when it surpassed 10,000 as the scikit-learn implementation crashed for a higher number of categories such as in the Lemma and Stem features.

Table III shows the inter-annotation agreement (**IAA**) scores for Lîsañ for each of the Yemeni, Libyan, Sudanese, and Iraqi dialects. The numbers are presented for each feature including the stem, lemma, dialect lemma (LemmaD), part of speech (POS), prefix, suffix, part of speech of prefix (POS-P), part of speech of the suffix (POS-X) and vocalization (VOC). **UNQ** and **OVP** denotes the number of normalized unique categories, and the total number of overlaps per feature, respectively.

B. Discussions

Our quantitative analysis shows very high agreement between the annotators. Annotators often mentioned an original dialect of the Lemma such as Egyptian, Palestinian, or Saudi outside the four dialects of Lîsañ which explains the values of 7 and 18 in the LemmaD row.

The number of Prefix and Suffix overlaps is almost double the prefix/POS as they show the agreement on the split prefixes and suffixes. While affix (prefix and suffix) segmentation scored high agreement across annotators, the two features showed very good, yet less agreement when we compared them along with their assigned POS tags. This is expected as the annotators needed to be more specific and used variations of affix POS tags assigned to each separate affix segment.

VII. CONCLUSION AND FUTURE WORK

In this paper, we presented four morphologically-annotated corpora for low-resourced Arabic dialects, which consisted of more than 1.2 million tokens. Each word in the four corpora was annotated with several morphological features. The annotation process was carried out by 35 native speakers of the target dialects who were trained on a set of guidelines and on how to use ADAT. We developed ADAT to assist the annotators and to ensure compatibility with SAMA and Curras tagsets. ADAT is developed as an open-sourced contribution and the four corpora will also be available online https://sina.birzeit.edu/currasat/.

As we have developed several morphologically annotated corpora for six dialects so far, we plan to represent and integrate all of them using the W3C Lemon model, similar to work on representing Arabic lexicons [31].

We plan to also build specialized lexicons for the four dialects, including normalizing and unifying subsets of the POS tagsets. We will enrich these lexicons with senses for each lemma. Last but not least, we plan to use the four corpora to extend Wojood [32] by annotating the corpora for Named Entity Recognition, similar to what we did with Curras and Baldi.

ACKNOWLEDGEMENTS

We would like to thank the 35 annotators and the Aklama the team who carried out the annotations of the four dialects. We also acknowledge efforts made by students at the University of Birzeit in reviewing and correcting annotations. We would like to thank Rayan Dankar in helping us develop the first version of the Adat tool and Archetlabs developers for their technical support.

REFERENCES

[1] K. Darwish, N. Habash, M. Abbas, H. Al-Khalifa, H. T. Al-Natsheh, H. Bouamor, K. Bouzoubaa, V. Cavalli-Sforza, S. R. El-Beltagy, W. El-Hajj, M. Jarrar, and H. Mubarak, "A panoramic survey of natural language processing in the arab worlds," *Commun. ACM*, vol. 64, p. 72–81, April 2021.

[2] M. Jarrar, N. Habash, F. Alrimawi, D. Akra, and N. Zalmout, "Curras: An annotated corpus for the palestinian arabic dialect," *Journal Language Resources and Evaluation*, vol. 51, pp. 745–775, September 2017.

[3] K. E. Haff, M. Jarrar, T. Hammouda, and F. Zaraket, "Curras + baladi: Towards a levantine corpus," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, (Marseille, France), June 2022.

[4] M. Maamouri, D. Graff, B. Bouziri, S. Krouna, A. Bies, and S. Kulick, "Ldc standard arabic morphological analyzer (sama) version 3.1," *LDC2010L01*, July 2010.

[5] M. Maamouri, A. Bies, T. Buckwalter, M. Diab, N. Habash, O. Rambow, and D. Tabessi, "Developing and using a pilot dialectal Arabic treebank," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, (Genoa, Italy), European Language Resources Association (ELRA), May 2006.

[6] M. Jarrar, N. Habash, D. Akra, and N. Zalmout, "Building a corpus for palestinian arabic: a preliminary study," in *Proceedings of the EMNLP 2014, Workshop on Arabic Natural Language*, pp. 18–27, Association For Computational Linguistics, October 2014.

[7] A. Canavan, G. Zipperlen, and D. Graff, "Callhome egyptian arabic speech," *LDC97S45*, 1997.

[8] H. Kilany, H. Gadalla, H. Arram, A. Yacoub, A. El-Habashi, and C. McLemore, "Egyptian colloquial arabic lexicon," *LDC99L22*, jul 2002.

[9] N. Habash, R. Eskander, and A. Hawwari, "A morphological analyzer for Egyptian Arabic," in *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, (Montréal, Canada), pp. 1–9, Association for Computational Linguistics, June 2012.

[10] M. Diab, N. Habash, O. Rambow, M. Altantawy, and Y. Benajiba, "Colaba: Arabic dialect annotation and processing," *LREC Workshop on Semitic Language Processing*, pp. 66–74, 01 2010.

[11] M. Maamouri, A. Bies, S. Kulick, M. Ciul, N. Habash, and R. Eskander, "Developing an Egyptian Arabic treebank: Impact of dialectal morphology on annotation and tool development," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, (Reykjavik, Iceland), pp. 2348–2354, European Language Resources Association (ELRA), May 2014.

[12] F. Alshargi, S. Dibas, S. Alkhereyf, R. Faraj, B. Abdulkareem, S. Yagi, O. Kacha, N. Habash, and O. Rambow, "Morphologically annotated corpora for seven Arabic dialects: Taizi, sanaani, najdi, jordanian, syrian, iraqi and Moroccan," in *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, (Florence, Italy), pp. 137–147, Association for Computational Linguistics, Aug. 2019.

[13] H. Bouamor, N. Habash, and K. Oflazer, "A multidialectal parallel corpus of Arabic," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, (Reykjavik, Iceland), pp. 1240–1245, European Language Resources Association (ELRA), May 2014.

[14] E. Gugliotta and M. Dinarelli, "TArC: Tunisian Arabish corpus, first complete release," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, (Marseille, France), pp. 1125–1136, European Language Resources Association, June 2022.

[15] S. Khalifa, N. Habash, F. Eryani, O. Obeid, D. Abdulrahim, and M. Al Kaabi, "A morphologically annotated corpus of emirati Arabic," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, (Miyazaki, Japan), European Language Resources Association (ELRA), May 2018.

[16] M. Abdul-Mageed, C. Zhang, A. Elmadany, H. Bouamor, and N. Habash, "NADI 2021: The second nuanced Arabic dialect identification shared task," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, (Kyiv, Ukraine (Virtual)), pp. 244–259, Association for Computational Linguistics, Apr. 2021.

[17] M. Abdul-Mageed, C. Zhang, H. Bouamor, and N. Habash, "NADI 2020: The first nuanced Arabic dialect identification shared task," in *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, (Barcelona, Spain (Online)), pp. 97–110, Association for Computational Linguistics, Dec. 2020.

[18] M. Salameh, H. Bouamor, and N. Habash, "Fine-grained Arabic dialect identification," in *Proceedings of the 27th International Conference on Computational Linguistics*, (Santa Fe, New Mexico, USA), pp. 1332–1344, Association for Computational Linguistics, Aug. 2018.

[19] K. Kageura and G. Kikui, "A self-referring quantitative evaluation of the ATR basic travel expression corpus (BTEC)," in *Language Resources and Evaluation (LREC'06)*, May 2006.

[20] M. Jarrar and H. Amayreh, "An arabic-multilingual database with a lexicographic search engine," in *The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019)*, vol. 11608 of *LNCS*, pp. 234–246, Springer, June 2019.

[21] D. Alhafi, A. Deik, and M. Jarrar, "Usability evaluation of lexicographic e-services," in *The 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1–7, IEE, November 2019.

[22] M. Jarrar, "The arabic ontology - an arabic wordnet with ontologically clean content," *Applied Ontology Journal*, vol. 16, no. 1, pp. 1–26, 2021.

[23] S. Malaysha, M. Jarrar, and M. Khalilia, "Context-gloss augmentation for improving arabic target sense verification," in *Proceedings of the 12th International Global Wordnet Conference (GWC2023)*, Global Wordnet Association, Jan 2023.

[24] M. Al-Hajj and M. Jarrar, "Arabglossbert: Fine-tuning bert on context-gloss pairs for wsd.," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, (Online), pp. 40–48, INCOMA Ltd., sep 2021.

[25] S. Ghanem, M. Jarrar, R. Jarrar, and I. Bounhas, "A benchmark and scoring algorithm for enriching arabic synonyms," in *Proceedings of the 12th International Global Wordnet Conference (GWC2023)*, pp. 215–222, Global Wordnet Association, Jan 2023.

[26] M. Jarrar, E. Karajah, M. Khalifa, and K. Shaalan, "Extracting synonyms from bilingual dictionaries," in *Proceedings of the 11th International Global Wordnet Conference (GWC2021)*, pp. 215–222, Global Wordnet Association, Jan 2021.

[27] M. Jarrar, F. Zaraket, R. Asia, and H. Amayreh, "Diacritic-based matching of arabic words," *ACM Asian and Low-Resource Language Information Processing*, vol. 18, pp. 10:1–10:21, December 2018.

[28] M. Jarrar, "Towards the notion of gloss, and the adoption of linguistic resources in formal ontology engineering," in *Proceedings of the 15th international conference on World Wide Web (WWW2006)*, pp. 497–503, ACM Press, New York, NY, May 2006.

[29] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, 2015.

[30] scikit learn, "sklearn.metrics.cohen_kappa_score," 2022.

[31] M. Jarrar, H. Amayreh, and J. P. McCrae, "Representing arabic lexicons in lemon - a preliminary study," in *The 2nd Conference on Language, Data and Knowledge (LDK 2019)*, vol. 2402, pp. 29–33, CEUR Workshop Proceedings, May 2019.

[32] M. Jarrar, M. Khalilia, and S. Ghanem, "Wojood: Nested arabic named entity corpus and recognition using bert," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, (Marseille, France), June 2022.