Construction and Annotation of the Jordan Comprehensive Contemporary Arabic Corpus (JCCA)

Majdi Sawalha,[†] Faisal Alshargi,[⋆] Abdallah Alshdaifat,[†]
Sane Yagi,[‡] and Mohammad A. Qudah[†]
[†]University of Jordan, Jordan [⋆]Universität Leipzig, Germany
[‡]University of Sharjah, UAE
sawalha.majdi@ju.edu.jo, alshargi@informatik.uni-leipzig.de,
a.shdaifat@ju.edu.jo, saneyagi@yahoo.com, m.qudah@ju.edu.jo

Abstract

To compile a modern dictionary that catalogues the words in currency, and to study linguistic patterns in the contemporary language, it is necessary to have a corpus of authentic texts that reflect current usage of the language. Although there are numerous Arabic corpora, none claims to be representative of the language in terms of the combination of geographical region, genre, subject matter, mode, and medium. This paper describes a 100-million-word corpus that takes the British National Corpus (BNC) as a model. The aim of the corpus is to be balanced, annotated, comprehensive, and representative of contemporary Arabic as written and spoken in Arab countries today. It will be different from most others in not being heavily-dominated by the news or in mixing the classical with the modern. In this paper is an outline of the methodology adopted for the design, construction, and annotation of this corpus. DIWAN (Al-Shargi and Rambow, 2015) was used to annotate a one-million-word snapshot of the corpus. DIWAN is a dialectal word annotation tool, but we upgraded it by adding a new tag-set that is based on traditional Arabic grammar and by adding the roots and morphological patterns of nouns and verbs. Moreover, the corpus we constructed covers the major spoken varieties of Arabic.

1 Introduction

A collection of texts in machine-readable format is called a corpus. The creation of a corpus is often motivated by interest in linguistic phenomena. Therefore, the design and creation of a corpus is always linked to purpose of usage. Thousands of corpora have been created and many are freely available. These corpora vary in size, type, format, usage, and purpose of creation. They are usually annotated with morphological, syntactic, semantic, discoursal, or prosodic information. Individual texts in a corpus often have meta-data in the header that give information about such attributes as genre of the text, author, source, date and country of publication, etc. (Baker et al., 2006).

Building a balanced and representative corpus remains an ideal goal for corpus creators. A balanced corpus includes a wide range of texts from the different genres and domains that the corpus claims to depict. Sometimes, this type of corpus is referred to as a reference, general, or core corpus. Similarly, a corpus is claimed to be representative if it contains the major linguistic variation in the concerned language. Although it is not an easy task to achieve balanceness and representiveness in a corpus, it can be done with a level of approximation and scalability (McEnery and Hardie, 2012; Baker et al., 2006).

The web provides a massive collection of texts which is growing rapidly. Constructing corpora by harvesting web pages is usually referred to as web-crawling. The web is an excellent information source with large amounts of data which one can select, organize, and compile into corpora of all types (McEnery and Hardie, 2012). Since the late 1980s, Arabic corpora have been constructed. However, not many of them are freely available as open-source. Most are for written Modern Standard Arabic (MSA). Morphosyntactically annotated Arabic corpora are very rare and not freely available to researchers.

This paper reports on the construction and annotation of a comprehensive 100-million-word corpus of contemporary Arabic. The purpose is to provide an open-source corpus of contemporary Arabic which is balanced, representative of the language, and comparable to the internationally recognized British National Corpus. The text of the corpus was selected from a wide range of genres, domains, and types. It consists of 83% written language and 17% spoken language. The texts of the corpus were collected primarily from text materials available online but also from the

Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 148–157
Florence, Italy, August 1, 2019.

transcripts of purpose-made recordings (see Section 3). The corpus was automatically annotated both morphologically and syntactically. A sample of one million words was manually and semi-manually verified; it was additionally annotated for sentiment and glossed in English. To accomplish this annotation, we used DIWAN (Al-Shargi and Rambow, 2015) but had to specifically develop for it morphological and syntactic annotation schemes on the basis of the long-established Arabic linguistic tradition (see Section 5). We also added new features to the DIWAN annotation tool to facilitate our semi-manual annotation process (see Section 6).

2 Literature Review

Arabic corpora vary in size, type, purpose, design, text type, etc. (Al-Sulaiti and Atwell, 2006). Zaghouani, 2017 surveyed freely available Arabic corpora and classified 66 of them into six main categories, namely (i) raw text corpora, (ii) annotated corpora, (iii) lexicons, (iv) speech corpora, (v) handwriting recognition corpora and (vi) miscellaneous corpora.

The Corpus of Contemporary Arabic (Al-Sulaiti and Atwell, 2006) was the first freely available Arabic corpus. Around one million words were collected from newspapers and magazines. Since then, most monolingual Arabic corpora have been constructed by collecting texts from news sources (i.e. newspaper articles). Examples of such corpora are: the Open Source Arabic Corpora (OSAC) which contain around 18 million words of written MSA and Classical Arabic (CA) texts (Saad and Ashour, 2010); Akhbar Al Khaleej 2004 Corpus consists of 3 million words of newspaper texts (Abbas and Smaïli, 2005); Al-Watan 2004 Corpus contains 10 million words of newspaper texts as well (Abbas et al., 2011); KACST Arabic Corpus includes more than 700 million words collected from 10 text source types such as newspapers, magazines, books, old manuscripts, university theses, refereed periodicals, websites, curricula, news agencies, and official prints (Al-Thubaity, 2015). There is also the International Corpus of Arabic (ICA) which was constructed by Bibliotheca Alexandrina and it contains 100 million words that were collected from the press, net articles, books, and academic text sources (Alansary and Nagi, 2014). The ArabiCorpus at Brigham Young University is one of the most popular web-based corpora. It consists of around 174 million words, 77% of which is from newspapers. It does, however, include around 9 million words of premodern literature, 1 million words of modern literature, 28 million words of non-fiction, and a token of colloquial Egyptian (0.164 million words).

The King Saud University Corpus of Classical Arabic (KSUCCA) consists of around 50 million words (Alrabia et al., 2014). The corpus includes texts of six genres, namely religion, linguistics, literature, science, sociology, and biography. The arTenTen corpus used web crawlers to automatically harvest 5.8 billion words from Arabic websites (Belinkov et al., 2013). Its purpose was linguistic and lexicographic in nature. It was automatically annotated using MADAMIRA and it is available on Sketch Engine.

The Historical Arabic Corpus (HAC) has 45 million words that were organized into primary and secondary resources, seven genres, and 100-year eras in the Gregorian calendar. Its intended purpose is historical semantics and etymological lexicography (Ismail et al., 2014).

Two specialized Arabic corpora use the Quran as a source of their textual content; hence, each consists of the same number of words in the Quran, 77430 words. The Quranic Arabic Corpus is morphologically and syntactically annotated. Its annotation was done automatically and verified collaboratively by the wider community (Dukes et al., 2013). The second corpus is the Boundary Annotated Quran Corpus. It is annotated with prosodic information and phrase boundaries (Brierley et al., 2012; Sawalha et al., 2012). It took advantage of boundary markups that flag starts and stops in the Quran (Sawalha et al., 2014; Brierley et al., 2016). Interest in dialectal Arabic corpora has recently surged. An example of such corpora is the Curras Palestinian Arabic corpus, a corpus of more than 56K tokens, which are annotated with morphological and lexical features (Jarrar et al., 2017). There are Arabic corpora that are only available for a fee, such as the Linguistic Data Consortium's[1] *The Penn Arabic Treebank*[2] and the European Language Resources Association's[3] *An-Nahar Newspaper Text Corpus*[4].

[1] https://www.ldc.upenn.edu/
[2] https://catalog.ldc.upenn.edu/LDC2016T02
[3] http://catalogue.elra.info/en-us/
[4] catalogue.elra.info/en-us/repository/browse/ELRA-W0027/

This brief review, which is based on a more extensive survey of the literature, points to the absence of resources that make the claim that they represent **in a comprehensive manner** the Arabic of today **as written and spoken** by contemporary native speakers. There is a great need for a corpus of modern Arabic as used by present-day native speakers of the language. The corpus must be truly representative of the language that the current inhabitants of the Arab World use, regardless of whether it is of the high or low variety. It must also be balanced in its representation of the written and spoken language, and of the various discourse genres. It must truly depict the language of the curricula and academia.

3 Methodology

To ensure that this corpus of modern Arabic is representative, balanced, comprehensive, and for general purposes, we followed the model of the British National Corpus (BNC)[5]. That is why this corpus contains slightly more than 100-million words of the same text types, domains, and genres. The corpus contains 87% of texts from written sources and 13% of transcribed spoken language. The written part includes texts from Applied Sciences, Arts, Belief and Thought, Commerce and Finance, Imaginative works, Leisure, Natural and Pure Sciences, Social Sciences, and World Affairs. The spoken subcorpus includes transcripts of Spontaneous Conversations (4.2%) and Context-Governed Spoken Language (6.2%) from the categories of Educational/Informative, Business, Public/Institutional, and Leisure. Tables 1 and 2 show the text categories of the corpus of the written and spoken subcorpora respectively.

Twenty million words of the category of World Affairs were selected from newspapers published in 20 Arab countries where around one million words were collected for each country from one or two newspapers published in that country. The different genres of newspaper articles include Politics; Arts and Culture; Economics; Local News; Opinions; Regional and International News; Sports; and Others (e.g., Weather Forecasts, News about Technology, Health, Tourism, etc.). The subcategory of Social Sciences includes around 14 million words of texts from books and online sources. It contains texts of the genres: Languages and Linguistics; Modern Arabic Dictionaries; Philosophy; Islamic Studies and Quran Interpretation; History; Geography; Anthropology and Sociology; Law; Education; Food and Nutrition; Travel; Lectures; Sports; etc. The subcategory of Belief and Thought consists of about three million words of texts of sacred books such as: the *Quran*; Quran Interpretation; the Hadith including *Hadith Qudsi*; the *Old Testament*; the *New Testament*; *Dictionary of the Bible*; and Interpretations of the Testaments, etc.

More than seven million words were collected from online sources to fill the subcategory of Commerce and Finance. These articles belong to a variety of topics within the commerce and finance genre. They include Accounting; Taxes; Investment; Finance; Financial Legal Issues; Inventory; Currency, etc. The subcategory of Imaginative Language consists of 16 million words. The texts were collected from written sources that include; stories; novels; poetry; plays; translations of international stories and novels. The subcategory of Leisure consists of 12 million words which include articles on topics such as Animals; Cars; Technology; Health; Women; Tourism; Cooking Recipes; How to; Arabian Cities; Jordanian Stories and Traditions; and Fitness. The subcategory of Arts was collected from web sources and comprises around seven million words. The texts of this category contain articles on Arts; Digital Photography; Film and Video Production; Printing; Area Planning and Landscaping; Sculpture; Ceramics and Metals; Computer Graphic Arts; Entertainment and Performance; Cinema and Theater; Photography; Music; Architecture; Fine Arts; Decorative Arts; International Arts; Arabic Calligraphy, etc. Around seven million words were collected from books and web resources for the category of Applied Sciences. The topics included in this category are Medicine; Engineering; Information Technology; Energy, etc. Finally, the Natural and Pure Sciences subcorpus consists of around four million words that come from Mathematics, Physics, Chemistry, Biology, etc.

The corpus is designed to have detailed metadata about each article. This is valuable knowledge that can be used to guide the search within the corpus. It can also be used in text classification and text data mining. Moreover, the corpus and its metadata constitute an excellent dataset for training machine learning algorithms on such tasks as genre identification. The metadata include infor-

[5] http://www.natcorp.ox.ac.uk/

Written Subcategory	Words(million)	Percentage
Applied Sciences	7	8%
Arts	7	7.50%
Belief and Thought	3	4%
Commerce and Finance	7.2	8.30%
Imaginative	16	18.50%
Leisure	12	14%
Natural and Pure Sciences	3.8	4%
Social Sciences	14	16%
World Affairs	20	20%
Total	**90**	**100** %

Table 1: Text categories of the written subcorpus

mation such as article ID, category, subcategory, country, author, date, source, and URL. Moreover, the title and the text of each article are stored in a traditional relational database and in XML format. Figure 1 shows a sample article in XML with the text, title and metadata clearly specified.

A corpus-representative snapshot of one million words are designated as the corpus gold standard. This is a sample of words semi-manually annotated and verified. Each word is morphologically decomposed into its prefixes, stem, suffixes, proclitics, and enclitics. Then, each morpheme is annotated with a morphological tag or possibly tags. The stem is labeled by one morphological tag, and its root and morphological pattern are specified. Other morphological attributes, such as the number and gender of a noun, are indicated as well. The tag set we used here was informed by traditional Arabic grammar (see Section 6). Moreover, each word was annotated for sentiment designation (i.e., positive, negative, or neutral sentiment). The annotation process was done using a specialized program, DIWAN (Al-Shargi and Rambow, 2015). Twenty annotators with expertise in Arabic linguistics were trained on the tag set and on the annotation tool and they were supervised by three linguists who ensured the accuracy of annotation and verification.

```
<Article Article_ID="11966" Category="World Affairs" SubCategory="محليات"
Date="2017-12-27 18:58:00" Author= "عمان -الرأي" source = "Alri جريدة الرأي"
URL="http://alrai.com/article/10418936" Country="Jordan" >
<Title>تكريم المشاركين في مسابقة الخطابة العربية في " الأردنية"</Title>
<Text>
كرمت الجامعة الأردنية الداعمين والمشاركين في فعاليات مسابقة الخطابة العربية للطلبة الناطقين بغير
العربية.ونظمت المسابقة لجنة الوافدين في كلية الآداب بدعم من مكتب شؤون الوافدين في الجامعة وأقيمت الشهر
الحالي تستهدف النهوض باللغة العربية لدى الطلبة الوافدين من جنسيات متعددة.وعرضت منسقة لجنة الوافدين
الطالبة نورماس القدومي أهداف ونشاطات اللجنة والدور الذي تقوم به في مساعدة طلبة من جنسيات عالمية لتعلم
العربية بهدف إكسابهم المهارات اللغوية لتكون عونا لهم في التواصل الاجتماعي مع طلبة الجامعة وتسهيل
دراستهم الجامعية وإقامتهم داخل المملكة الأردنية الهاشمية</Text>
</Article>
```

Figure 1: A stored article with XML markup.

Context-Governed Subcategory	Words(million)	Percentage
Educational / Informative	1.6	26 %
Business	1.3	21 %
Public/ Institutional	1.7	27 %
Leisure	1.6	26 %
Total	**6.2**	**100** %

Table 2: Categories of the Context-Governed Spoken Language subcorpus.

4 Copyrights

The texts of the written subcorpus were primarily selected from sources available online. To get around copyrights, we followed Eckart's example by 'scrambling' the texts such that the original structure of a document would be destroyed. "This inhibits the reconstruction of the original documents. With respect to German copyright legislation this approach is considered safe" (Eckart et al., 2014). We assume this is satisfactory to copyright laws in most countries around the world.

5 Annotation

To create and annotate the comprehensive corpus of contemporary Arabic, we followed the principles presented in (AlShargi et al., 2016). This approach consists of several main steps. We started out by deciding on the categories, subcategories, and sizes in millions of words of the components of the corpus. To ensure balance, we simply followed the BNC proportions. Then we collected the target textual material from sources similar to those of the BNC, as well. We added texts from the social media, forums, and websites according to the various topical categories (cf. Tables 1, 2. Then, we modified the DIWAN annotation tool (Al-Shargi and Rambow, 2015) by adding new annotation tags such as root, pattern, and sentiment, by creating an elaborate CODA, and by developing a user interface that reflects these modifications. (See Tables 4, 5, 6 where the new tags we added appear in bold). After the primary annotation of the entire corpus was run automatically, we conducted an error detection round to find and correct annotation errors. (Figure 2 shows the workflow).

DIWAN assists human annotators in tagging each token with the relevant morphological, syntactic, and semantic information. DIWAN has the following annotation fields: 1) ***Diac***: where the word to be annotated is shown with diacritics. 2) ***Lex***: Here the lemma in its citation form appears. For example, the lemma of واحبابها *wAHbAbhA*

'and her lovers' will have the lex حبيب *Hbyb* 'lover' 3) ***BWhash***: In this field, the Buckwalter rendition of the lemma is split into prefix, stem, and suffix. The stem is marked by the symbol # on both sides, 4) ***Gloss***: the English translation of the lemma appears in this field.

There are features in DIWAN that indicate the proclitics and enclitics of words. The clitics are assigned slots: prc3, prc2, prc1, and prc0 for proclitics, and enc0; enc1, and enc2 for enclitics. A lower index indicates closer proximity to the stem. Additionally, there are features that mark the part of speech (POS), functional number and gender of nouns, and aspect of verbs. Functional number and gender refer to the function of a word, rather than its form. For example قادة *qAdp* 'leaders' is functionally masculine and plural, even though it ends in ة, which is the marker of feminine singular nouns.

We added three new features to DIWAN, (i) **root** which is a base form, for example لمس *lms* to touch is the root of these two words سيلمسونها *sylmswnhA* they will touch it and يلمس *ylms* 'he touches', (ii) **sentiment** which shows the attitude towards a word as to whether it is negative, positive, or neutral; for example, the sentiment annotation of the word 'sabba' in سب العدو *sb AlEdw* 'he cursed the enemy' is negative while that of the word 'ahabba' in أحب المرأة *>Hb Almr>p* 'he loved the woman' is positive and that of عمان *EmAn* 'Amman' is neutral. And (iii) **pattern** the morphological mold that the root is formed by; e.g., the word كاسِر *kAsir* breaker is derived by the mold فاعِل *fAEil* doer and the root كَسَرَ *kasara* he broke . To show the details of the annotation, we present table 3.

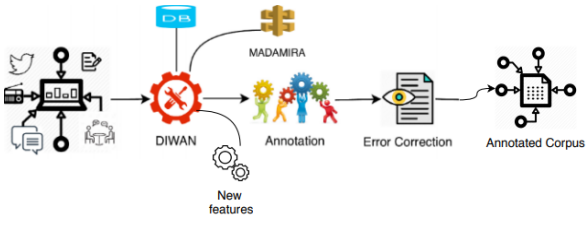

Figure 2: Steps to Creating a Comprehensive Corpus for Contemporary Arabic

6 Morphology

Morphological annotation of the whole corpus was automatically performed using MADAMIRA (Pasha et al., 2014). We isolated a one-million word snapshot of the corpus for manual verification. Twenty-five B.A. students of Arabic at the University of Jordan carried out the manual verification and two professors of linguistics supervised their work and vetted their annotation. The annotators used DIWAN (Al-Shargi and Rambow, 2015) to review and verify MADAMIRA's analysis. The morphological annotation required (1) Development of a new tag-set with detailed morphological description. Fourteen new noun-tags were added to Madamira. These new tags fall into three groups: i) derived nouns: Active participle, Passive participle, Exaggeration, Qualificative adjective, Noun of time/place, Noun of Instrument, and Elative noun; ii) underived nouns: Concrete noun and Abstract noun; and iii) gerunds: Original gerund, Gerund with initial miim, Gerund of instance, Gerund of state, and Gerund of profession. (2) Providing the roots of the nouns and verbs, since such a root conveys the core lexical meaning of a word. It normally consists of three consonants, and less frequently of two or four consonants. The majority of Arabic words (nouns and verbs) are derived from triliteral roots, uncommonly from biliteral or quadriliteral roots. For instance, the consonantal root د . ر . س *d.r.s* has the basic lexical meaning of studying, from which these words are derived: دَرْسٌ *darosN* 'lesson', مُدَرِّس *mudar~is* 'teacher ', دِرَاسَة *diraAsap* 'sutdying', مَدْرَسَة *madorasap* 'school', دَارِس *daAris* 'student'. In all these derived words, the consonants d-r-s constitute their root (McCarthy, John, 1981; Prunet et al., 2000; Davis and Zawaydeh, 2001). (3) Providing the morphological pattern of each noun and verb. This pattern constitutes a canonical template that consists of a series of discontinuous consonants including those of the root, a series of discontinuous vowels, and a templatic pattern. It carries a schematic meaning and grammatical information together including the word's part of speech. For instance, the morphological pattern C1VVC2VC3 together with the vowel melody - a - i - represents the active participle of Form I verbs (Bat-El, 1994, 2001; Ratcliffe, Robert , 1998; Ussishkin, Adam, 1999, 2005).

7 Spoken vs Written Language

Languages often have a low variety that is used in everyday communication and a high variety that is used in formal settings. The spoken language

Analyze	Sentence					
Sentence	الدولي	بالقانون	الاستهتار	في	الشركة	امعنت
BW	Aldwly	bAlqAnwn	AlAsthtAr	fy	Al$rkp	<mEnt
gloss	international	law	negligence	in	company	insisted
lex:	dawoliy_1	qAnuwn_1	AisothotAr_1			<imoEn_1
pfx:	Al/DET	b/PREP+Al/DET	Al/DET	-	Al/DET	-
stm:	dwl/NOUN_RELATIVE	qAnwn/NOUN_ABSTRACT	AsthtAr/NOUN	fy/PREP	$rk/NOUN	>mEn/PV
sfx:	(null)/CASE_DEF_GEN	-	-		p/NSUFF_FEM_SG	t/PVSUFF_SUBJ:2FS
gen:, num:	m,s	m,s	m,s	-	f,s	f,s
root	دول	قنن	هتر	none	none	معن
sntmnt	neutral	positive	negative	neutral	neutral	neutral
ptrn	فَعْلِيّ	فَاعُول	اسْتِفْعَال	none	فَعِلَة	أَفْعَل
Sentence		فرجت	حلقاتها	استحكمت	فلما	ضاقت
BW		frjt	HlqAthA	>stHkmt	flmA	DAqt
gloss		opened	rings	completed	when	intensified
lex:		far~aj_1	Haloqap_1	AstHkm_1	lam~A_1	dAq_1
pfx:		-		-	f/SUB_CONJ	-
stm:		frj/PV	Hlq/NOUN_ABSTRACT	AstHkm/PV	lmA/ADV	DAq/PV
sfx:		t/PVSUFF_SUBJ:3FS	At/NSUFF_FEM_PL+ (null)/CASE_DEF_GEN+ hA/POSS_PRON_3FS	t/PVSUFF_SUBJ:2FS	none , none	t/PVSUFF_SUBJ:3FS
gen:, num:		f,s	f,p	f,s	none, none	f,s
root		فرج	حلق	حكم	none	ضيق
sntmnt		positive	neutral	negative	neutral	negative
ptrn		فَعَل	فَعَلَة	اسْتَفْعَل	none	فَعَل
Sentence		العليم	السميع	وهو	الله	فسيكفيكهم
BW		AlElym	AlsmyE	whw	AllAh	fsykfykhm
gloss		all-knowing	all-hearing	he	God	will suffice
lex:					All~‘h$_1$	<imoEAn_1
pfx:		Al/DET	Al/DET	w/CONJ	-	f/CON+s/FUT_PART+ y/IV3MS
stm:		Elym/ADJ_INTENS	smyE/ADJ_INTENS	hw/PRON_3MS	Allh/NOUN_PROP	kfy/IV
sfx:		-	222	-	-	k/IVSUFF_DO:2MS+ hm/IVSUFF_DO:3MP
gen:, num:		m,s	m,s	m,s	m,s	m,s
root		علم	سمع	none	أله	كفي
sntmnt		positive	positive	neutral	positive	positive
ptrn		فَعِيل	فَعِيل	none	عَال	يَفْعِل
Sentence	دقيقةً	دراسةَ	المرصودة	الظاهرة	الباحثون	دَرَسَ
BW	dqyqp	drAsp	AlmrSwdp	AlDAhrp	AlbAHvwn	drs
gloss	closely	studying	observed	phenomenon	researchers	studied
lex:	daqiyq_1	dirAsap_1	maroSuwd_1	ZAhir_1	bAHiv_1	darasa_1
pfx:		-	Al/DET	Al/DET	Al/DET	-
stm:	dqyq/ADJ_QUALIT	drAs/GERUND	mrSwd/NOUN_ PASSIVE_PART	ZAhr/NOUN_ ACTIVE_PART	bAHv/NOUN_ ACTIVE_PART	drs/PV
sfx:	p/NSUFF_FEM_SG	p/NSUFF_FEM_SG	p/NSUFF_FEM_SG	-	p/NSUFF_FEM_SG -	-
gen:, num:	f,s	f,s	f,s	f,s	m,p	m,s
root	دقق	درس	رصد	ظهر	بحث	درس
sntmnt	positive	positive	neutral	neutral	positive	positive
ptrn	فَعِيلَة	فِعَالَة	مَفْعُولَة	فَاعِلَة	فَاعِل	فَعَل
Sentence	الشارع	نبض	مع	يتماهى	الرسميّ	الموقف
BW	Al$ArE	nbD	mE	ytmAhY	Alrsmy~	Almwqf
gloss	public	pulse	with	identify	official	position
lex:	$AriE_1	naboD_1	maE_1	tamahY_1	rasomiy~_1	mawoqif_1
pfx:	Al/DET	-	-	y/IV3MS	Al/DET	Al/DET
stm:	$ArE/NOUN_ CONCRETE	nbD/GERUND	mE/ADV	tmAhY/IV	rsmy/NOUN_ RELATIVE	mwqf/GERUND_ MEEM
sfx:	-	-	-	-	-	-
gen:, num:	m,s	m,s	none,none	m,s	m,s	m,s
root	شرع	نبض	none	مهي	رسم	وقف
sntmnt	neutral	positive	neutral	positive	neutral	neutral
ptrn	فَاعِل	فَعْل	none	يَتَفَاعَل	فُعْلِيّ	مَفْعِل
Sentence		الماخوذ	يجيب	عَخواله	القاروط	وذيت
BW		AlmAxw*	yjyb	ExwAlh	AlqArwT	wd~yt
gloss		the thingy	to get	to his uncles	the kid	i sent
msa		المأخوذ	يَجيء بـ	على أخواله	القاروط	وأَدَيْتُ
lex:		maAxuw*_1	jaAb_1	xaAl_1	qaArwT_1	wd~Y_1
pfx:		AL/DET	y/IV3MS	El Y/PREP	AL/DET	-
stm:		mAxw*/NOUN_ CONCRETE	jyb/IV	AxwAl/NOUN_ CONCRETE	qArwT/NOUN_ CONCRETE	wd~Y/PV
sfx:			-	h/POSS_PRON_3MS	-	t/PVSUFF_SUBJ:1S
gen:, num:		m,s	m,s	m,p	m,s	m,s
root		أخذ	جيء	خول	قرط	أدي
sntmnt		neutral	neutral	positive	negative	neutral
ptrn		مَفْعُول	يَفْعِل	فَعَال	فَاعُول	فَعَل

Table 3: Annotated sentences of JCCA Corpus. In this table, the abbreviation *BW* represents Buckwalter transliteration, *gloss* the English meaning, *lex* the lexical entry, *pfx* the prefix, *stm* the stem, *sfx* the suffix, *gen* the gender, *root* the consonantal roots,*sntmnt* the sentiment designation, and *ptrn* the morphological pattern.

tends to be more liberal and more prone to change, the written variety more coded and more conservative. Arabic has three major varieties, two written and one spoken: Classical Arabic, the language of scholarship until the end of the eighteenth century; Modern Standard Arabic, the language of ed-

Tag	Description	Arabic
DET	Definite Article	أداة تعريف
PREP	Prepositions	حرف جر
CONJ	Conjunction	حرف عطف
INTERROG	Interrogative particles	حرف استفهام
FUT_PART	Particles of futurity	حرف استقبال
PREFIX	Prefix	زيادة في أول الكلمة
CV_PREF	Imperative prefix	حرف أمر
IMPERF_PREF	Imperfect prefix	حرف مضارعة
INF_PART	Infinitive particle	حرف مصدري
INF_SUBJUNC_PART	Infinitive/Subjunctive particle	حرف مصدري ونصب
INF_ANNUL_PART	Infinitive/Annulling particle	حرف مصدري ناسخ
NON_GOVERN	Non-Governing particle	حرف غير عامل
NEG_PART	Negative particle	حرف نفي
OTHER	Non-Governing particle	سابقة أخرى

Table 4: Prefix Tags (Bold is new)

Tag	Description	Arabic
GERUND	Gerund	المصدر
GERUND_MEEM	Gerund with initial miim	المصدر الميمي
GERUND_INSTANT	Gerund of instance	مصدر المرة
GERUND_STATE	Gerund of state	مصدر الهيئة
GERUND_PROFESSION	Gerund profession	مصدر صناعي
NOUN_CONCRETE	Concrete noun	اسم ذات
NOUN_ABSTRACT	Abstract noun	اسم معنى
NOUN_ACTIVE_PART	Active participle	اسم فاعل
NOUN_PASSIVE_PART	Passive participle	اسم مفعول
ADJ_INTENS	Form of exaggeration	صيغة المبالغة
ADJ_QUALIT	Adjective	الصفة المشبهة
NOUN_TIME_PLACE	Noun of time/place	اسم الزمان والمكان
NOUN_INSTRUMENT	Instrumental noun	اسم الآلة
ADJ_COMP	Elative noun	اسم التفضيل
NOUN_RELATIVE	Relative noun	اسم منسوب
NOUN_PROP	Proper noun	اسم علم
NOUN_PROP_FOREIGN	Foreign proper noun	اسم علم أجنبي
ADV	Adverb	الظرف
PRON	Pronoun	الضمير المنفصل
DEM_PRON	Demonstrative pronoun	اسم الإشارة
REL_PRON	Relative pronoun	اسم موصول
INTERROG_PRON	Interrogative pronoun	اسم استفهام
REL_ADV	Conditional noun	اسم شرط
NOUN_VERB_LIKE	Verb-like noun	اسم الفعل
NOUN_FIVE	Five nouns	الأسماء الخمسة
NOUN_DIMINUTIVE	Diminutive	اسم تصغير
NOUN_BLEND	Blend noun	اسم منحوت
NOUN_NUM	Numeral	اسم عدد
EXCEPT_NOUN	Exceptive Noun	اسم استثناء
COMP_NOUN	compound noun	اسم مُرَكَّب
FOREIGN	Foreign word	كلمة أجنبية
ABBREV	Abbreviation	اختصار
PV	Perfect verb	فعل ماض
PV_PASS	Passive Perfect v.	فعل ماض مجهول
IV	Imperfect v.	فعل مضارع
IV_PASS	Passive Imperfect v.	فعل مضارع مجهول
UNINFLECTED_VERB	Uninflected Verb	فعل جامد
CV	Imperative verb	فعل أمر
PREP	Preposition	حرف جر
NEG_PART	Preposition	حرف نفي
CONJ	Conjunction	حرف عطف
INTERROG_PART	Interrogative particle	حرف استفهام
SUBJUNC_PART	Subjunctive particle	حرف نصب
JUSSIVE_PART	Jussive particle	حرف جزم
ANNUL_PART	Annulling particle	حرف ناسخ
VOC_PART	Vocative particle	حرف نداء
EXCEPT_PART	Exceptive par.	حرف استثناء
FUTUR_PART	Par. of futurity	حرف استقبال
YES_NO_RESP_PART	Yes/No particle	حرف جواب
CONDITION_PART	conditional particle	حرف شرط
CERT_PART	Certain/Uncertain particle	حرف تحقيق
PART	other particles	حروف أخرى
PUNC	Punctuation mark	علامة ترقيم
NUMBER	Number	رقم
CURRENCY	Currency	عملة
DATE	Date	تاريخ
NON_ARABIC	Non-Arabic word	كلمة غير عربية
OTHER	OTHER	أخرى

Table 5: Stem Tags (Bold is new)

ucation and formal written communication from the Arab renaissance in the nineteenth century onward; and the dialects, the colloquial regional varieties that are spoken in everyday communication. Since the corpus constructed here is comprehensive and since it claims to be representative of contemporary Arabic, it has to exclude Classical Arabic, but include Modern Standard Arabic, and the regional dialects. We define Contemporary Arabic as the language both written and spoken by living native speakers of Arabic; therefore, the dialects need to be represented. We are not alone in this view, check out *A Frequency Dictionary of Arabic* (Buckwalter and Parkinson, 2011) and the *Oxford Arabic Dictionary* (Arts et al., 2014).

The major spoken varieties are, therefore, represented in the corpus: North Africa is represented by the Moroccan dialect; the Nile region by Egyptian; the Arabian Peninsula by Taizi, Sanaani, and Najdi; Greater Syria by Shami, Jordanian, and Palestinian. The data in the form of contextualized sentences were collected from (1) personal communication in Facebook and Whatsapp family groups; (2) jokes, songs, videoclips, movie scripts, and TV interviews in the local dialects; and (3) personal interviews of old speakers, especially those with minimal education. The data were collected by students who came from these regions. Like any other language, Arabic has differences between the dialects and the standard variety, between the spoken and written varieties. There is variation in the pronunciation of some consonants and vowels (e.g., q, D, Z, v, *, A); suppression of word final inflections; fixed word-order (i.e., subject-verb-object (SVO)); contracted forms (e.g., مظلش *maZal~i$* for ما ظل شيءٌ *mA Zal~a $ay'N* 'nothing remains'); use of high frequency lexical items(e.g., قاعد *qAEid* rather than جالس *jAlis* 'sitting'); use of some lexical items that are archaic in MSA (e.g., افلح *AifliH* 'Partake of food' in Jordanian Arabic in addition to the senses in Standard Arabic of Plough! and Succeed!); liberal incorporation of foreign words (e.g., مسّج *mas~aj* 'sent a message'); abandonment of the dual and the passive voice (e.g., إنكَسَر *<inkasar* 'broke' rather than كُسِرَ *kusira* 'it got broken'); abandonement of the yes-no question

Tag	Type	Arabic	Tag	Type	Arabic
POSS_PRON	Proclitic	ضمير متصل بالاسم	SUBJ_PRON	Suffix	ضمير متصل بالفعل
OBJ_PRON	Proclitic	ضمير متصل بالفعل (مفعول به)	SUFF_FEM_TA	Proclitic	تاء التأنيث
NSUFF_FEM_SG	Proclitic	تاء مربوطة	RELATIVE_YA	Proclitic	ياء النسبة
CASE_INDEF_ACC_GEN	Suffix	التنوين	**SUFF**	Suffix	زيادة في آخر الكلمة
NSUFF_FEM_PL	Proclitic	حروف جمع المؤنث	NSUFF_MASC_PL_NOM	Proclitic	حروف جمع مذكر مرفوع
NSUFF_MASC_PL_ACC	Proclitic	حروف جمع مذكر منصوب	NSUFF_MASC_PL_GEN	Proclitic	حروف جمع مذكر مجرور
NSUFF_MASC_DU_NOM	Proclitic	حروف المثنى مذكر مرفوع	NSUFF_MASC_DU_ACC	Proclitic	حروف المثنى مذكر منصوب
NSUFF_MASC_DU_GEN	Proclitic	حروف مثنى مذكر مجرور	NSUFF_FEM_DU_NOM	Proclitic	حروف المثنى مؤنث مرفوع
NSUFF_FEM_DU_ACC	Proclitic	حروف مثنى مؤنث منصوب	NSUFF_FEM_DU_GEN	Proclitic	حروف مثنى مؤنث مجرور
EMPHATIC_NUN	Suffix	نون التوكيد	**PROTECT_NUN**	Suffix	نون الوقاية
REL_PRON	Relative Pronoun	اسم موصول	**ADV**	Adverb	ظرف
SINGLAR	Number/Singular	مفرد	DUAL	Number/Dual	مثنى
PLURAL	Number/Plural	جمع سالم	**BROKEN_PLR**	Number/Broken plural	جمع تكسير
COLCV_NOUN	Number/Collective noun	اسم الجمع			

Table 6: Tags for suffixes (Bold is new)

particles هل *hal* and أ >; use of the suffix شـ *$* at the end of a verb (e.g., ما قعدش *mA qaEadi$* rather than ما قَعَدَ *mA qaEada*he did not sit); loss of gender distinction, especially in the language of females (e.g., إِجو البنات <*ijw AlbanAt* rather than جاءت البنات *jA'at AlbanAt* 'the girls came'). Arabic has a free word order because of grammatical inflections. When all words' grammatical functions are marked with appropriate inflections, it is not necessary to restrict the arrangement of words in a sentence; hence, Classical Arabic exhibits a totally free word order. Modern Standard Arabic shows preference for verb-subject-object even though inflections are amongst its distinctive features. The spoken varieties continue a historical tradition that we suspect had started as early as Islamic times, where case inflection had lost grounds to fixed word order. Preference in Classical Arabic for the default word order (i.e., verb-subject-object) in an otherwise free word order system was a portent of developments to come. As Islamic conquest brought Arabs in contact with foreigners who soon adopted the language, and as the diglossic gap widened, grammatical inflection lost favor in the low variety while it retained its glamour in the high variety, under the influence of the Quran. The spoken, the low, variety started to favor the subject-verb-object word order as a result of the loss of case inflections and to set apart the agent from the patient of the predicate. The written variety manifested in MSA, on the other hand, used the verb-subject-object order as the unmarked default and retained other combinations for special purposes. All modern regional varieties are descendants of old spoken varieties of Arabic in much the same way as Modern Standard Arabic is a successor of Classical Arabic, the written variety. Regional varieties of Arabic share great many syntactic features. For example, they have two negation patterns: single negation and discontinuous negation (Alqassas, 2015). The first uses the negative particle ما *mA* followed by the verb phrase, whilst the second adds the negative marking suffix شـ *$* to the verb in addition to the negative particle that precedes it. Thus, I didnt say may be expressed as ماقلتِش *mA qult-i$* or ماقُلت *mA qult*. To negate the future, however, there are three options: (1) the negative particle followed by the imperfect verb as in ما أسافر *mA >asAfir* 'I will not travel'; (2) or followed by the imperfect inflicted with the negative marking suffix as in ما أسافِرش *mA >asAfr-i$* ; (3) or followed by the future particle رَح *raH* and the imperfect verb as in ما رح أسافر *mA raH >asAfir*. JCCA consists in part of a spoken language component that is annotated morphologically and syntactically, glossed with MSA forms, and translated into English. This is especially useful with contractions, the hallmarks of spoken Arabic. The gloss is often the non-contracted equivalent in MSA as demonstrated in Table 7.

8 Conclusion and Future Work

This paper outlined the methodology for the design, construction, and annotation of the Jordan Comprehensive Contemporary Arabic Corpus (JCCA). The corpus is balanced, comprehensive, and representative of contemporary Arabic as written and spoken in Arab countries today. It consists of 100 million words that reflect current usage of the language. The corpus consists of 87% written and 13% spoken language. The text of the corpus was selected such that it would be representative of a wide range of geographical regions, genres, subject matters, modes, and media. DI-

Contracted	BW	Full Form	Gloss
شلونك	$lwnk	أي شيء لونك	how are you?
اصطفل	ASTfl	اصطفِ الذي تريد	whatever you want
إيش	<y$	أي شيء	pardon me?
ليش	ly$	لأي شيء	why?
شو	$w	أي شيء هو	what?
بيش	by$	بأي شيء	for how much?
قديش	qdy$	قدر أي شيء	how much?
معلش	mEly$	ما عليك شيء	it's OK!
مظلش	mZl$	ما ظل شيء	nothing left
إللي	<lly	الذي، التي	that/which

Table 7: Contracted words in colloquial Arabic, In this table, the abbreviation *Contracted* represents examples of spoken words (i.e. contractions), *BW* is Buckwalter transliteration, *Full Form* is the non-contracted equivalent in MSA, *gloss* the English meaning.

WAN was upgraded and used to annotate and manually verify the annotation of a one-million-word snapshot of the corpus, making it a gold standard of superior quality that can serve as a resource against which automatic annotation may be compared. JCCA construction made these additional contributions: (i) Development of a new and elaborate tag-set that is based on the morphology of traditional Arabic grammar; (ii) Addition of the roots and morphological patterns of nouns and verbs; (iii) Coverage of the major spoken varieties of Arabic: North Africa; the Nile; the Arabian Peninsula; and Levant. Future work is to make this corpus a monitor corpus where new texts are added proportionally every year. This will facilitate tracking language change and will render the corpus more amiable to lexicography.

9 Acknowledgment

The research reported here was supported by the Scientific Research Fund of the Ministry of Higher Education and Scientific Research, Jordan (Grant No. Soci/2/1/2016).

References

Mourad Abbas and Kamel Smaïli. 2005. Comparison of Topic Identification methods for Arabic Language. In *International Conference on Recent Advances in Natural Language Processing - RANLP 2005*, 14-17, Borovets, Bulgaria.

Mourad Abbas, Kamel Smaili, and Berkani. 2011. Evaluation of topic identification methods on Arabic corpora. *Journal of Digital Information Management*, 9(5):185–192.

Faisal Al-Shargi and Owen Rambow. 2015. DIWAN: A dialectal word annotation tool for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 49–58, Beijing, China. Association for Computational Linguistics.

Latifa Al-Sulaiti and Eric Steven Atwell. 2006. The design of a corpus of Contemporary Arabic. *International Journal of Corpus Linguistics*, 11(2):135–171.

Abdulmohsen Al-Thubaity. 2015. A 700m+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation*, 49(3):721–751.

Sameh Alansary and Magdy Nagi. 2014. The International Corpus of Arabic: Compilation, Analysis and Evaluation. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages pages 8–17, Doha, Qatar.

Ahmad Alqassas. 2015. Negation, tense and npis in jordanian arabic. *Lingua*, 156:101–128.

Maha Sulaiman Alrabia, AbdulMalik Al-Salman, Eric Atwell, and Nawal Alhelewh. 2014. KSUCCA: A Key To Exploring Arabic Historical Linguistics. *International Journal of Computational Linguistics (IJCL)*, 5(2):27–36.

Faisal AlShargi, Aidan Kaplan, Ramy Eskander, Nizar Habash, and Owen Rambow. 2016. Morphologically annotated corpora and morphological analyzers for moroccan and sanaani yemeni arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Tressy Arts, Yonatan Belinkov, Nizar Habash, Adam Kilgarriff, and Vit Suchomel. 2014. artenten: Arabic corpus and word sketches. *Journal of King Saud University - Computer and Information Sciences*, 26(4):357 – 371. Special Issue on Arabic NLP.

Paul Baker, Andrew Hardie, and Tony McEnery. 2006. *A Glossary of Corpus Linguistics*. Edinburgh University Press.

Outi Bat-El. 1994. *Stem modification and cluster transfer in Mmodern Hebrew*. Natural Language Linguistic Theory, 12(4), 571-596.

Outi Bat-El. 2001. *In search for the roots of the C-root: The essence of Semitic morphology*. Workshop on Root and Template Morphology. Los Angeles: University of South California.

Yonatan Belinkov, Nizar Habash, Aadm Kilgarriff, Noam Ordan, Ryan Roth, and Vit Suchomel. 2013. arTenTen12: A new, vast corpus for Arabic. In *Second Workshop on Arabic Corpus Linguistics, WACL'S*, Lancaster University, UK.

Claire Brierley, Majdi Sawalha, and Eric Atwell. 2012. Open-Source Boundary-Annotated Corpus for Arabic Speech and Language Processing. In *LREC*, pages 1011–1016.

Claire Brierley, Majdi Sawalha, Barry Heselwood, and Eric Atwell. 2016. A Verified Arabic-IPA Mapping for Arabic Transcription Technology, Informed by Quranic Recitation, Traditional Arabic Linguistics, and Modern Phonetics. *Journal of Semitic Studies*, 61(1):157–186.

Tim Buckwalter and Dilworth Parkinson. 2011. *A frequency dictionary of Arabic: Core vocabulary for learners.* London: Routledge.

Stuart Davis and Bushra Zawaydeh. 2001. *Arabic hypocoristics and the status of the consonantal root.* Linguistic Inquiry, 32(3): 512-520.

Kais Dukes, Eric Atwell, and Nizar Habash. 2013. Supervised collaboration for syntactic annotation of quranic arabic. *Language Resources and Evaluation*, 47(1):33–62.

Thomas Eckart, Faisal Al-shargi, Uwe Quasthoff, and Dirk Goldhahn. 2014. Large arabic web corpora of high quality: The dimensions time and origin. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools, LREC*, Reykjavk.

Omaima Ismail, Sane Yagi, and Basam Hammo. 2014. Corpus Linguistic Tools for Historical Semantics in Arabic. *International Journal of Arabic-English Studies (IJAES)*, 15:135–152.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: an annotated corpus for the palestinian arabic dialect. *Language Resources and Evaluation*, 51(3):745–775.

McCarthy, John. 1981. *A prosodic theory of nonconcatenative morphology*. Linguistic Inquiry, 12, 373-418.

Tony McEnery and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *LREC*, volume 14, pages 1094–1101.

Jeam-Francois Prunet, Renee Bland, and Ali Idrissi. 2000. *The mental representation of Semitic words.* Linguistic Inquiry, 31(4). 609-648.

Ratcliffe, Robert . 1998. *The broken plural problem in Arabic and comparative Semitic: allomorphy and analogy in non-concatenative morphology.* Amsterdam/Philadelphia: John Benjamins.

Motaz Saad and Wesam Ashour. 2010. Osac: Open source arabic corpora. In *EEECS10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science*, pages 118–123.

Majdi Sawalha, Claire Brierley, and Eric Atwell. 2012. Prosody Prediction for Arabic via the Open-Source Boundary-Annotated Quran Corpus. *Journal of Speech Sciences*, 2(2):175–191.

Majdi Sawalha, Claire Brierley, and Eric Atwell. 2014. Automatically generated, phonemic Arabic-IPA pronunciation tiers for the Boundary Annotated Qur'an Dataset for Machine Learning (version 2.0). In *proceedings of LRE-Rel 2: 2nd Workshop on Language Resource and Evaluation for Religious Texts, LREC 2014 post-conference workshop 31st May 2014, Reykjavik, Iceland*, page 42. LRA.

Ussishkin, Adam. 1999. *The inadequacy of the consonantal root: Modern Hebrew denominal verbs and output-output correspondence.* Phonology, 16(3), 401-442.

Ussishkin, Adam. 2005. *A Fixed prosodic theory of nonconcatenative templatic morphology.* Natural Language Linguistic Theory, 23(1), 169-218.

Wajdi Zaghouani. 2017. Critical Survey of the Freely Available Arabic Corpora. *CoRR*, abs/1702.07835.