A Grammatically Annotated Corpus for Sana’ani Arabic Dialect

[1][2]Sabah Al-Shehabi, [3][4]Mohammed Sharaf-Addin
[1]CALTS, University of Hyderabad, Hyderabad, India
[2]Department of English, Faculty of Education, Mahweet, Sana’a University, Sana’a, Yemen.
[3]CAS in Linguistics, Osmania University, Hyderabad, India
[4] Department of English, Faculty of Arts, Thamar University, Thamar, Yemen
[1][2] sabahmohammed986@gmail.com, [3][4]ma.alshami22@gmail.com

Article Info
Volume 83
Page Number: 4953 - 4961
Publication Issue:
March - April 2020

Article History
***Article Received**: 24 July 2019*
***Revised**: 12 September 2019*
***Accepted**: 15 February 2020*
***Publication**: 27 March 2020*

Abstract

In this paper, we introduce a new resource for Sana’ani Arabic dialect. This grammatically tagged corpus is basically a collection of social media texts that is primarily developed as a training data for developing Sana’ani Arabic Part Of Speech (POS) tagger. The corpus consists of 7,295 tokenized sentences with an average of 15 tokens in each sentence and with a total number of 112,517 tokens and 15,940 types. The corpus is manually annotated using a modified tagset from The Biestagset which covers 24 tags. The manual annotation performed is rather a grammatical annotation ignoring morphological inflections and concentrating on the syntactic features using the context to identify the part of speech of each token.

Index Terms; *Corpus Annotation, Dialectal Arabic,Parts of Speech, Sana’ani Arabic, Tagset*

I. INTRODUCTION

Arabic language is one of the most spoken languages of the world. One of the markers of Arabic language is the diglossic nature of the language [1] where two varieties (Modern Standard Arabic (MSA) and Dialectal Arabic (DA) exists side-by-side and are closely related. MSA is a predominant variety over dialectal Arabic in formal settings which restrict almost all the written content to the standard variety. However, recently and with the advent of technology and the vast spread of social media networking sites, a strong presence of DA is noticed and more individual-driven data becomes accessible and available as users of these sites feel free and encouraged to jot down their thoughts, interact or comment about their daily social life in their own dialects. The challenge, however, remains in obtaining such dialectal datasets which can be viable, and usable by machines. This challenge is tested when it comes to building Natural language Processing (NLP) tools and applications. Therefore, obtaining a clean, preprocessed, valid and machine readable text is a crucial necessity for developing any NLP applications. Online data can be collected from the networking sites either manually or automatically using tools for crawling and compiling. This collection of texts, after being cleaned and preprocessed, which is now called a raw corpus can be considered a standard reference for the language variety which it is supposed to represent. This type of corpus can be used for developing many NLP tools and applications. However, machines are still not smart enough to disambiguate similar contents unless being provided with some added values to the texts. This process is called corpus annotation which [2] defines as the process of ‘adding such interpretative, linguistic information to an electronic corpus of spoken and/or written language data’. The advantages of such annotated corpus is suggested by both Leech and McEnery and cited in [3] which include: (1) it is much easier to extract information; (2) reusable resource; (3) a multi-functional resource; and (4) it records a linguistic analysis explicitly.

The main objective of this study is to present a new annotated corpus for Sana'ani Dialect with linguistic information that can be used for developing further NLP applications. This paper is structured with the following headings: □. INTRODUCTION, □□.RELATED WORK, □□□.SANA'ANI ARABIC, □V. CORPUS DESCRIPTION, V. CORPUS ANNOTATION, and V□.CONCLUSION AND FUTURE WORK.

II. RERLATED WORK

The literature directed to Arabic Dialects is increasing on each successive day over a long period, after the bulk of major works on Arabic language was centered to MSA. However, researches on Arabic dialects are still lagging far behind that of MSA either in terms of data availability, coverage or validity for machine use. This may be due of the paucity of data readily available for researchers as MSA is still predominant over dialectal Arabic in formal settings. However, with the advent of technology and the vast spread of social media networking sites, more individual-driven data becomes accessible and available.

A recent critical survey of the freely available Arabic Corpora was conducted by [4] where he listed about 66 free resources of Arabic Corpora. All these corpora exist in the form of 6 categories: i.e., 23 Raw Text Corpora (i.e., 11 Monolingual Corpora List; 4 Multilingual Corpora List; 2 Dialectal Corpora; and 6 Web-based Corpora List); 15 Annotated Corpora (i.e., 6 Named Entities Corpora List; 3 Errors Annotated Corpora List; and 6 Miscellaneous Annotated Corpora List); 16 Lexicon Corpora (i.e., 9 Lexical Databases List and 7 List of Words Lists); 1 Speech Corpora; 4 Handwriting Recognition Corpora and 7 Miscellaneous Corpora types (e.g., Questions/Answers, comparable corpora, plagiarism detection and summaries). As it is noted among this collection of texts, the focus can be summarized in terms of quantity, quality, coverage, and accessibility which are the criteria or the main motives which Arabic researchers opt for better resources. Out of this collection of texts, this survey mentioned only two dialectal corpora which exist in the form raw text resources (i.e., Tunisian Dialect Corpus [5] and Arabic Multi Dialect Text Corpora [6]). The Tunisian Dialect Corpus consists of 3,403 words which have been transcribed from spoken dialogues between staffs and clients. While the Arabic Multi Dialect Text Corpora has a huge volume of about 2 million unique words gathered from 55K webpages obtained from main Arabic regional dialectal varieties (i.e., Gulf, Levantine, North Africa, Egypt).

A number of other studies have been conducted on Arabic dialects. Most of them focus on preparing dialectal corpora for machine learning use and training as well as for developing dialect-based NLP applications. These corpora either evolved as (1) raw texts dialectal corpora [7][8]; (2) annotated dialectal corpora [4],[6],[9-11][15]; or (3) Parallel dialectal corpora [12][3].

The studies that focus on raw corpora include [8] who developed a monolingual social media based text corpus for Sana'ani Yemeni dialect, one of the most popular spoken dialects of Yemen. Their corpus size is 447,401 tokens and 51,073 types extracted from Facebook and Telegram Apps that represent daily fictional conversations written during the years 2017 and 2018. While [7] is a balanced multi-Arabic dialectal text corpus built by using CMC and social media sources: Twitter, comments from online newspapers, and Facebook. Their corpus size is 13,876,504 word tokens collected from five groups of Arabic dialects: Gulf, Iraqi, Egyptian, Levantine, and North African.

A number of other researches were conducted on Arabic annotated corpora (category 2) for the purpose of creating standard reference resources that provide a stable base of linguistic analyses. These studies include [9][14][11] and [15] focused on morphological annotation. [9] presented new resources for two Arabic dialects: namely Moroccan and Sana'ani Yemeni Arabic. The corpus for each dialect was morphologically annotated using the DIWAN tool [16] which requires manual annotation. Their corpus size is 64K and 32.5K tokens for Morrocan and Sana'ani Yemeni Arabic respectively. While [14] developed a corpus for Palestinian Arabic dialect called Curras. This corpus consists of 56,700 tokens and 16,416 types. They annotated about 98.7 % tokens and (97.6 %) types which were valid. Each token was annotated morphologically with part-of-speech (POS), stem, prefix, suffix, lemma, and gloss. They collected their corpus from Facebook, Twitter, Forums, Palestinian stories, Palestinian terms, and TV Shows. [11] introduced another annotated large-scale resource for Emirati Arabic with a manual morphological annotation including tokenization, part-of-speech, lemmatization, English glosses and dialect identification. This corpus covers 200K words chosen from eight Gumar corpus novels of Emirati Arabic. [15] presented a collection of morphologically annotated corpora for seven Arabic dialects: Taizi Yemeni, Sanaani Yemeni, Najdi, Jordanian, Syrian, Iraqi and Moroccan Arabic. Their corpora collections cover 200,000 words provided with orthography, diacritized lemmas, tokenization, morphological units and English glosses. The other type of dialectal corpora, on the other hand, used different annotation [10]. They presented a multi-dialectal corpus that covers 11 distinctive Arabic regional dialectal varieties spoken in 16 Arabic countries that was extracted from Twitter platforms and they called it 'Arap-Tweet'. However, later on they developed an improved version (version 2.0) with various improvements in terms of volume and quality of annotation [17]. The annotation adopted in these corpora was based on three criteria: Dialect, Age and Gender.

The third corpora collections concentrated more on parallel dialectal corpora [12-13]. [12] presented two resources: the MADAR Corpus (a parallel corpus) and MADAR Lexicon. In MADAR Corpus they translated some selected sentences from the Basic Traveling Expression Corpus (BTEC) [18] into Arabic multi-dialects covering about 25 cities; whereas in MADAR Lexicon, they cover about 1,045 entries taken from the same cities. [13] on the other hand, presented a comprehensive 3-way large-scale parallel lexicon of English, MSA and Egyptian Arabic with deep linguistic annotation that includes part of speech (POS), number, gender, rationality, and morphological root and pattern forms. This lexicon consists of about 73,000 Egyptian entries.

As our focus is on Sana'ani Yemeni Arabic, the only reported work on this dialect is done by [9] [8] and [15]. The first annotated corpus for Sana'ani dialect was attempted by [9] where a collection of 32.5K tokens was obtained from both online and print materials. They covered as much genres as they could. This include Oral interviews, Social texts, Wisdoms and tales, Sana'ani folktales, Sermons, Poems, Humor, Explanation and Politic text. They used the DIWAN tool which assigns the following annotations for each word in the corpus: Diac, Lex, Bwhash, Gloss, Clitics, Other features (part of speech, gender, functional gender, formal number, and functional number.) The other study seems to be similar to [9] conducted by the same authors and using the same corpus size and the same tool [15]. However, this study includes two Yemeni dialects, Sana'ani and Taizi along with other 5 Arabic dialects. Each word in the corpus was annotated with CODA, Lemma, Morph, Prefix, Stem and Suffix to bridge a common ground with MSA and other Arabic dialects. A more recent study on Sana'ani corpus was conducted by [8]. They developed a mono-dialectal social media based text corpus for Sana'ani Yemeni dialect. Their corpus size is 447,401 tokens and 51,073 types extracted from Facebook and Telegram Apps that represent daily fictional conversations written during the years

2017 and 2018. Since this corpus is recent with a high volume and plain text, we build our study based on the data presented in their corpus. We selected about 112,517 tokens and manually annotated them with our adopted POS tags.

III. SANA’ANI ARABIC

Sana’ani Arabic is one of the three main dialects spoken in Yemen. It belongs to the Yemeni dialects which are spoken in South of Arabian Peninsula namely, Yemen and south of Kingdom of Saudi Arabia. It is mainly spoken by 30 per cent of the whole population of Yemen which would approximate 9 million speakers [8]. Sana’ani is considered as a spoken informal variety where Modern standard Arabic (MSA) is the formal written form for all Arabic speakers. These two forms are used in complementary distribution which is known as diglossia. Though Sana’ani Arabic has common linguistic features with Classical Arabic and MSA, it shows a linguistic peculiarity of its own. In the following section we will show some of the disguising linguistic features of Sana’ani Arabic in comparison to MSA.

A. Linguistic Details

Sana’ani Arabic Phonology was described in details by Watson [19]. Phonologically Sana’ani speakers show a unique pronunciation of some MSA consonants such as the voiceless uvular plosive /q/ which is replaced by a voiced velar plosive /g/. For examples the word /qa:la/ ‘he said’ in MSA is pronounced in Sana’ani Arabic as /ga:la/. Another distinguishing feature of Sana’ani Arabic is replacement of the voiced dental-alveolar plosive /d/ in word medial position with an emphatic voiceless dental-alveolar plosive /t̪/ e.g., /s̪adr/ ‘chest’ in MSA is pronounced as /s̪at̪r/. In addition, word initial or intervocalic voiceless dental-alveolar plosive /t̪/ is pronounced as the voiced dental-alveolar plosive /d/ e.g., /t̪ajja:rah/ ‘airplane’ is pronounced as /daja:rah/. Such phonological treats of Sana’ani Arabic can influence the dialect orthography.

From a morphological point of view, Sana’ani Arabic has several distinguishing features. For instance, the dual marker of MSA is mostly nonfunctional in Sana’ani Arabic. Instead, the plural marker is being used. e.g., in MSA the second person dual pronoun is /ʔantuma:/ ‘you dual’ is realized in Sana’ani Arabic as /ʔntum/ ‘you plural’. Another distinguishing morphological feature of Sana’ani is the imperfective tense marker which includes continuous/habitual and future markers, is totally different from MSA as presented in Table □.

Table I. It shows imperfective tense marker

	Sana’ani Arabic		MSA	
person	Continuous/ habitual	future	Continuous/habitual	future
1st	biyt- biyn-	ʃa- ʕd-	ʔa- na-	sa- sawfa
2nd&3rd	b-	ʕa-	ja- ta-	

Syntactically Sana’ani Arabic acts freely and deviates from many of Classical Arabic and MSA syntactic rules. For example the Sana’ani Arabic shows a freer word order in general. For instance, adjectives in MSA are to come after nouns and not to precede them but in Sana’ani Arabic it can come prior to a noun for the purpose of emphasis. e.g., /t̪ajja:rahkabi:rah/ ‘a large airplane’ it can also come as /kabi:raht̪ajja:rah/ ‘airplane large’. Besides, nouns and adjectives can be separated by the indefinite demonstrative /hakaða/ ‘like this’ e.g., /t̪ajja:rahhakaðakabi:rah/ literally ‘airplane like this large’ which means ‘a large airplane like this’.

IV. CORPUS DESCRIPTION

The corpus used for this paper is taken from “social media” raw corpus developed by [8] which is a collection of fictional dialogues that representing different settings and topics of Sana‘ani dialectal data during the years 2017 and 2018. Out of 447,401 tokens and 51,073 types, we manually annotated about 112,517 tokens and 15,940 types with 24 distinguished POS tagset. The main aim of developing such tagged corpus is to use it for training POS tagger for Sana’ani Dialect.

V. CORPUS ANNOTATION

Since our corpus is a social media corpus adopted from an earlier research presented by [8], a number of decisions are made prior to the grammatical annotation. Firstly, we have to take into account the type of tagset used for the annotation so we decided to use a coarse tagset which ignores any inflectional features related to the text morphology. Secondly, our annotation adheres to the following maxims of corpus annotation by [20]:

(1) It should always be easy to dispense with annotations, and revert to the raw corpus. The raw corpus should be recoverable.

(2) The annotations should, correspondingly, be extractable from the raw corpus, to be stored independently, or stored in an interlinear format.

(3) The scheme of analysis presupposed by the annotations—the annotation scheme—should be based on principles or guidelines accessible to the end-user. (The annotation scheme consists of the set of annotative symbols used, their definitions, and the rules and guidelines for their application.)

(4) It should also be made clear how, and by whom, the annotations were applied.

(5) There can be no claim that the annotation scheme represents 'God's truth'. Rather, the annotated corpus is made available to a research community on a caveat emptor principle. It is offered as a matter of convenience only, on the assumption that many users will find it useful to use a corpus with annotations already built in, rather than to devise and apply their own annotation schemes from scratch (a task which could take them years to accomplish).

(6) Therefore, to avoid misapplication, annotation schemes should preferably be based as far as possible on 'consensual', theory-neutral analyses of the data.

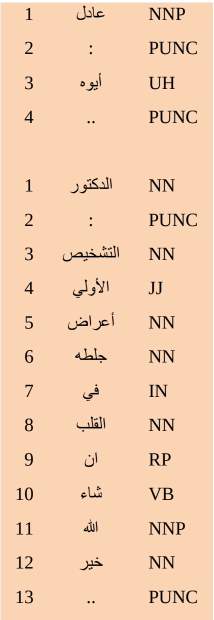
(7) No one annotation scheme can claim authority as a standard, although de facto interchange 'standards' may arise, through widening availability of annotated corpora, and perhaps should be encouraged. [4]

Thirdly, the orthographical variations are dealt with using the normalization software tool developed by Sharaf-Addin [22] for the purpose of Sana'ani text normalization. Finally, our annotation is guided by the PATB annotation guidelines that are described by [23]. In this section we describe the annotation process performed including the used tagset, sentence tokenization and annotation statistics.

A. Annotation Process

The annotation is mainly performed manually. First, the corpus is preprocessed using Sana'ani dialect normalizer [22] the text is then manually revised to check for any possible orthographic variation that skips normalization. Second, the corpus is manually tokenized into sentences as described in part 5.3. Third, annotation format is designed in columns where the first column presents the word no, the second shows the token while the third presents the POS tag of the token. Each token is presented in a row and the sentences are separated by an empty row. Table □□.is an example of annotation format.

Table II. It shows an example of POS tags annotation format.



1	عادل	NNP
2	:	PUNC
3	أيوه	UH
4	..	PUNC
1	الدكتور	NN
2	:	PUNC
3	التشخيص	NN
4	الأولي	JJ
5	أعراض	NN
6	جلطه	NN
7	في	IN
8	القلب	NN
9	ان	RP
10	شاء	VB
11	الله	NNP
12	خير	NN
13	..	PUNC

B. Tagset

The tagset used is a modified tagset from The Biestagset which was developed by Ann Bies and Dan Bikel as a reduced form of the Buckwaltertagset that is used in the Penn Arabic Treebank (PATB) [24]. It is also known as the reduced tagset (RTS) consisting of 24 tags. We chose to use a linguistically coarse tagset to only account for the syntactic features rather than the morphological ones. The modifications that are applied to the RTS tagset are meant for refining the tags and making them suitable for Sana'ani Arabic as well as the purpose of this annotation which is preparing a training corpus for performing Parts of Speech Tagging task. The alterations in the Biestagset are shown in Table □□□. In addition, combined tags are also used wherever needed depending on the tokens and the context.

Table III. It shows both Bies as well as adopted tagset.

BiesTagset		**Adopted Tagset**	
NOMINALS			
Nouns			
NN	singular common noun or abbreviation	NN	common noun or abbreviation
NNS	plural/dual common noun		
NNP	singular proper noun	NNP	proper noun
NNPS	plural/dual proper noun		
Pronouns			
PRP	personal pronoun	PRP	Personal & possessive pronoun
PRP$	possessive personal pronoun		
WP	relative pronoun	WP	relative pronoun
		D_PRP	demonstrative pronoun
Other		JJ	adjective
JJ	adjective		
RB	adverb	RB	adverb
WRB	relative adverb	WRB	relative adverb
CD	cardinal number	CD	cardinal number
		FCD	foreign cardinal number
		OD	Ordinal number
FW	foreign word	FW	foreign word
PARTICLES		CC	coordinating conjunction
CC	coordinating conjunction		
		SC	subordinating conjunction
DT	determiner/demonstrative pronoun	DT	determiner
RP	particle	RP	particle
		INTG_RP	Interrogative particle
IN	preposition or subordinating conjunction	IN	preposition
VERBS		AUX_VB	Auxiliary verb
VBP	active imperfect verb		
VBN	passive imperfect/perfect verb		
VBD	active perfect verb	VB	Main verbs
VB	imperative verb		
OTHER		UH	interjection
UH	interjection		
PUNC	punctuation	PUNC	punctuation
NUMERIC_COMMA	the letter ر r used as a comma	SYM	
NO_FUNC	unanalyzed word	NO_FUNC	unanalyzed word

C. Sentence Tokenization

One of the major issues that faced us is identifying the sentence boundaries. Since punctuation marks are generally unreliable in written Arabic text many Arabic researchers avoid using them for analysis instead they prefer using clauses as units of analysis [21]. Hence, for the purpose of grammatical annotation we had to do the sentence tokenization manually following the criterion suggested by [21]. The criterion is a syntactic-semantic criterion which suggests that the sentence has to be structurally independent and expresses a complete thought. Based on this, our corpus was tokenized into 7,295 sentences. Each sentence is annotated in a linear format where sentences are separated from each other with a blank row.

D. Annotation Statistics

The current corpus calculated to 112,517 token out of which 15,940 is the number of types. Fig.□ shows the ratio of tokens to types. The corpus was annotated fully and the frequency of each tag is shown in Table □V. The number of nominals is the highest as it approximates 47,391 out of which nouns (NN, & NNP) equal 36,397, pronouns (PRP, WP & D_PRP) equal 4,271 and others (JJ,RB,WRB, CD,FCD, OD, & FW) equal 6,723. Then verbs (VB & AUX_VB) come with the second highest frequency which is calculated as 27,988. Particles

(CC, SC, DT, RP, INTG_RP & IN), on the other hand, are calculated to 18,311. The rest of the tags which are known as others (UH, PUNC & SYM) are calculated as 18,827.

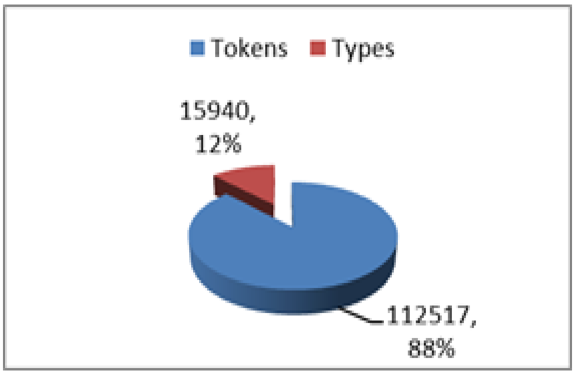


Fig. I It shows token-type ratio

Table IV. It shows the frequency of tags

Tag	Frequency
VB	27,253
NN	24,400
PUNC	16,966
NNP	11,997
IN	9,075
RP	5,234
JJ	3,533
PRP	2,809
RB	2,013
INTG_RP	1,977
UH	1,830
CC	969
D_PRP	969
SC	902
AUX_VB	735
WP	493
OD	309
FCD	304
FW	272
CD	173
DT	154
WRB	119
SYM	31

VI. CONCLUSION AND FUTURE WORK

In this paper we present a grammatically annotated corpus for Sana'ani Arabic with distinctive 24 POS tags. This corpus covers more than 110,000 tokens. The corpus was tokenized into 7,295 sentences and annotated manually using a modified coarse tagset. The annotation performed is rather a grammatical annotation ignoring morphological inflections and concentrating on the context to identify the part of speech of each token. The corpus is normalized and tokenized and then annotated following a set of established rules and schemes. As the purpose of this corpus is to make a training corpus, in future work we are planning to perform a supervised Part Of Speech Tagging using machine learning algorithms. We also plan to extend the size of the annotated corpus to cover as much as needed for the training task.

REFERENCES

[1]. Habash, Nizar. "Arabic morphological representations for machine translation." In Arabic computational morphology, pp. 263-285. Springer, Dordrecht, 2007.

[2]. Garside, Roger, Geoffrey N. Leech, and Tony McEnery, eds. Corpus annotation: linguistic information from computer text corpora. Taylor & Francis, 1997.

[3]. McEnery, Tony, Richard Xiao, and Yukio Tono. Corpus-based language studies: An advanced resource book. Taylor & Francis, 2006.

[4]. Zaghouani, Wajdi. "Critical survey of the freely available Arabic corpora." arXiv preprint arXiv:1702.07835 (2017).

[5]. Graja, Marwa, Maher Jaoua, and L. HadrichBelguith. "Lexical study of a spoken dialogue corpus in tunisian dialect." In The international arab conference on information technology (acit), benghazi–libya. 2010.

[6]. Almeman, Khalid, and Mark Lee. "Automatic building of arabic multi dialect text corpora by bootstrapping dialect words." In 2013 1st

International Conference on Communications, Signal Processing, and their Applications (ICCSPA), pp. 1-6. IEEE, 2013.

[7]. Alshutayri, Areej& Atwell, Eric. A Social Media Corpus of Arabic Dialect Text. (2018).

[8]. SharafAddin M., Al-Shehabi S. (2020) Developing Social-Media Based Text Corpus for San’ani Dialect (SMTCSD). In: Satapathy S., Raju K., Shyamala K., Krishna D., Favorskaya M. (eds) Advances in Decision Sciences, Image Processing, Security and Computer Vision. ICETE 2019. Learning and Analytics in Intelligent Systems, vol 3. Springer, Cham

[9]. Al-Shargi, Faisal, Aidan Kaplan, RamyEskander, NizarHabash, and Owen Rambow. "Morphologically annotated corpora and morphological analyzers for Moroccan and Sanaani Yemeni Arabic." 2016.

[10]. Zaghouani, Wajdi, and AnisCharfi. "Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification." arXiv preprint arXiv:1808.07674 (2018).

[11]. Khalifa, Salam, NizarHabash, FadhlEryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. "A morphologically annotated corpus of Emirati Arabic." In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.

[12]. Bouamor, Houda, NizarHabash, Mohammad Salameh, WajdiZaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid et al. "The MADAR Arabic dialect corpus and lexicon." In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018.

[13]. Diab, Mona T., Mohamed Al-Badrashiny, Maryam Aminian, Mohammed Attia, HebaElfardy, NizarHabash, AbdelatiHawwari, WaelSalloum, PradeepDasigi, and RamyEskander. "Tharwa: A Large Scale Dialectal Arabic-Standard Arabic-English Lexicon." In LREC, pp. 3782-3789. 2014.

[14]. Jarrar, Mustafa, NizarHabash, FaeqAlrimawi, DiyamAkra, and Nasser Zalmout. "Curras: an annotated corpus for the Palestinian Arabic dialect." Language Resources and Evaluation 51, no. 3 (2017): 745-775.

[15]. Alshargi, Faisal, ShahdDibas, SakharAlkhereyf, ReemFaraj, BasmahAbdulkareem, Sane Yagi, OuafaaKacha, NizarHabash, and Owen 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, pp. 137-147. 2019.

[16]. Al-Shargi, Faisal, and Owen Rambow. "DIWAN: A dialectal word annotation tool for Arabic." In Proceedings of the Second Workshop on Arabic Natural Language Processing, pp. 49-58. 2015.

[17]. Charfi, Anis, WajdiZaghouani, Syed Hassan Mehdi, and Esraa Mohamed. "A Fine-Grained Annotated Multi-Dialectal Arabic Corpus." In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp. 198-204. 2019.

[18]. Takezawa, Toshiyuki, GenichiroKikui, Masahide Mizushima, and EiichiroSumita. "Multilingual spoken language corpus development for communication research." In International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006, pp. 303-324. 2007.

[19]. Watson J. San'ani Arabic. In: Versteegh K., Eid M., Elgibali A., Woidich M. and Zaborski A.(eds) Encyclopaedia of Arabic Language and Linguistics vol 4. Brill. (2009)

[20]. Leech, Geoffrey. "Corpus annotation schemes." Literary and linguistic computing 8, no. 4 (1993): 275-281.

[21]. Alkohlani, Fatima A. "THE ARABIC

SENTENCE: TOWARDS A CLEAR VIEW." E-Journal of Arabic Studies & Islamic Civilization Volume 2 – 2015

[22]. SharafAddin, Mohammed. "Developing a Normalizer for San'ani Arabic Social Media Texts." International Journal of Interdisciplinary Research and Innovations 7, no. 2 (2019). http://www.researchpublish.com/journal/IJIRI/Issue-2-April-2019-June-2019/15

[23]. Maamouri, Mohamed, Ann Bies, and Seth Kulick. "Enhancing the Arabic Treebank: a Collaborative Effort toward New Annotation Guidelines." In LREC, pp. 3-192. 2008.

[24]. Habash, Nizar Y. "Introduction to Arabic natural language processing." Synthesis Lectures on Human Language Technologies 3, no. 1 (2010): 1-187