BAAC: Bangor Arabic Annotated Corpus

Ibrahim S Alkhazi
College of Computers & Information Technology
Tabuk University, Tabuk, Saudi Arabia

William J. Teahan
School of Computer Science Bangor University
United Kingdom

***Abstract*—This paper describes the creation of the new Bangor Arabic Annotated Corpus (BAAC) which is a Modern Standard Arabic (MSA) corpus that comprises 50K words manually annotated by parts-of-speech. For evaluating the quality of the corpus, the Kappa coefficient and a direct percent agreement for each tag were calculated for the new corpus and a Kappa value of 0.956 was obtained, with an average observed agreement of 94.25%. The corpus was used to evaluate the widely used Madamira Arabic part-of-speech tagger and to further investigate compression models for text compressed using part-of-speech tags. Also, a new annotation tool was developed and employed for the annotation process of BAAC.**

Keywords—Component; arabic language; corpus; annotated corpora; analysis results

I. Background and Motivation

The Arabic language "العربية" is acknowledged to be one of the most largely used languages, with 330 million people using the language as their first language, as shown in Table 1, plus 1.4 billion more using it as a secondary language [1]. The majority of the speakers are located across twenty-two nations, primarily in the Middle East, North Africa and Asia, and the United Nations considers the Arabic language as one of its five official languages. The Arabic language is part of the Semitic languages that includes Tigrinya, Amharic, Hebrew, etc., and shares almost the same structure as those languages. It has 28 letters, two genders – feminine and masculine, as well as singular, dual and plural forms. The Arabic language has a right-to-left writing system with the basic grammatical structure that consists of verb-subject-object and other structures, such as VOS, VO and SVO [2]–[4].

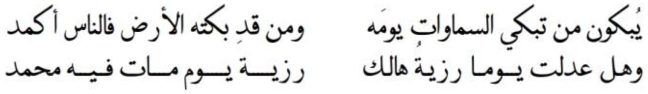
TABLE I. The Most Universally Used Languages

Rank	Language	Users (millions)
1	Mandarin	1051
2	English	508
3	Hindi	497
4	Spanish	392
5	Arabic	330
6	Russian	277
7	Bengali	211
8	Portuguese	191
9	Malay	159
10	French	129

The non-colloquial written text for the Arabic language can be divided into two types: Classical Arabic and Modern Standard Arabic [5]–[8]. The Classical Arabic (CA) epoch, as shown in Figure 1, is usually measured from the sixth century which is the start of Arabic literature. It is the language of the Holy Quran, the 1,400-year-old primary religious book of Islam with 77,430 words [9] and other ancient Islamic books from that era, such as the Hadith books [10]. With the beginning of journalism and the spread of literacy in the eighteenth century came Modern Standard Arabic or MSA. MSA is the language of current printed Arabic media and most Arabic publications.

Most Arabic natural language processing (NLP) tasks perform better for MSA [11]. One example of those tasks is parts-of-speech tagging (POS) of the Arabic language as reported in [10], [12], [13], where the performance of the taggers is best when tagging MSA text. The reason for the variation in performance between MSA and CA is that most Arabic language NLP systems were trained using MSA text [14], [15]. More effort is currently being made, such as the creation of manually annotated CA corpora [16] and the evaluation of different Arabic POS taggers on CA text by Alosaimy and Atwell [12], to fill this gap in research.

The term corpus can be defined as a computerised set of genuine texts or discourses provided by language speakers and saved in a machine-readable form [17]–[20]. Xiao [21] argues that a corpus is not a randomly collected collection of texts nor an archive, but a file that manifests four essential aspects: a corpus is a set of (1) machine-readable (2) genuine texts (that includes transcripts of spoken data) that are (3) tested to be (4) representative of a specific or a group of languages.



يُبكون من تبكي السماواتِ يومَه　　ومن قدِ بكته الأرض فالناس أكمد
وهل عدلت يـوما رزية هالك　　رزيـــة يـــوم مـــات فيـــه محمد

Fig. 1. A Classical Arabic Poem.

Corpora play a significant factor in the development, improvement and evaluation of many NLP applications such as machine translation [22], [23], part-of-speech tagging [24] and text-classification [14], [23]. The design of any corpus depends on its intended applications [25]. Some corpora are for general use and can be utilised in many applications, and others may serve a specific purpose, such as building dictionaries or examining the language of a specific author or duration of time [10].

There are several kinds of annotations which could be applied to corpora, and each annotation is usually designed to handle a certain aspect of the language [26]. One type of corpora annotation is the structural annotation of the corpus by attaching descriptive information about the text, like mark-ups that specify the boundaries of the sentence, section and chapter, or a header file that names the author of the text or adds information about participants, such as the age and gender. Another type of annotation is the morphological annotation, where information about the text, like the stems or root based in a language like Arabic, is added to the corpora. This research applies the most common type of corpora annotation, which is POS tagging of the text [26], where a tag, such as a noun, verb or particle is combined with each term in the corpus, and the number of tags used in the annotation varies from a few to 400 tags or more [27].

Based on the type of text and creation purposes, the corpus can be categorised into six categories: Raw Text Corpora, Annotated Corpora, Lexicon Corpora, Annotated Corpora and Miscellaneous Corpora. Examples of corpora for the Arabic language are provided below.

1) Raw Text Corpora can be Divided into:

A. ***Monolingual corpora***, such as the BACC [28], Ajdir Corpora [29], the King Saud University corpus of Classical Arabic [30], Alwatan [31], Tashkeela [32] and the Al Khaleej Corpus [33]. The monolingual corpora consist of a raw text written in a single language.

B. ***Multilingual corpora***, also known as comparable corpora or parallel corpora, are corpora that are written in two or more languages. Multilingual corpora, such as the UN corpus [34] which is the most important and widely known free corpus [23], Corpus A [22], the Hadith Standard Corpus [35], [36] and MEEDAN Translation Memory [37], are widely used in NLP fields such as machine translation [22], [23].

C. ***Dialectal Corpora***, where the corpus is written in a specific language dialect, such as the Bangor Twitter Arabic Corpus for the Egyptian, Gulf, Iraqi, Maghrebi and Levantine Arabic dialects [38]. Such corpora are used in fields such as text-classification [14].

D. ***Web-based corpora***, such as the KACST Arabic Corpus [39], the Leeds Arabic Internet Corpus [40] and the International Corpus of Arabic [41], where the corpora are only accessible online by an inquiry interface and the corpora cannot be downloaded.

2) The second type is Lexicon corpora, that can be divided into:

A. ***Lexical Databases***, such as the BAMA 1.0 English-Arabic Lexicon [42] and the Arabic-English Learner's Dictionary [43].

B. ***Words Lists*** such as the Word Count of Modern Standard Arabic [43] and the Arabic Wordlist for Spellchecking [44], [45].

These types of corpora act like a vocabulary or a list of words and can be employed by linguists to study many aspects of a language or combined with the lexicons of systems, like spell checking applications, to improve their performance [23].

*3) **Miscellaneous Corpora**, such as Speech Corpora [46], Handwriting Recognition Corpora [47], are beneficial for a number of NLP correlated tasks such as plagiarism detection [48], speech recognition systems [46] and question answering [49].*

*4) **Annotated corpora** are essential for the development of many NLP systems, such as part-of-speech tagging [24], text parsing [50]. Annotated corpora are divided into:*

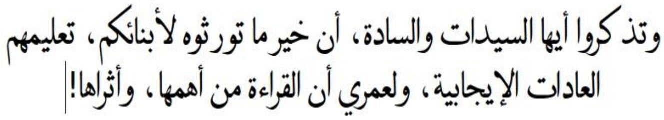
A. ***Named Entities Corpora*** such as JRC-Names [51] and ANERCorp [52]. Most corpora of this type include the names of persons with the company or organisation name and the locations.

B. ***Error-Annotated Corpora***, such as the KACST Error corpus [53], is a beneficial resource for systems such as spelling correction and machine translation corrected output [54].

C. ***Miscellaneous Annotated Corpora***, such as the OntoNotes corpus [55] and the Arabic Wikipedia Dependency Corpus [56] which are semantically annotated corpora [55].

D. ***Part-of-Speech (POS)*** tagged corpora are an important resource for the training and development of POS systems [24]. Some of the existent resources will be presented in detail in the existing resources section below.

POS annotated corpora are essential for the development of many NLP systems, such as part-of-speech tagging [24], statistical modelling [57] and tag-based compression which provides more effective compression for Arabic text than word or character-based compression methods [13]. The lack of such resources limits some researchers from progressing further in their efforts. The limited availability of some existing annotated corpora and the cost of acquiring others are one of the main reasons that contribute to resource scarcity. Several efforts have been made to overcome the lack of resources [12], [16], [20].



وتذكروا أيها السيدات والسادة، أن خيرما تورثوه لأبنائكم، تعليمهم
العادات الإيجابية، ولعمري أن القراءة من أهمها، وأثراها!

Fig. 2. A Social News from Press Sb-corpus [28] in MSA text.

There exist some annotated corpora for the Arabic language that cannot be utilised by many researchers, such as the tag-based text compression research applied by Alkhazi, Alghamdi and Teahan [13] due to availability, and cost issues, such as the Arabic Treebank corpus [58]. Other resources are designed to be used for particular research or annotated using a distinctive tagset produced for an explicit purpose. The Qur’anic Arabic Dependency Treebank is one example where the text is written in CA text and the corpus uses a tagset which is designed to tag CA text using traditional Arabic grammar [16], [22]. This need for annotated corpora, which are necessary for the development of many NLP systems, provided the motivation to create a manually annotated corpus for the Arabic language.

This research produces a manually annotated POS tagged corpus that is written in MSA. The tagset used in the new corpus was suggested by Alkhazi, Alghamdi and Teahan [13]; further details about the tagset will be discussed in the annotation tagset section (section III-B), and the annotation process follows the annotations guidelines prescribed by Maamour [59] .

II. Existing Resources

In 2001, the Linguistic Data Consortium (LDC) published the first versions of the Penn Arabic Treebank (ATB) [58]. This resource is widely used in many Arabic NLP applications such as the training of POS taggers, like the Madamira Arabic POS tagger [60] and the Stanford Arabic POS tagger [3]. The corpus consists of three parts with a total of 1 million annotated words. The first part v2.0 was a newswire text written in Modern Standard Arabic and consisted of 166K terms acquired from the Agence France Presse corpus. The second part was obtained from the Al-Hayat corpus which was distributed by Ummah Arabic News Text and consists of 144K [58]. The last part of the ATB corpus, part 3 v1.0, as shown Figure 3, is a newswire text obtained from the An-Nahar corpus and consists of about 350K morphologically annotated words. For non-members of the LDC, the cost of acquiring any part of the ATB corpus exceeds several thousand US dollars which prevents access to researchers with a limited budget [57], [58].

Khoja [61]–[63] has published a 50,000 terms manually annotated POS tagged corpus written in MSA text. According to the author, the corpus is divided into two parts; the first part is a newspaper text consisting of 1,700 terms that are manually tagged using a tagset that differentiates between the three moods of the verb and case structures of the noun [64]. The second part of the corpus is tagged using a simple tagset that includes only the following POS tags: noun, verb, particle, punctuation or number [62]. However, access to this resource was not provided.

Another annotated corpus was published by Mohit [56]. The AQMAR Arabic Wikipedia Dependency Tree Corpus is a manually annotated corpus that contains 1262 sentences collected from ten Arabic Wikipedia articles and the 36K terms of the corpus are manually annotated using the Brat annotation tool [56]. The ten articles were annotated for named entities beforehand [65]–[67] and cover topics such as Linux, Internet, Islamic Civilisation, Football, etc. The tagset used in this corpus contains a small number of tags and therefore cannot be used for the research concerning tag-based text compression.

The Columbia Arabic Treebank (CATiB) [27] is another manually annotated Treebank corpus that consists of newswire feeds, from the year 2004 to 2007 and written in MSA. The corpus was initially tokenized and then POS tagged by the MADA&TOKAN toolkit [15], [27]. The TrEd annotation interface [68] was utilised in the annotation process. The number of tags used by CATiB is relatively small as it consists only of six POS tags, NOM, PROP, VRB, VRB-PASS and PRT, where each tag comprises a group of subtags, for example, the tag "NOM" can be used to tag nouns, adverbs, pronouns and adjectives.

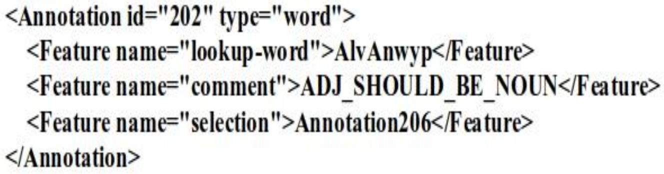
III. BAAC: The Bangor Arabic Annotated Corpus

The goal of this annotated corpus is to contribute by filling the gap created by the scarcity of freely available Arabic resources, manually annotated POS tagged corpora in particular, which is caused by the lack of availability and cost issues. Another goal is to provide a new resource required by many kinds of research, such as the ongoing tag-based text compression research conducted by Alkhazi, Alghamdi and Teahan [13], where the only annotation required at this stage is POS tags. The tagset used to annotate the new corpus is the same as used by the Madamira Arabic tagger, for reasons that will be discussed in the annotation tagset section (section B). Since the Madamira Arabic POS tagger is trained by the Arabic Treebank corpus [13], [14], and that corpus is written in MSA, the newly annotated corpus must also be written in MSA.

A. The Data Source.

The data source for the new corpus is the Press sub-corpus from the BACC corpus [28]. The BACC corpus was created originally to test the performance of various text compression algorithms on different text files. The results of the text classification performed by Alkhazi and Teahan [14] reveal that the Press sub-corpus is 99% written in MSA, as shown in Figure 2. According to the authors, the sub-corpus is a newswire text consisting of 51K terms, gathered from various news websites between 2010 and 2012 and covers many topics such as political and technology news.

B. The Annotation Tagset.



```
<Annotation id="202" type="word">
   <Feature name="lookup-word">AlvAnwyp</Feature>
   <Feature name="comment">ADJ_SHOULD_BE_NOUN</Feature>
   <Feature name="selection">Annotation206</Feature>
</Annotation>
```

Fig. 3. A sample POS tag from the ATB Part 3 v 1.0.

The tagset used in the BAAC corpus is the same as used by the Madamira tagger [60], which was used initially by the MADA tagger [15]. The tagset is the subset of the English tagset which was presented with the English Penn Treebank and consists of 32 tags and was initially proposed by Diab, Hacioglu and Jurafsky [69]. The experiments conducted by Alkhazi, Alghamdi and Teahan [13] have concluded that the quality of tag-based compression varies from one tagset to another. The different tagsets, some of which are shown in Table 3, were used to compress MSA text using POS tags, and tag-based compression using the Madamira tagset outperforms other tagsets such as Stanford [70] and Farasa [71]. Since one of the main goals of creating a gold-standard POS annotated text is to investigate the effect of manual annotation on the tag-based text compression, as described below in the experiments, therefore, the Madamira tagset, which outperformed other tagsets and consists of only 32 tags that are shown in Table 2, is used to annotate the BAAC POS tag and to create the ground-truth data which will be used later for training and evaluation purposes.

TABLE II. THE AGREEMENTS, DISAGREEMENTS AND BSERVED AGREEMENT

Tag	**Agreements**	**Disagreements**	**Observed Agreement %**
noun	23570	529	97.80
verb	5714	44	99.24
prep	5574	10	99.82
adj	4632	1235	78.95
noun_prop	2272	520	81.38
conj_sub	1534	17	98.90
conj	1148	79	93.56
pron_rel	992	37	96.40
pron_dem	767	11	98.59
noun_quant	574	1	99.83
part_neg	498	2	99.60
pron	367	6	98.39
adv	166	195	45.98
adj_comp	265	15	94.64
noun_num	252	7	97.30
part_verb	221	0	100.00
verb_pseudo	203	0	100.00
adj_num	156	26	85.71
adv_interrog	25	111	18.38
adv_rel	83	3	96.51
abbrev	60	2	96.77
part_restrict	59	16	78.67
part	25	27	48.08
pron_interrog	19	30	38.78
part_focus	14	9	60.87
part_interrog	22	0	100.00
part_fut	12	0	100.00
part_voc	10	0	100.00
part_det	8	2	80.00
interj	2	0	100.00
Total	**49244**	**2934**	**94.38%**

C. *Automatic POS Tagging.*

Madamira [60] was utilised to automatically tag the corpus by POS. The manual annotation process of the BAAC corpus followed annotation guidelines proposed by Maamouri [72] for annotating POS tags. All the previous corrections that are made to a tag are shown to the annotators during the process of annotation, as illustrated in section III-E, and the Madamira tagset used to annotate this corpus applies the criteria proposed by the author.

D. *The Annotation Tool.*

Most existing tools, such as TrEd tool [68], [73] which was used in the annotation of The Prague Dependency Treebank, are developed to annotate Treebank types of corpora, such as dependency trees corpora, that contain other information about the term, such as the gloss or a comment from an annotator, as shown in Figure 3. As mentioned earlier, the first stage of the BAAC annotation process will only add the POS tags to the corpus. Other linguistic information, such as the structural annotation, will be adapted in future work, therefore, the tool which will be used to manually annotate this corpus will only annotate POS tags. During the preparation for the annotation process, many constraints arose and defined four requirements that had to be met by the annotation tool. First, as the annotators are native Arabic speakers, a well-detailed Arabic translation of the tagset was provided with examples during the annotation process. Second, the software used for the annotation had to comply with the hardware and software requirements of the computer used to perform the annotation. Thirdly, the annotation tool, as shown in Figure 4, had to be executed on different operating systems, therefore, the tool was designed to be portable. Finally, online backing up procedures with the ID of the annotators was done to ensure the safety of the data.

TABLE III. DIFFERENT ARABIC TAGSETS

Term	**Madamira Tag**	**Stanford Tag**	**Farasa Tag**
الادارة	noun	DTNN	NOUN-FS
ترحب	noun_prop	VBP	E/ES/SV
بالتزام	verb	NN	NOUN-MS
الامين	noun	DTNN	NOUN-MS
العام	noun	DTJJ	ADJ-MS
بزيادة	adj	NN	NOUN-FS
عنصر	noun	NN	NOUN-MS
الميزانية	noun	DTNN	NOUN-FS
العادية	noun	DTJJ	ADJ-FS
لمكتب	noun	NN	NOUN-MS
الامم	noun	DTNN	NOUN-MP
المتحدة	noun	DTJJ	ADJ-MP

The previous requirements were met by developing a new annotation tool. First, a detailed Arabic translation of the tagset, which was obtained from Alrabiah [10] and then examined by Arabic specialists, was coded in the annotation tool as shown in Figure 4. The annotation tool also offers examples of the tag if required by the annotator. To comply with the hardware requirements and reduce memory dependency, the tool loads only one sentence to be modified at a time. To follow the Maamouri [72] annotation guidelines, the tool also displays the history of annotation by showing two types of modifications, the original tag assigned by the Madamira tagger and any tag chosen by previous annotators, if they exist. A current status of the annotation process is also displayed to the annotator, such as the number of annotated tags in the current session and the number of modified tags in the total document. The Java programming language was used to develop the annotation tool, and therefore, the tool can be executed on different operating systems. The tool also provided online backing up procedures each time the annotator modified a tag to eliminate any data loss.

E. *Data Preparation.*

After using Madamira [60] to automatically POS tag the corpus, a copy of the corpus was given to each annotator. Each copy was split into batches of documents that have 10-20 sentences and the ID of the annotator was coded with each batch to be used later in the evaluation section. The two annotators, who are native Arabic speakers and postgraduate students in Arabic Studies, started working to manually annotate the corpus on a full-time basis in two stages.

In the first stage of the annotation process, the annotators were required to work on-site to resolve any issues with the annotation tool and the annotation of the corpus was completed using the facilities provided by Tabuk Public Library. When the annotation process was finished, the two versions were evaluated and the Inter Annotator Agreement was calculated using two metrics, as will be discussed below in the BAAC evaluation section. The differences between the two versions were examined and adjusted off-site by a third annotator, who is a native Arabic speaker and PhD candidate student in Arabic Studies, to produce a final version of the corpus. The total time needed to annotate the corpus was two months – three weeks for the first stage and the rest for the final stage.

IV. BAAC Evaluation

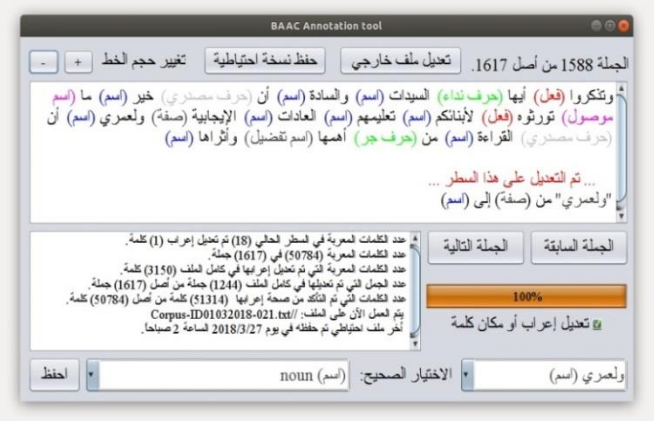
The quality of the annotated corpus affects the quality of the NLP application that utilises it. For instance, Reidsma and Carletta [74] has illustrated that the errors produced by machine learning tools are the same errors made by the annotators of the corpus that was used for training those tools. Two metrics were used to evaluate the quality of the BAAC, the Kappa coefficient [75] to calculate the inter-annotator agreement (IAA) among the two annotators and a direct percent agreement for each tag [76]. Using the data in Table 4, the obtained Kappa value is 0.956, which is recognised as perfect according to Landis and Koch [77]. The total observed agreement from Table 2, which displays the number of agreements and disagreements of different tags between the two annotators in a reverse frequency order, is 94.25%. Taking the number of tag occurrences into consideration, Table 2 shows that the tag verb or 'فعل' has the highest agreement between the annotators with 99.24% agreement. It also shows that the annotators agreed only 25 times out of 136 (18%) on the tag 'adv_interrog' or 'حال'. Also, the annotators agreed only on (45.98%) on the tag 'adv', and (38.78%) on the tag 'pron_interrog'. The reasons for such variation between the annotators were:

TABLE IV. The Bacc Agreement Table

	abbrev	adj	adj_comp	adj_num	adv	adv_interrog	adv_rel	conj	conj_sub	interj	noun	noun_num	noun_prop	noun_quant	part	part_det	part_focus	part_fut	part_interrog	part_neg	part_restrict	part_verb	part_voc	prep	pron	pron_dem	pron_interrog	pron_rel	verb	verb_pseudo	Total
abbrev	59								1															2							62
adj		4363	1	1							247		22											1					4		4639
adj_comp		7	260	1							4		2																6		280
adj_num				142								12	2																		156
adv		92		12	108		1	6	67		63		3	1										8							361
adv_interrog		106		6		9		1	3		7																		4		136
adv_rel		8					74																				1				83
conj								1148																							1148
conj_sub								52	1455																			32			1539
interj										2																					2
noun	1	1151		4	42		4	5	18		22762	2	98		1								1					4	41		24134
noun_num		3		15							5	235	1																		259
noun_prop		166	4	1	1			1	1		450	3	2121											11					36		2795
noun_qua		1									2			57																	576

nt													3																		
part					8			2	1		1		1		23						16										52
part_det		1									1					8															10
part_focus								9									14														23
part_fut																		12													12
part_interrog																			22												22
part_neg		1						1												498											500
part_restrict																					58			1							59
part_verb																						221									221
part_voc											1												9								10
prep		18						1			6		1								1			5556	1				2		5586
pron		4									1				1										366				1		373
pron_dem		1			7			1					1													767			1		778
pron_interrog						16	7		1															2			18	5			49
pron_rel													1											3				988			992
verb		18							4		20		19																5663		5724
verb_pseudo																														203	203
Total	60	5940	265	182	166	25	86	1227	1551	2	23570	252	2272	574	25	8	14	12	22	498	75	221	10	5584	367	767	19	1029	5758	203	

- The different understanding of the tag and, in some cases, its subset of tags by the annotators. For example, Table 4 shows that the two annotators disagreed concerning the tag 'noun' and the tag 'adj' in many instances. The different understanding of the tag 'adv_interrog' and the tag 'adj' has also caused a noticeable number of disagreements between the two annotators.
- Human error in the annotation process contributed to some of the errors in the annotated corpus. This was confirmed by random samples taken to be re-annotated by the same annotator.



الجملة 1588 من أصل 1617.

وتذكروا (فعل) أيها (حرف نداء) السيدات (اسم) والسادة (اسم) أن (حرف مصدري) خير (اسم) ما (اسم
موصول) تورثوه (فعل) لأبنائكم (اسم) تعليمهم (اسم) العادات (اسم) الإيجابية (صفة) ولعمري (اسم) أن
(حرف مصدري) القراءة (اسم) من (حرف جر) أهمها (اسم تفضيل) وأثراها (اسم)

... تم التعديل على هذا السطر ...
"ولعمري" من (صفة) إلى (اسم)

عدد الكلمات المعربة في السطر الحالي (18) تم تعديل إعراب (1) كلمة.
عدد الكلمات المعربة (50784) في (1617) جملة.
عدد الكلمات المعربة التي تم تعديل إعرابها في كامل الملف (3150) كلمة.
عدد الجمل التي تم تعديلها في كامل الملف (1244) جملة من أصل (1617) جملة.
عدد الكلمات التي تم التأكد من صحة إعرابها (51314) كلمة من أصل (50784) كلمة.
يتم العمل الآن على الملف: //Corpus-ID01032018-021.txt
آخر ملف احتياطي تم حفظه في يوم 2018/3/27 الساعة 2 صباحا.

ولعمري (اسم) ▾ الاختيار الصحيح: noun (اسم) ▾

Fig. 4. The Annotation tool.

TABLE V. THE TEN MOST FREQUENT TAGS BY THE FIRST ANNOTATOR

Tag	Frequency	%
noun	24099	47.52
verb	5714	11.27
prep	5574	10.99
adj	4632	9.13
noun_prop	2792	5.51
conj_sub	1534	3.02
conj	1148	2.26
pron_rel	992	1.96
pron_dem	778	1.53
noun_quant	575	1.13

The previous reasons were taken into consideration, and all the disagreements were highlighted, which was then given to the third annotator who went through all the disagreements and modified them based on his judgment. Finally, a final version of the corpus, which contains the agreements from the first two annotators and the agreements of the third one, was produced and used for further applications, as illustrated in the experiments section.

V. CORPUS STATISTICS

As stated, the text of the BAAC corpus was obtained from the sub-corpus Press of the BACC. The first annotator made 3150 changes to the originally tagged corpus and the second made 2959 modifications. Table 5 and Table 8 list the first ten most frequent tags for the annotators. The most frequent tag is 'noun' representing 47.52% for the first annotator and 46.48% for the second. The least used tag is 'noun_quant' being 1.13% of the tags for both annotators. A noticeable difference between the two annotators is the use of the tag 'adj' which represents 11.57% for the first annotator and occurring 1235 more times for the second annotator (9.13%).

Table 6 shows the ten most frequently used terms in the BAAC. The first and second most frequent words in the BAAC are 'في' which is a 'prep', that translates as 'in', and 'من', which is also a 'prep', that translates as 'from' representing 2.83% and 2.65% of the text respectively. The table also shows that the most commonly used bigram is 'من خلال', which translates as 'through' occurring 37 times in the corpus. Since the Press sub-corpus, which is the source of the BAAC, was gathered between 2010 and 2012 from several Arabic news websites, the most commonly used trigrams in the BAAC are ' في ميدان التحرير' which translates as 'In Tahrir Square', and ' الأعلى للقوات المسلحة' which translates as 'Higher Council of the Armed Forces', which were mentioned 12 times, and both trigrams relate to the events that happened in Egypt during the same period.

Figure 5 plots using log scales the ranked tag, bi-tag and tri-tag sequences versus their frequencies in the BAAC. There are 32 unique tags used in the annotated corpus, as mentioned earlier. The corpus also has 433 unique bi-tags where the sequence 'noun noun' dominates most of the bi-tags sequences. Finally, there are 2,113 distinct tri-tags used in the BAAC. The figure shows a Zipf's Law-like behaviour which mirrors the behaviour of a similar plot for the English language [78]. More details about the BAAC n-tag sequences are found in Table 7 and will be discussed below.

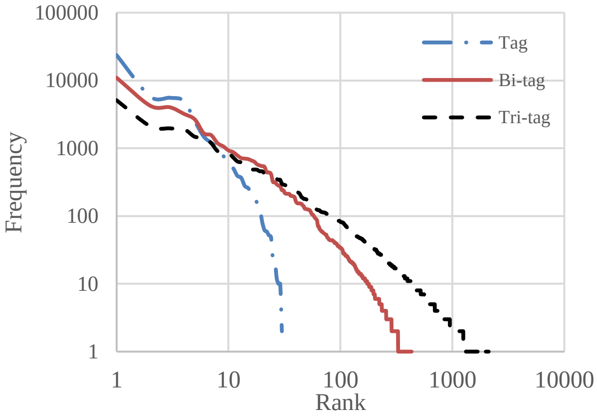


Fig. 5. Rank versus Tag, Bi-tag and Tri-tag Frequencies for the BAAC.

TABLE VI. WORD N-GRAM STATISTICS FROM THE BAAC

Rank	Word	Freq	%	Bigram	Freq	%	Trigram	Freq	%
1	في	1437	2.83	من خلال	37	0.07	في ميدان التحرير	12	0.02
2	من	1345	2.65	إلى أن	37	0.07	الأعلى للقوات المسلحة	12	0.02
3	و	735	1.45	الولايات المتحدة	34	0.07	المجلس الأعلى للقوات	11	0.02
4	أن	698	1.38	ميدان التحرير	30	0.06	القانون رقم لسنة	10	0.02
5	على	615	1.21	في مصر	28	0.05	غفر الله له	9	0.02
6	إلى	401	0.79	عدد من	28	0.05	قال أبو عبدالله	8	0.02
7	التي	352	0.69	من قبل	26	0.05	عبدالله غفر الله	8	0.02
8	عن	351	0.69	ثورة يناير	26	0.05	اللجنة الوطنية للاستقدام	8	0.02
9	أو	275	0.54	بعد أن	26	0.05	الكسب غير المشروع	8	0.02
10	لا	245	0.48	أن يكون	25	0.05	أبو عبدالله غفر	8	0.02

TABLE VII. MOST FREQUENT TAG, BI-TAG AND TRI-TAG SEQUENCES FROM THE BAAC

Rank	Tag	Freq	%	Bi-tag	Freq	%	Tri-tag	Freq	%
1	noun	23782	46.9	noun noun	11035	21.8	noun noun noun	5133	10.1
2	verb	5801	11.4	prep noun	4255	8.39	noun prep noun	2121	4.18
3	prep	5574	11	noun adj	4037	7.96	prep noun noun	1970	3.88
4	adj	4995	9.85	verb noun	3229	6.37	noun noun adj	1918	3.78
5	noun_prop	2532	4.99	noun prep	2679	5.28	noun adj noun	1482	2.92
6	conj_sub	1501	2.96	adj noun	1676	3.31	verb noun noun	1467	2.89
7	conj	1212	2.39	noun verb	1566	3.09	noun noun prep	1195	2.36
8	pron_rel	1025	2.02	verb prep	1190	2.35	noun verb noun	909	1.79
9	pron_dem	774	1.53	noun noun_prop	1066	2.1	verb prep noun	886	1.75
10	noun_quant	573	1.13	noun_prop noun_prop	932	1.84	adj noun noun	858	1.69

TABLE VIII. THE TEN MOST FREQUENT TAGS BY THE SECOND ANNOTATOR

Tag	Frequency	%
noun	23570	46.48
adj	5867	11.57
verb	5758	11.35
prep	5584	11.01
noun_prop	2272	4.48
conj_sub	1551	3.06
conj	1227	2.42
pron_rel	1029	2.03
pron_dem	767	1.51
noun_quant	574	1.13

Table 7 illustrates the ten most frequently used tag, bi-tag and tri-tag sequences in the BAAC. The tag 'noun' was utilised 23,782 times (46.9%) followed by the tag 'verb' that appeared in 11.44% of the text. The sequence of two nouns, the bi-tag 'noun noun', appeared in 11,035 occasions (21.76%), followed by the bi-tag 'prep noun' which was used 4,255 times in the BAAC. The sequence of three nouns came 5,133 times in the text, which represents 10.12% of the text, followed by the tri-tag 'noun prep noun' which came in 4.18% of the BAAC.

TABLE IX. MOST FREQUENT TAG, BI-TAG AND TRI-TAG SEQUENCES OF THE KHALEEJ SUB-CORPUS 'NEWS'

Rank	Tag	Freq	%	Bi-tag	Freq	%	Tri-tag	Freq	%
1	noun	485250	50.2	noun noun	243525	25.2	noun noun noun	122386	0.13
2	adj	120187	12.4	noun adj	91607	9.47	noun noun adj	49187	0.05
3	prep	104158	10.8	prep noun	81537	8.43	prep noun noun	43107	0.04
4	verb	91064	9.41	verb noun	52016	5.38	noun prep noun	39116	0.04
5	noun_prop	51985	5.37	noun prep	48968	5.06	noun adj noun	35544	0.04

To further analyse the n-tag results of the BAAC, Table 9 shows the tag, bi-tag and tri-tag statistics of the News sub-corpus from a different corpus, the Khaleej corpus [31], which also was tagged using Madamira tagger for comparison purposes. The sub-corpus contains 967K terms gathered from news websites. The table shows that both corpora, the News and the BAAC, share the same most frequent tag, bi-tag and tri-tag sequence, where the tag 'noun' in the sub-corpus News represents 50.2% of the text, the bi-tag 'noun noun' was used 243,525 times (25.2%) and the tri-tag 'noun noun noun' appeared in 0.13% of the text. These results confirm that the tag statistics are comparable between the different corpora.

TABLE X. TAG-BASED COMPRESSION RESULTS

Annotator	File size	Compressed size (bytes)	Compression ratio (bpc)
1	824,151	111,009	1.0776
2	819,482	110,954	1.0832
Original File	818,508	110,874	1.0837

VI. EXPERIMENTS AND FUTURE WORK

We have utilised the BAAC corpus in two applications, to evaluate the performance of the Madamira tagger, and to further investigate the tag-based text compression models as applied in by Alkhazi and Teahan [13]. Using the BAAC corpus, the Madamira tagger achieved an accuracy of 93.1%. To evaluate the effect of manual annotation on the tag-based text compression, the two versions of the BAAC were compressed using tag-based text compression models. The results of the compression were then compared to the compressed results of the original Madamira auto-tagged corpus. Table 10 illustrates the compression size (in bytes) and ratio (in bits per charactar) of all three files, and the results confirm that (1) manual annotation of the text reduces the quality of tag-based compression, as mentioned by Teahan and Alkhazi [13], [78]–[82], and (2) compressing the text using other text compression algorithms outperforms the tag-based text compression when compressing small text files, such as the BAAC corpus, as mentioned by Alkhazi and Teahan [13].

Further investigation is required to study the effect of using POS tagging systems, such as the OpenNLP project [83], trained using the BAAC on the tag-based text compression. Future work will add more annotated MSA text and will expand to cover CA text. More linguistic information, such as the structural annotation, will also be added to the BAAC to increase the possible NLP applications of the corpus.

VII. CONCLUSION

A new corpus, BAAC, was presented in this paper. It is an MSA corpus that contains 50K words manually annotated by part-of-speech tags. The annotated corpus used the same tagset utilised by the Madamira tagger and followed annotation guidelines proposed by Maamouri for annotating the POS tags. Also, a new annotation tool was developed and employed for the annotation process of BAAC which obtained a Kappa value of 0.956, and an average observed agreement of 94.25%. The BAAC was used to evaluate the Madamira tagger and to study the effect of the manual annotation on the performance of the tag-based Arabic text compression.

ACKNOWLEDGEMENT

Many thanks to our sponsor, Tabuk Public Library, for providing the facilities required by this research, and to its head, Mr Khalaf Alonazy, for his support. I would also like to show my appreciations to Ms Khawla S Alkhazi for the time and effort she spent contributing to this research.

References

[1] A. Soudi, A. Farghaly, G. Neumann, and R. Zbib, Challenges for Arabic machine translation, vol. 9. John Benjamins Publishing, 2012.

[2] M. A. Alghamdi, I. S. Alkhazi, and W. J. Teahan, "Arabic OCR Evaluation Tool," in Computer Science and Information Technology (CSIT), 2016 7th International Conference on, 2016, pp. 1–6.

[3] S. Green and C. Manning, "Better Arabic parsing: Baselines, evaluations, and analysis," COLING '10 Proc. 23rd Int. Conf. Comput. Linguist., no. August, pp. 394–402, 2010.

[4] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity, M. S. Khorsheed, and A. Al-Rajeh, "Automatic Arabic text classification," in Proceedings of The 9th International Conference on the Statistical Analysis of Textual Data, 2008.

[5] P. Damien, N. Wakim, and M. Egea, "Phoneme-viseme mapping for Modern, Classical Arabic language," in ACTEA'09. International Conference on Advances in Computational Tools for Engineering Applications, 2009., 2009, pp. 547–552.

[6] M. M. Najeeb, A. A. Abdelkader, and M. B. Al-Zghoul, "Arabic natural language processing laboratory serving Islamic sciences," Int. J. Adv. Comput. Sci. Appl., vol. 5, no. 3, 2014.

[7] K. C. Ryding, A reference grammar of modern standard Arabic. Cambridge university press, 2005.

[8] M. A. Alghamdi and W. J. Teahan, "A New Thinning Algorithm for Arabic Script," Int. J. Comput. Sci. Inf. Secur., vol. 15, no. 1, p. 204, 2017.

[9] K. Dukes and N. Habash, "Morphological Annotation of Quranic Arabic.," in LREC, 2010.

[10] M. S. Alrabiah, "Building A Distributional Semantic Model for Traditional Arabic and Investigating its Novel Applications to The Holy Quran," Ph.D. thesis, King Saud University, 2014.

[11] K. Dukes, "Statistical parsing by machine learning from a Classical Arabic treebank," Ph.D. thesis, University of Leeds, 2013.

[12] A. Alosaimy and E. Atwell, "Tagging Classical Arabic Text using Available Morphological Analysers and Part of Speech Taggers," J. Lang. Technol. Comput. Linguist., 2017.

[13] I. S. Alkhazi, M. A. Alghamdi, and W. J. Teahan, "Tag based models for Arabic Text Compression," in 2017 Intelligent Systems Conference (IntelliSys), 2017, pp. 697–705.

[14] I. S. Alkhazi and W. J. Teahan, "Classifying and Segmenting Classical and Modern Standard Arabic using Minimum Cross-Entropy," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 4, pp. 421–430, 2017.

[15] N. Habash, O. Rambow, and R. Roth, "MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization," in Proceedings of the 2nd international conference on Arabic language resources and tools (MEDAR), Cairo, Egypt, 2009, pp. 102–109.

[16] K. Dukes and T. Buckwalter, "A dependency treebank of the Quran using traditional Arabic grammar," in Informatics and Systems (INFOS), 2010 The 7th International Conference on, 2010, pp. 1–7.

[17] R. R. Jablonkai and N. Čebron, "Corpora as Tools for Self-Driven Learning: A Corpus-Based ESP Course," in Student-Driven Learning Strategies for the 21st Century Classroom, IGI Global, 2017, pp. 274–298.

[18] R. Mitkov, The Oxford handbook of computational linguistics. Oxford University Press, 2005.

[19] M. Wynne, "Archiving, distribution and preservation," Dev. Linguist. corpora A Guid. to good Pract., pp. 71–78, 2005.

[20] R. A. Abumalloh, H. M. Al-Sarhan, and W. Abu-Ulbeh, "Building Arabic Corpus Applied to Part-of-Speech Tagging," Indian J. Sci. Technol., vol. 9, no. 46, 2016.

[21] R. Z. Xiao, "Theory-driven corpus research: using corpora to inform aspect theory," Lüdeling, A., Kytö, M. Corpus Linguist. An Int. Handb., vol. 2, pp. 987–1008, 2008.

[22] S. Alkahtani, "Building and verifying parallel corpora between Arabic and English," Ph.D. thesis, Bangor University, 2015.

[23] W. Zaghouani, "Critical survey of the freely available Arabic corpora," in In LREC'14 Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT), 2017, pp. 1–8.

[24] F. Al Shamsi and A. Guessoum, "A hidden Markov model-based POS tagger for Arabic," in Proceeding of the 8th International Conference on the Statistical Analysis of Textual Data, France, 2006, pp. 31–42.

[25] S. Atkins, J. Clear, and N. Ostler, "Corpus design criteria," Lit. Linguist. Comput., vol. 7, no. 1, pp. 1–16, 1992.

[26] C. F. Meyer, English corpus linguistics: An introduction. Cambridge University Press, 2002.

[27] N. Habash and R. M. Roth, "Catib: The columbia arabic treebank," in Proceedings of the ACL-IJCNLP 2009 conference short papers, 2009, pp. 221–224.

[28] K. M. Alhawiti, "Adaptive models of Arabic text," Ph.D. thesis, Bangor University, 2014.

[29] "Ajdir Corpora." [Online]. Available: http://aracorpus.e3rab.com/argistestsrv.nmsu.edu/AraCorpus/. [Accessed: 22-Sep-2018].

[30] "King Saud University Corpus of Classical Arabic (KSUCCA) | Maha Al-Rabiah – Blog." [Online]. Available: https://mahaalrabiah.wordpress.com/2012/07/20/king-saud-university-corpus-of-classical-arabic-ksucca/. [Accessed: 22-Sep-2018].

[31] M. Abbas, K. Smaili, and D. Berkani, "Evaluation of Topic Identification Methods for Arabic Texts and their Combination by using a Corpus Extracted from the Omani Newspaper Alwatan," Arab Gulf J. Sci. Res., vol. 29, no. 3–4, pp. 183–191, 2011.

[32] T. Zerrouki and A. Balla, "Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems," Data Br., vol. 11, pp. 147–151, Apr. 2017.

[33] "Khaleej corpus." [Online]. Available: https://sourceforge.net/projects/arabiccorpus/. [Accessed: 22-Sep-2018].

[34] "UN Corpus(Arabic portion)." [Online]. Available: http://www.sibawayh-nlp.com/?q=node/1158. [Accessed: 23-Sep-2018].

[35] "Islamic books | Leicester | Al Kunuz." [Online]. Available: https://www.alkunuz.co.uk/. [Accessed: 23-Sep-2018].

[36] A. El Shamsy, "Al-Shāfiʾī's Written Corpus: A Source-Critical Study," J. Am. Orient. Soc., vol. 132, no. 2, pp. 199–220, 2012.

[37] "Meedan's Open Source Arabic/English Translation Memory." [Online]. Available: https://github.com/anastaw/Meedan-Memory. [Accessed: 23-Sep-2018].

[38] M. Altamimi, O. Alruwaili, and W. J. Teahan, "BTAC: A Twitter Corpus for Arabic Dialect Identification," in Proceedings of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-corpora 2018), 2018, pp. 5–10.

[39] A. O. Al-Thubaity, "A 700M+ Arabic corpus: KACST Arabic corpus design and construction," Lang. Resour. Eval., vol. 49, no. 3, pp. 721–751, 2015.

[40] L. Al-Sulaiti and E. S. Atwell, "The design of a corpus of Contemporary Arabic," Int. J. Corpus Linguist., vol. 11, no. 2, pp. 135–171, 2006.

[41] S. Alansary and M. Nagi, "The international corpus of Arabic: Compilation, analysis and evaluation," in Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), 2014, pp. 8–17.

[42] Linguistic Data Consortium., Buckwalter Arabic morphological analyzer : Version 1.0. Linguistic Data Consortium, 2002.

[43] "Arabic-English Learner's Dictionary." [Online]. Available: http://www.perseus.tufts.edu/hopper/opensource/download. [Accessed: 22-Sep-2018].

[44] M. Attia, P. Pecina, A. Toral, L. Tounsi, and J. van Genabith, "A lexical database for modern standard Arabic interoperable with a finite state morphological transducer," in International Workshop on Systems and Frameworks for Computational Morphology, 2011, pp. 98–118.

[45] "Arabic Wordlist for Spellchecking." [Online]. Available: https://sourceforge.net/projects/arabic-wordlist/. [Accessed: 29-Sep-2018].

[46] K. Almeman, M. Lee, and A. A. Almiman, "Multi dialect Arabic speech parallel corpora," in Communications, Signal Processing, and their

Applications (ICCSPA), 2013 1st International Conference on, 2013, pp. 1–6.

[47] S. Al Maadeed, W. Ayouby, A. Hassaine, and J. M. Aljaam, “QUWI: an Arabic and English handwriting dataset for offline writer identification,” in Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on, 2012, pp. 746–751.

[48] I. Bensalem, P. Rosso, and S. Chikhi, “A new corpus for the evaluation of arabic intrinsic plagiarism detection,” in International Conference of the Cross-Language Evaluation Forum for European Languages, 2013, pp. 53–58.

[49] Y. Benajiba, P. Rosso, and J. M. G. Soriano, “Adapting the JIRS passage retrieval system to the Arabic language,” in International Conference on Intelligent Text Processing and Computational Linguistics, 2007, pp. 530–541.

[50] D. Chiang, M. Diab, N. Habash, O. Rambow, and S. Shareef, “Parsing Arabic dialects,” in 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006.

[51] R. Steinberger, B. Pouliquen, M. Kabadjov, and E. der Goot, “JRC-Names: A freely available, highly multilingual named entity resource,” CoRR, vol. abs/1309.6, 2013.

[52] Y. Benajiba, P. Rosso, and J. M. Benedíruiz, “Anersys: An arabic named entity recognition system based on maximum entropy,” in International Conference on Intelligent Text Processing and Computational Linguistics, 2007, pp. 143–153.

[53] M. I. Alkanhal, M. A. Al-Badrashiny, M. M. Alghamdi, and A. O. Al-Qabbany, “Automatic stochastic arabic spelling correction with emphasis on space insertions and deletions,” IEEE Trans. Audio. Speech. Lang. Processing, vol. 20, no. 7, pp. 2111–2122, 2012.

[54] S. Jeblee, H. Bouamor, W. Zaghouani, and K. Oflazer, “CMUQ@QALB-2014: An SMT-based System for Automatic Arabic Error Correction,” in Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), 2014, pp. 137–142.

[55] S. Pradhan and L. Ramshaw, “OntoNotes: Large Scale Multi-Layer, Multi-Lingual, Distributed Annotation,” in Handbook of Linguistic Annotation, Springer, 2017, pp. 521–554.

[56] B. Mohit, “Named Entity Recognition,” in Natural Language Processing of Semitic Languages, I. Zitouni, Ed. Springer, USA, 2014.

[57] J. Hajic, O. Smrz, P. Zemánek, J. Šnaidauf, and E. Beška, “Prague Arabic dependency treebank: Development in data and tools,” in Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools, 2004, pp. 110–117.

[58] M. Maamouri, A. Bies, T. Buckwalter, H. Jin, and W. Mekki, “Arabic Treebank: Part 3 (full corpus) v 2.0 (MPG + Syntactic Analysis),” LDC2005T20, 2005. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2005T20. [Accessed: 25-Nov-2016].

[59] M. Marcus et al., “The Penn Treebank: annotating predicate argument structure,” in Proceedings of the workshop on Human Language Technology, 1994, pp. 114–119.

[60] A. Pasha et al., “MADAMIRA : A Fast , Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic,” Proc. 9th Lang. Resour. Eval. Conf., pp. 1094–1101, 2014.

[61] S. Khoja, R. Garside, and G. Knowles, “A tagset for the morphosyntactic tagging of Arabic,” Proc. Corpus Linguist. Lancaster Univ., vol. 13, 2001.

[62] S. Khoja, “APT: An automatic arabic part-of-speech tagger,” Lancaster University, 2003.

[63] Y. O. M. Elhadj, A. Abdelali, R. Bouziane, and A. H. Ammar, “Revisiting Arabic Part of Speech Tagsets,” Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA, vol. 2014, pp. 793–802, 2015.

[64] J. A. Haywood, H. M. Nahmad, and G. W. Thatcher, A new Arabic grammar of the written language. Lund Humphries London, 1965.

[65] N. Schneider, B. Mohit, K. Oflazer, and N. A. Smith, “Coarse lexical semantic annotation with supersenses: an Arabic case study,” in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, 2012, pp. 253–258.

[66] Behrang Mohit, “AQMAR Arabic Dependency Corpus.” [Online]. Available: http://www.cs.cmu.edu/~ark/ArabicDeps/. [Accessed: 18-Sep-2018].

[67] Behrang Mohit, “brat rapid annotation tool.” [Online]. Available: http://brat.nlplab.org/. [Accessed: 16-Sep-2018].

[68] “TrEd User’s Manual.” [Online]. Available: https://ufal.mff.cuni.cz/tred/documentation/tred.html. [Accessed: 20-Sep-2018].

[69] M. Diab, K. Hacioglu, and D. Jurafsky, “Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks,” in Proceedings of HLT-NAACL 2004: Short papers, 2004, pp. 149–152.

[70] S. Green, M.-C. de Marneffe, and C. D. Manning, “Parsing Models for Identifying Multiword Expressions,” Comput. Linguist., vol. 39, no. 1, pp. 195–227, Mar. 2013.

[71] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, “Farasa: A fast and furious segmenter for arabic,” in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 2016, pp. 11–16.

[72] M. Maamouri, A. Bies, and S. Kulick, “Enhancing the Arabic Treebank: a Collaborative Effort toward New Annotation Guidelines.,” in LREC, 2008, pp. 3–192.

[73] P. Pajas and J. Štěpánek, “Recent advances in a feature-rich framework for treebank annotation,” in Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, 2008, pp. 673–680.

[74] D. Reidsma and J. Carletta, “Reliability measurement without limits,” Comput. Linguist., vol. 34, no. 3, pp. 319–326, 2008.

[75] J. Cohen, “A coefficient of agreement for nominal scales,” Educ. Psychol. Meas., vol. 20, no. 1, pp. 37–46, 1960.

[76] M. L. McHugh, “Interrater reliability: the kappa statistic,” Biochem. medica Biochem. medica, vol. 22, no. 3, pp. 276–282, 2012.

[77] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” Biometrics, pp. 159–174, 1977.

[78] W. J. Teahan and J. G. Cleary, “Tag Based Models of English Text.,” in Data Compression Conference, 1998, pp. 43–52.

[79] W. J. Teahan, “Modelling English text,” Ph.D. thesis, Waikato University, 1998.

[80] W. J. Teahan and D. J. Harper, “Using compression-based language models for text categorization,” in Language modeling for information retrieval, Springer, 2003, pp. 141–165.

[81] Z. Chang, “A PPM-based Evaluation Method for Chinese-English Parallel Corpora in Machine Translation,” no. September, pp. 1–106, 2008.

[82] W. J. Teahan, Y. Wen, R. McNab, and I. H. Witten, “A compression-based algorithm for Chinese word segmentation,” Comput. Linguist., vol. 26, no. 3, pp. 375–393, 2000.

[83] T. Morton, J. Kottmann, J. Baldridge, and G. Bierner, “Opennlp: A java-based nlp toolkit,” 2005.