The Guidelines of Building a Treebank for Modern Standard Arabic

Amena Dheif
Phonetics and linguistics Department,
Faculty of Arts, Alexandria University
Teaching assistant of computational linguistics
amena.helmy@alexu.edu.eg

Ahmed Abd El Ghany
Phonetics and linguistics Department,
Faculty of Arts, Alexandria University
lecturer of computational linguistics
abdelghany.ma@umk.edu.my

Sameh Al Ansary
Phonetics and linguistics Department,
Faculty of Arts, Alexandria University
Professor of computational linguistics
Sameh.Alansary@bibalex.org

***Abstract*— Treebanks are one of the most needed and used linguistic resources in the fields of Natural language processing (NLP) and Natural language understanding (NLU). Arabic has only two constituency-based treebanks and a number of dependency treebanks. The current research presents the guidelines for building a parsed Arabic treebank for Modern Standard Arabic (MSA). The guidelines show, firstly the choice of the grammar formalism, then the genre and size of the treebank, and finally the annotation layers of the treebank. The study also shows that using the traditional Arabic grammar syntactic theory to describe the Arabic syntax has proven to be more suitable than using any of the modern syntax theories. Working with the traditional Arabic grammar also helps avoid the errors that the available treebank fell in as a result of using guidelines that don't suit the Arabic grammar. The study adopts three layers of annotations: the morphological layer, the syntactic layer, and the grammatical function layer. The resultant tree is a very detailed and rich syntactic tree, which is preferable by the researcher over having a huge amount of data poorly and shallowly annotated.**

Keywords*— *Treebanking, Arabic grammatical functions, Parsed Treebank.

I. INTRODUCTION

Treebanks can be defined as linguistically annotated corpora mainly at the morphological and syntactic levels [1], [2], [3]. The term treebank is usually used to refer to syntactic annotations, yet recently the term has been used to refer to corpora annotated with semantic annotations such as the predicate argument structures annotations [3]. The term treebank was created by Geoffrey Leech [4], driven by the fact that the syntactic annotations are represented by the tree structure, however, the term now applies to all the syntactically annotated corpora whether represented in tree structures or not [2]. In [2], it was indicated that the term treebank is different from the term parsed corpora, where the treebank term usually describes corpora that are manually annotated or manually post-edited, whereas the term parsed corpora defines corpora that are automatically analyzed.

According to NLP researchers, treebanks are one of the most needed and used linguistic resources in the tasks of NLP and NLU [5]. Treebanks play also a crucial role in pure linguistic studies, thus corpus linguists prefer building treebanks that can serve both ends since the creation of a treebank isn't easy and is both time and experts consuming [1].

A. *Treebanks' importance in pure linguistic research*

Treebanks are invaluable resources for corpus-based linguistic studies, especially linguistic studies that are related to syntax or dependent on syntax [2]. They can be used in qualitative research, as repositories that can be searched to find examples of certain linguistic constructions, or counter-examples to a syntactic hypothesis or even an established theory [2]. Treebanks can also be used to evaluate whether a grammatical formalism is suitable for describing certain languages, as in the case of the Prague Dependency Treebank [6], where one of the project's main goals was to find out whether the Functional Generative Description is appropriate for the describing contemporary Czech [7]. Treebanks can also be used to evaluate the coverage and consistency of available grammars. In [8] an English handcrafted grammar based on the Tree Adjoining Grammar XTAG formalism was tested for both coverage and consistency by comparing it with a grammar of the same format extracted from the Penn Treebank [7].

B. *Treebanks' importance in language pedagogy*

In language pedagogy, corpora in general play a major role in the process of foreign language teaching, as they help in, firstly, answering the students' questions and queries through accessing available data [9], secondly, in the construction of teaching materials, and finally, in the construction of examination and training resources for the learners.

There is a number of projects that have been put up using treebanks to achieve the previous goals, as in the case of building the Clause Pattern Database (CPDB) [10], where a wide representative sample of the clause patterns of English was collected and represented in the form of a treebank to help learners acquire that difficult construction. The Englicious web platform [11] is another project that relied on the parsed British section of the International Corpus of English, to build training and testing materials for the students [11].

Although the usage of supporting tools and corpora in the field of teaching Arabic as a foreign language is still in its preliminary stages, and the grammar materials are mostly represented in the shape of grammar textbooks [12], some researchers lately consider using corpora and treebanks in the process of teaching Arabic as a foreign language, as in the case of the Quranic Arabic corpus QAC [13], since the QAC corpus builders sought using the QAC corpus in the development of Arabic grammar teaching materials [13].

C. *Treebanks' importance in natural language processing:*

The field of Arabic natural language processing (ANLP) [14] has benefited from the availability of Arabic treebanks such as the Penn Arabic Treebank (PATB) [5] in the development of many technologies, as it provides annotations on various levels. Many Tokenizers were trained on the PATB like [15], [16], [17], [18], [19], [20]. Also a number of statistical-based part of speech taggers were carried out using

the PATB as in [16], [17], [21], [22], [23], [24] [18], [19], [25], [26]. A number of chunkers were also developed based on the PATB as in [16], [24] [18], [27], [19]. And finally and most importantly, PATB was used in building statistical syntactic parsers through the induction of probabilistic grammars from the treebank [1] as in the projects of [28], [29], [30], [31], [32].

II. Literature Review

Since Arabic is a widespread language with a population of around 273.9 Million speakers [33], and the second language of another 60 million speakers [34], and since Arabic has a rich and complex grammar, these characteristics entitle Arabic to have numerous linguistic resources. However, the state of art shows otherwise, as Arabic has only two constituency treebanks namely the Penn Arabic Treebank (PATB) [5] and the syntactic Arab Treebank (SATB) [35], and a few number of dependency treebanks. And therefore Arabic natural language processing (ANLP) is more challenging [14].

A. *The Penn Arabic Treebank (PATB)*

The Penn Arabic Treebank (PATB) [5] is a long-term project built to support Arabic natural language processing (ANLP) research [5]. The PATB project was developed originally to work on the newswire genre of the written modern standard Arabic (MSA), and then it was expanded to include both dialectical Arabic [36] and spoken modern standard Arabic [37].

The PATB [5] project adopted the Penn treebank annotation scheme, firstly, to produce an Arabic treebank compatible with the other available Penn treebanks, secondly, to be able to work with the available tools and software used with the other treebanks, and finally, because the LDC team found out that training the annotators to work with this annotation style is easier than training them to work with the traditional Arabic grammar [38].

The PATB [5] provides data with both morphological and syntactic annotations.

The Morphological layer, on one hand, is more of a morphological analysis process than it is a part-of-speech tagging. It provides the tokens with their morphological, morphosyntactic, and gloss information [39]. The morphological analysis phase was performed automatically using the Tim Buckwalter Arabic morphological analyzer (BAMA) [40]. The tagset adopted by Buckwalter consists of 54 atomic tags [40]. These tags were used to provide each token with a compound tag for the core word, the affixes, and the clitics attached to it to represent the rich and complex morphology of Arabic [41]. The number of compound tags used in the ATB: Part 3 v 2.0 was 1290 tags [42] which is quite a large number.

The syntactic layer, on the other hand, provides syntactic, semantic, and logical annotations of each sentence, as it provides the constituent structures of the word sequences along with their constituent tags, the function categories (semantic role tags) of some constituents, and finally determine the null categories if available [39]. The syntactic annotation was performed semi-automatically using Dan Bikel's parsing engine [43] and was then corrected by the annotators [38]. The Constituents tagset used in the PATB is composed of 23 tags, and the function tagset consists of 20 tags [42].

The PATB has undergone a revision and correction process [44] since the reported parsing scores of the parsers trained on the PATB were very low compared to the scores of the parsers trained on the English and Chinese Penn treebanks, and as the project received many critics concerning applying the English grammar rules to Arabic which yielded many erratic annotations both in the morphological and the syntactic layer [44].

The researcher believes that the annotation scheme of the PATB suffers from many drawbacks. Firstly, the constituents tagset adopted by the PATB is inconsistent, as it borrows two tags from the X-Bar theory namely the SBAR and SBARQ tags, while the other tags are adopted from the phrase structure grammar theory. Secondly adopting a redundant constituent tag like the QP tag which is simply an NP. Thirdly, the inconsistent annotations of the function tags in the PATB, as some NPs and PPs get to have a function tag and others don't as shown in figure (1). Also adding the function tags on the same level of the constituent tags is another drawback, as they should be tagged on a separate level. Thirdly, the inconsistent syntactic annotation of some constructions, like the modified NPs, where an ADJP postmodifier of an NP is annotated as a sister to the Noun of the NP construction, but if the postmodifier is a PP or SBAR constituent, it is annotated as a sister of the whole NP as shown in figure (2). Finally, using annotation guidelines that differ from and conflict with the Arabic traditional syntax.

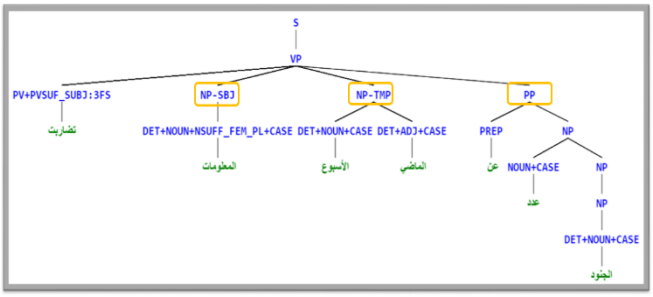

Figure 1: A part from a PATB tree showing the inconsistency of adding the function tags *[42]*.

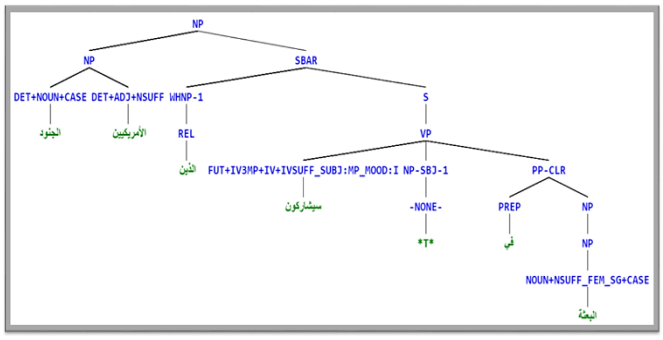

Figure 2: A part from a PATB tree showing the inconsistency of adding Post-modifiers to an NP construction *[42]*.

B. *Syntactic Arab Treebank (SATB)*

The Syntactic Arab Treebank (SATB), [35] is a 100,000 words project built using the Islam online corpus. The SATB provides 2 layers of annotation. The first layer provides the POS tagging where 62 tags were adopted following the tagging methodology provided in [45]. The second layer provides syntactic annotations following the Phrase Structure grammar. The second layer also provides semantic functions of some constituents. The semantic tags are of three classes namely function tags, semantic roles tags, and dependency tags [35].

The SATB share some of the drawbacks the PATB suffered from, like adding the function tags and the constituent tags at the same level, and the inconsistent adding of the semantic tags as some constituents get to have semantic tags and other constituents don't.

C. *Dependency Arabic Treebanks*

Arabic has many Dependency treebanks. The Prague dependency Arabic Treebank (PADT) [46] was built following the Prague dependency Czech treebank guidelines with some necessary adjustments to account for the differences between Arabic and Czech. [47]. It provides morphological, syntactic, and tectogrammatical annotations of Arabic following the theory of Functional Generative Description [46]. According to [48], PADT adopts 31 dependency relations. One main issue with PADT is that the adopted relations are very complex to comprehend and theory-dependent.

The Columbia Arabic Tree Bank (CATiB) [49] is another Arabic dependency treebank. The main goal of CATiB was to annotate a large amount of data syntactically in a short time. Thus a small number of eight dependency relations were adopted [49]. The researcher sees that using such a small number of dependency relations results in a poor representation of the richness of the Arabic grammar.

Finally, The Quranic dependency Arabic treebank (QDAT) [50] project adopted the Arabic traditional grammar theory (Irab) to represent the treebank dependency relations. The researcher believes that adopting the traditional Arabic grammar functions is a strength point for the QDAT project.

III. METHODOLOGY

In treebanking, The treebank design is established based on a number of principles, such as the intended usage of the treebank, the availability of the workforce and the analysis tools, the timeframe set for the completion and release of the treebank, and most importantly the number of annotation layers the treebank provides [2]. The following subsections provide the guidelines followed to build the ongoing project of the Arabic treebank for Modern Standard Arabic.

A. *Choosing the grammar formalism*

Choosing the grammar formalism that the treebank annotation shall follow is a cornerstone in any treebank scheme. In general, the choice falls between constituency representation and dependency representation. The constituency representation is more suitable for languages with fixed word order, yet it is more difficult and demanding [1]. Dependency representation, on the other hand, are more adequate for languages with free word order [1]. In the current study, the constituency representation was chosen, since Arabic has only two constituency-based treebanks, and for the wide range of usages the constituency-based treebanks offer to the NLP and NLU fields.

B. *Choosing treebank size and genres*

Linguistic annotation is a laborious task, even with the recent improvements in automating the annotation process. Thus, there is a natural trade-off between the treebank size and the depth of the annotation layers used to annotate the data [2]. Thus treebanks with rich syntactic annotations tend to be smaller in size. In the current study, it was preferred to build a small treebank with detailed and rich annotations like the case with the SUSANNE corpus [7].

Regarding the genre of the treebank, the researcher believes that choosing different genres helps in covering different types of syntactic constructions that are yielded by the genre difference, thus the ongoing project draws data from three different genres namely the news, economy, and sports genres.

C. *Choosing the treebank annotation layers*

In treebanking, it is preferable to add rich annotation layers to data so that the resource becomes more reusable and sharable [1]. The ongoing project consists of three layers of annotations which are: the word layer, the syntactic annotation layer, and the function layer.

The word layer or the lower layer of annotation often provides the morphosyntactic features of the terminal units of the treebank [2]. In agglutinative languages like Arabic, the terminal nodes tend not to be words tokenized with whitespaces or punctuations delimiters, rather a process of tokenization occurs where the morphs of a single word are separated. For Arabic, there are many tokenization schemes available, like the ones shown in figure (3). The model of the ongoing project adopts the D3Tok scheme, as it separates all the proclitics and enclitics that constitute a constituent in the syntactic tree. The word layer also provides the part of speech tags, morphological, morpho-syntactic, and semantic attributes of each unit. Table (I) and Table (II) present the POS tagset adopted for the ongoing project along with the features specified for each POS tag. This strategy uses a small number of tags and adds a rich number the features for each tag following the classification found in [51].

word	ATB Tok	D3 Tok	bw Tok
مثل	مثل	مثل	مثل
المؤمنين	المؤمنين	ال_+مومنين	ال_+مومن+_ين
في	في	في	في
توادهم	تواد+_هم	تواد+_هم	تواد+_هم
،	،	،	،
و	و	و	و
تراحمهم	تراحمهم	تراحمهم	تراحمهم
،	،	،	،
و	و	و	و
تعاطفهم	تعاطف+_هم	تعاطف+_هم	تعاطف+_هم
مثل	مثل	مثل	مثل
الجسد	الجسد	ال_+جسد	ال_+جسد
،	،	،	،
إذا	إذا	إذا	إذا
اشتكى	اشتكى	اشتكى	اشتكى
منه	من+_ه	من+_ه	من+_ه
عضو	عضو	عضو	عضو
تداعي	تداعي	تداعي	تداعي
له	ل+_ه	ل+_ه	ل+_ه
سائر	سائر	سائر	سائر
الاعضاء	الأعضاء	ال_+أعضاء	ال_+اعضاء
بالسهر	ب_+السهر	ب_+ال_+سهر	ب_+ال_+سهر
و	و	و	و
الحمي	الحمى	ال_+حمى	ال_+حمى

Figure 3: Popular tokenization schemes of Arabic

TABLE I. : THE ADOPTED POS TAGSET FOR ONGOING STUDY.

POS Tagset
Noun
Adjective
Verb
Incomplete verb
Preposition
Particle
Conjunction
Adverb
Interrogative Pronoun
Part_N
Punctuation

TABLE II. : THE ADOPTED MORPHOLOGICAL, MORPHOSYNTACTIC, AND SEMANTIC FEATURES OF THE CURRENT STUDY

Features used	Description	POS tags applied to
Noun Type	Shows the class of the noun whether it is a common noun, proper noun, etc.	Noun Adjective
Gender	Shows the word gender.	Noun Adjective Verb Incomplete verb
Number	Shows the number of the word whether it is singular, dual, etc.	Noun Adjective Verb Incomplete verb
Person	Shows the person each word presents whether it is first, second, etc.	Noun Adjective Verb Incomplete verb
Definiteness state	Shows the definiteness state of a word whether it's definite, indefinite, definite by itself, etc.	Noun Adjective
Case	Shows the case of a word whether it's nom, acc, or gen.	Noun Adjective Adverb Interrogation pronoun Part_N
Noun semantic class	Shows a semantic feature of the word whether it is rational, irrational, person, etc.	Noun Adjective Adverb Part_N
Interrogative class	Shows whether a word represents time interrogative, location interrogative, etc.	Interrogation pronoun
Aspect	Shows the verb if it indicates a complete event or incomplete event	Verb Incomplete verb
Mood	Shows the mood of a verb, whether it is indicative, jussive, etc.	Verb Incomplete verb
Voice	Shows the voice of a verb whether it is in the active or passive voice.	Verb Incomplete verb
Transitivity	Shows if a verb is intransitive, transitive, etc.	Verb
Work	Shows the word effect on the next word, if it has any.	Particle Preposition Conjunction
Specificity	Shows whether the word works on nouns, verbs, or both of them.	Particle Preposition Conjunction
Lemma	Shows the lemma of the word	Particle Preposition Conjunction Adverb
Particle semantic meaning	Shows the semantic meaning of the particle, whether it indicates negation, prohibition, emphasis, etc.	Particle Preposition Conjunction

The second layer is the syntactic layer which provides the constituency analysis of each sentence. The current project adopts the constituents' labels shown in table (III). The project also adopts the strategy of building a parsed treebank, where the treebank is generated automatically using a parser built by the researcher, with no postediting. The parser generates a number of trees, and the annotator's job is to choose the perfect tree that best fits the sentence's meaning.

TABLE III. : THE ADOPTED LIST OF CONSTITUENTS' LABELS USED IN THE CURRENT PROJECT.

Chosen Categories labels	Description
S_n	Nominal sentence
S_v	Verbal Sentence
NP	Noun phrase
PP	Prepositional phrase
ADJP	Adjectival phrase
ADVP	Adverbial phrase
Rel Clause	Relative clause
Coord NP	Coordinated noun phrase
Coord ADJP	Coordinated Adjectival phrase
Coord PP	Coordinated prepositional phrase
Coord ADVP	Coordinated adverbial phrase

The third layer is the functions layer, and it provides the grammatical function tag of each constituent in the tree. The functions were chosen from the traditional grammar Arabic theory (الإعراب), as it describes the Arabic rich grammatical constructions better than any modern syntactic theory since many of these theories were developed to describe the English language. Although these theories should be able to describe other languages' syntax, this notion stays unaccomplished. Table (IV) shows the adopted function tags along with their name in the Arabic grammar if found and a sentence/s from the current project that contains a constituent representing the function tag in question.

TABLE IV. THE ARABIC GRAMMATICAL FUNCTION TAGS USED IN THE ONGOING PROJECT

Grammatical functions tags	Description in Arabic	Examples
Topic	مبتدأ	- تنظيم الدولة"يهدد بإعدام مختطفات السويداء"
Comment	خبر	- تنظيم الدولة"يهدد بإعدام مختطفات السويداء"
Subject x	اسم إن أو كان	- إن المستحيل ليس تونسيا
Predicate x	خبر إن أو كان	- إن المستحيل ليس تونسيا
Pre determiner		- إن المستحيل ليس تونسيا
Post determiner	نعت أو متعلق بالاسم أو الصفة	- وتابعت"إنه أمر مضحك وميك." - ولم يعرف على الفور محتوى البيان الذي سيلقيه المهاجمون . - وأفادت تقارير أن الحريق طال الحاسبات والأجهزة الإلكترونية الخاصة بالتسجيل البارومتري للناخبين
Post modifier	مضاف إليه	- تنظيم الدولة"يهدد بإعدام مختطفات السويداء"
Apposition	بدل	- لكن ترفض واشنطن دوما فكرة المساومة بشأن القس برونسون. - ويقول منتقدو هذه الخطوة إنها ستحد من حرية التعبير .
Specifier	تمييز	- وأضاف أن قطاع الإسكان يتصدر القطاعات الأكثر احتياجاً للتمويل. - وارتفع اليورو مقابل الدولار 0.3 بالمئة إلى 1.2335 دولار .
Num_complement	--	- وألقي القبض على ثلاثة من أفراد العائلة بتهمة سرقة عملة ذهبية كبيرة.

Confirmation	توكيد	- وتابع المصدر نفسه أن قطاعات التكنولوجيا المتقدمة أو المواد الاستهلاكية سجلت نشاطا قوياً.
Subject	فاعل	- وتابع المصدر نفسه أن قطاعات التكنولوجيا المتقدمة أو المواد الاستهلاكية سجلت نشاطا قوياً.
Subject of passive Verb	نائب فاعل	- وتقام لقاءات الإياب الثلاثاء والأربعاء القادمين.
Object	مفعول به	- وتابع المصدر نفسه أن قطاعات التكنولوجيا المتقدمة أو المواد الاستهلاكية سجلت نشاطا قوياً.
Object 2	مفعول به ثاني	- أمير قطر تميم"يهدي "أردوغان"قصراً طائراً".
Indirect Object	مفعول به غير مباشر	- نجوم يغيبون عن مونديال روسيا
Object of setting	مفعول فيه	- وتقام لقاءات الإياب الثلاثاء والأربعاء القادمين. - إسرائيل تشن هجمات داخل سوريا ردا على"قصف إيراني".
Cognate object	مفعول مطلق	- وأشار أردوغان أيضا إلى أن تركيا تدرس التجارة مع قيرغيزستان بالعملة المحلية أيضا.
Cognate object substitute	نائب المفعول المطلق	- رئيس الوزراء العراقي يعلن النصر رسميا.
Object of purpose	مفعول لأجله	- إسرائيل تشن هجمات داخل سوريا ردا على"قصف إيراني".
Comitative object	مفعول معه	Didn't occur once in the data
Verbal attachement	متعلق بالفعل	- أردوغان يهاجم وزير خارجية الإمارات لنشره تغريدة"مسيئة للعثمانيين"
Verbal Adverbial	الحال	- ما الذي قاله عباس بالضبط ؟ - موريتيو وكونتي"فقدا عقليهما تماما"
Sentence Adverbial	--	- وفي السياق ، يستضيف كافاليرز المباراة الرابعة الجمعة القادم.
Pre Verb Modifier	--	- وقالت كيتاروفيتش للصحفيين"لا أعرف كيف سأصمد حتى الأحد." - وقال"سوف نطبق ضغطا ماليا غير مسبوق على النظام الإيراني."
Tense Modifier	--	- وكان نحو 5400 مسلح ومدني خرجوا من جنوب الغوطة ، الاثنين.
Nhead	--	- أردوغان يهاجم وزير خارجية الإمارات لنشره تغريدة"مسيئة للعثمانيين"
Vhead	--	- أردوغان يهاجم وزير خارجية الإمارات لنشره تغريدة"مسيئة للعثمانيين"
Adjhead	--	- أردوغان يهاجم وزير خارجية الإمارات لنشره تغريدة"مسيئة للعثمانيين"
Phead	--	- أردوغان يهاجم وزير خارجية الإمارات لنشره تغريدة"مسيئة للعثمانيين"
Advhead	--	- وأشار أردوغان أيضا إلى أن تركيا تدرس التجارة مع قيرغيزستان بالعملة المحلية أيضا.
Relhead	--	- ولم يعرف على الفور محتوى البيان الذي سيلقيه المهاجمون .
Part_head	--	- ولم يعرف على الفور محتوى البيان الذي سيلقيه المهاجمون .
P_Complement	--	- وأشار أردوغان أيضا إلى أن تركيا تدرس التجارة مع قيرغيزستان بالعملة المحلية أيضا.
Rel_complement	صلة الموصول	- ولم يعرف على الفور محتوى البيان الذي سيلقيه المهاجمون .
Adv complement	--	- وأشار أردوغان أيضا إلى أن تركيا تدرس التجارة مع قيرغيزستان بالعملة المحلية أيضا.

The final look of any syntactic tree in the current project looks like the one represented in figure (4), which represents sentence (1).

[1] ويقول مدافعون عن حقوق الإنسان إن من بين المتهمين معارضين سلميين.

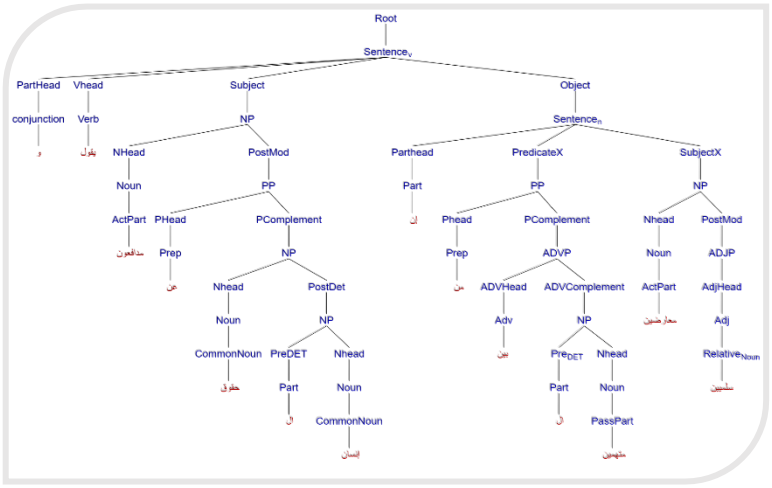

Figure 4: The syntactic tree of sentence (1).

CONCLUSION

The current study presents the guidelines for building a parsed treebank for Modern Standard Arabic. The Study shows the grammar formalism chosen for the treebank representation, the treebank genres and size, and finally the annotation layers adopted in the current study.

REFERENCE

[1] G. Leech, "EAGLES Recommendations for the Syntactic Annotation of Corpora," 1996.

[2] J. Nivre, "Treebanks," Corpus Linguistics: An International Handbook, pp. 225-241, 2007.

[3] M.-F. Moens, "Information Extraction from Blogs," in Handbook of Research on Web Log Analysis, Information Science Reference, 2009, pp. 469-487.

[4] G. Sampson, "Thoughts on Two Decades of Drawing Trees," in Treebanks: Building and Using Parsed Corpora, Dordrecht, Springer Netherlands, 2003, pp. 23-41.

[5] M. Maamouri, A. Bies, T. Buckwalter and W. Mekki, "The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus," NEMLAR Conference on Arabic Language Resources and Tools, pp. 102-109, 2004.

[6] A. Böhmová, J. Hajič, E. Hajičová and B. Hladká, "The Prague Dependency Treebank," in Treebanks: Building and Using Parsed Corpora, Springer Netherlands, 2003, pp. 103-127.

[7] A. Abeille, Treebanks: Building and Using Parsed Corpora, Springer Netherlands, 2003.

[8] F. Xia and M. Palmer, "Comparing and Integrating Tree Adjoining Grammars," in Proceedings of the Fifth International Workshop on Tree Adjoining Grammar and Related Frameworks ({TAG}+5), 2000.

[9] T. Johns, "Should You Be Persuaded - Two Samples Of Data-Driven Learning Materials," 1991.

[10] N. Martin, E. Pujadas, M. Villanueva and M. Guinjoan, "Corpora and New Technologies in the Linguistics Classroom: A Pedagogical Use of a Clause Pattern Database," in Revista Electrónica de Lingüística Aplicada, 2016.

[11] S. Wallis, I. Cushing and B. Aarts, "Exploiting parsed corpora in grammar teaching," in Linguistic Issues in Language Technology, Volume 18, 2019 - Exploiting Parsed Corpora: Applications in Research, Pedagogy, and Processing, 2019.

[12] K. M. Wahba, "Materials Development in Arabic Language Learning and Teaching," in Handbook for Arabic Language Teaching Professionals in the 21st Century, Routledge, 2017.

[13] M. L. Arifianto, "Utilizing the Quranic Arabic Corpus as a Supplementary Teaching and Learning Material for Arabic Syntax:

An Overview of a Web-based Arabic Linguistics Corpus," KnE Social Sciences, p. 403–412, 2021.

[14] A. Farghaly and K. Shaalan, "Arabic Natural Language Processing: Challenges and Solutions," ACM Transactions on Asian Language Information Processing, vol. 8, no. 4, December 2009.

[15] Y.-S. Lee, K. Papineni, S. Roukos, O. Emam and H. Hassan, "Language Model Based Arabic Word Segmentation," in Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Sapporo, 2003.

[16] M. Diab, K. Hacioglu and D. Jurafsky, "Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks," in HLT-NAACL 2004: Short Papers, 2004.

[17] N. Habash and O. Rambow, "Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop," in Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), 2005.

[18] M. Diab, "Second Generation AMIRA Tools for Arabic Processing: Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking," in Proceedings of the Second International Conference, 2009.

[19] A. Pasha, M. Al-Badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow and R. Roth, "MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic," in Proceedings of the Ninth International Conference on Language Resources and Evaluation, Reykjavik, 2014.

[20] 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.

[21] J. Hajic, O. Smrz, T. Buckwalter and H. Jin, "Feature-Based Tagger of Approximations of Functional Arabic Morphology," 2005.

[22] R. Roth, O. Rambow, N. Habash, M. Diab and C. Rudin, "Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking," in Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, 2008.

[23] 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 second International Conference on Arabic Language Resources and Tools (MEDAR), 2009.

[24] M. Diab, "Improved Arabic Base Phrase Chunking with a new enriched POS tag set," in Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Prague, 2007.

[25] K. Darwish, H. Mubarak, A. Abdelali and M. Eldesouki, "Arabic POS Tagging: Don't Abandon Feature Engineering Just Yet," in Proceedings of the Third Arabic Natural Language Processing Workshop, Valencia, 2017.

[26] S. Khalifa, N. Zalmout and N. Habash, "YAMAMA : Yet Another Multi-Dialect Arabic Morphological Analyzer," in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, Oska, 2016.

[27] N. Khoufi, C. Aloulou and L. H. Belguith, "Chunking Arabic texts using Conditional Random Fields," in 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), 2014.

[28] N. Khoufi, A. Chafik and L. Belguith, "ARSYPAR : A tool for parsing the Arabic language," 2013.

[29] N. Khoufi, A. Chafik and L. H. Belguith, "Supervised learning for parsing the Arabic language," 2013.

[30] M. Al-Emran, S. Zaza and K. Shaalan, "Parsing modern standard Arabic using Treebank resources," in 2015 International Conference on Information and Communication Technology Research (ICTRC), 2015.

[31] N. Khoufi, C. Aloulou and L. H. Belguith, "Parsing Arabic using induced probabilistic context free grammar," International Journal of Speech Technology, pp. 313-323, 2016.

[32] R. Maalej, N. Khoufi and C. Aloulou, "Parsing Arabic using deep learning technology," in Proceedings of the Tunisian-Algerian Joint Conference on Applied Computing (TACC 2021), 2021.

[33] "What are the top 200 most spoken languages?," 04 June 2022. [Online]. Available: https://www.ethnologue.com/guides/ethnologue200.

[34] J. Owens, "A House of Sound Structure, of Marvelous form and Proportion: An Introduction," in The Oxford Handbook of Arabic Linguistics, Oxford, Oxford University Press, 2013.

[35] A. Ruby, "Building Syntactic Treebank for Modern Standard Arabic," Egyptian Journal of Language Engineering, pp. 86-101, 2015.

[36] M. Maamouri, A. Bies, S. Kulick, M. Ciul, N. Habash and R. Eskander, "Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development," in Proceedings of the Ninth International Conference on Language Resources and Evaluation ({LREC}'14), 2014.

[37] M. Maamouri, A. Bies, S. Kulick, W. Zaghouani, D. Graff and M. Ciul, "From Speech to Trees: Applying Treebank Annotation to Arabic Broadcast News," in Proceedings of the Seventh International Conference on Language Resources and Evaluation ({LREC}'10), Malta, 2010.

[38] M. Maamouri and A. Bies, "Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools," in Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages, Geneva, 2004.

[39] M. Maamouri, A. Bies and S. Kulick, "Creating a Methodology for Large-Scale Correction of Treebank Annotation: The Case of the Arabic Treebank," pp. 138-144, 2011.

[40] T. Buckwalter, "Buckwalter Arabic Morphological Analyzer Version 1.0," Linguistic Data Consortium, Philadelphia, 2002.

[41] M. Maamouri, A. Bies, S. Krouna, F. Gaddeche and B. Bouziri, "Arabic Treebanking Morphological Analysis & POS Annotation Version 3.8," 2009.

[42] M. Maamouri, A. Bies, T. Buckwalter, H. Jin and W. Mekki, "Arabic Treebank: Part 3 (full corpus) v 2.0 (MPG + Syntactic Analysis) LDC2005T20," Linguistic Data Consortium, Philadelphia, 2005.

[43] D. M. Bikel, "Design of a Multi-Lingual, Parallel-Processing Statistical Parsing Engine," in Proceedings of the Second International Conference on Human Language Technology Research, San Francisco, 2002.

[44] M. Maamouri, A. Bies and S. Kulick, "Enhancing the Arabic Treebank: A Collaborative Effort toward New Annotation Guidelines," in LREC, 2008.

[45] M. Attia, "Arabic Tokenization System," in Proceedings of the 2007 Workshop on Computational Approaches to {S}emitic Languages: Common Issues and Resources, 2007.

[46] J. Hajič, O. Smrz, P. Zemánek, J. Šnaidauf and E. Beška, "Prague Arabic dependency treebank: Development in data and tools," Proceedings of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools, pp. 110-117, 2004.

[47] D. Taji, N. Habash and D. Zeman, "Universal Dependencies for Arabic," 2017.

[48] "PADT For Arabic Manual For Chosen Problems Of Syntactical Analysis Of Arabic," 2002.

[49] R. M. Roth and N. Habash, "CATiB: The Columbia Arabic Treebank," in Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, 2009.

[50] K. Dukes, E. Atwell and A. B. M. Sharaf, "Syntactic annotation guidelines for the quranic Arabic dependency treebank," Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010, pp. 1822-1827, 2010.

[51] A. Dahdah, Dictionary of Arabic Grammar in Charts and Tables, Librairie du Liban, 1985.