I3rab: A New Arabic Dependency Treebank Based on Arabic Grammatical Theory

DANA HALABI, Princess Sumaya University for Technology, Amman, Jordan
EBAA FAYYOUMI, Computer Science Department, Hashemite University, Zarqa, Jordan
ARAFAT AWAJAN, Princess Sumaya University for Technology, Amman, Jordan; Mutah University, Karak, Jordan

Treebanks are valuable linguistic resources that include the syntactic structure of a language sentence in addition to part-of-speech tags and morphological features. They are mainly utilized in modeling statistical parsers. Although the statistical natural language parser has recently become more accurate for languages such as English, those for the Arabic language still have low accuracy. The purpose of this article is to construct a new Arabic dependency treebank based on the traditional Arabic grammatical theory and the characteristics of the Arabic language, to investigate their effects on the accuracy of statistical parsers. The proposed Arabic dependency treebank, called I3rab, contrasts with existing Arabic dependency treebanks in two main concepts. The first concept is the approach of determining the main word of the sentence, and the second concept is the representation of the joined and covert pronouns. To evaluate I3rab, we compared its performance against a subset of Prague Arabic Dependency Treebank that shares a comparable level of details. The conducted experiments show that the percentage improvement reached up to 10.24% in UAS and 18.42% in LAS.

CCS Concepts: • Theory of computation → Grammars and context-free languages;

Additional Key Words and Phrases: Arabic language, dependency treebank, dependency parsing, arabic grammatical theory, nominal sentence, verbal sentence

ACM Reference format:
Dana Halabi, Ebaa Fayyoumi, and Arafat Awajan. 2021. I3rab: A New Arabic Dependency Treebank Based on Arabic Grammatical Theory. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 21, 2, Article 23 (November 2021), 32 pages.
https://doi.org/10.1145/3472295

1 INTRODUCTION

Treebanks are annotated corpora that serve as valuable resources in many data-driven Natural Language Processing (NLP) applications. Typically, sentences in a treebank are annotated with
part of speech (POS), morphological features, and syntactic structure [Frank et al. 2012; Volk et al. 2005]. Treebanks are mainly used for modeling statistical parsers [Kübler et al. 2009]. Furthermore, they are also used in other NLP applications such as question-answering [Li and Xu 2016; Comas, Turmo, and Márquez 2010; Athreya et al. 2021], machine translation [Galley and Manning 2009; Katz-Brown et al. 2011; Renduchintala and Williams 2021], evaluation of machine translation [McCaffery and Nederhof 2016; Yu et al. 2015; Owczarzak et al. 2007], and information retrieval [Gillenwater et al. 2013]. There are many grammatical formalisms to represent the syntax structure. The most commonly used are phrase structure [Chomsky 2002] and dependency structure [Xia and Palmer 2001].

Arabic Language suffers from low performance of parsing. This motivates us to tackle the phenomenon of poor Arabic parsers performance compared to other languages’ parsers. This either refer to the quality of the linguistic resources (i.e., treebanks) used in modeling the statistical parsers, or the approaches and algorithms used in developing the parsers. This article investigates and focuses on the quality of the linguistic resources (i.e., treebanks) to improve the parser performance. Generally speaking, the quality of treebank is affected by two factors: (1) the level and quality of annotation, and (2) the concepts and theories involved in analyzing the structure of sentence then mapping it into either phrase or dependency structure parsers. It is clearly understood that the first factor is within the accepted way due to the deep analysis of sentences [Maamouri et al. 2004; Smrz et al. 2008] and performing the annotation process with manual rechecking [Habash and Roth 2009].

To the best of our knowledge, the existing Modern Standard Arabic (MSA) treebanks were inspired by the characteristics of other languages, mainly English and Czech [Ryding 2005; Maamouri et al. 2004; Smrz et al. 2002]. This is achieved by considering the verb as the main word in the sentence regardless of its position. Consequently, there is a need for a new Arabic treebank constructed according to linguistic and grammatical theories covering the Arabic features, and simultaneously compatible with the concepts and rules of constructing treebanks. The first step to addressing this issue is determining the most appropriate grammatical formalisms that the Arabic linguistic and grammatical theories should coincide with. In this article, we selected the dependency structure on the basis of the substantial attention that has been paid to dependency-structure treebanks in the past two decades. The reason for this decision was the usefulness of bi-lexical relations between individual words (head and modifier words) in solving different ambiguity problems in POS and parsing tasks [Nivre 2005; Kübler et al. 2009]. Moreover, the Arabic language is a Semitic language that is highly inflectional and has rich morphological features [Al-Sughaiyer and Al-Kharashi 2004; Halabi et al. 2017]. In addition, it is considered to have relatively flexible word order [Ryding 2005]. Bharati et al. [Bharati, 1995] has suggested that free word order and rich morphological languages can be handled better by using a dependency-based rather than phrase-structure-based framework.

The Arabic language, like other languages, has several linguistic and grammatical theories dedicated to describing its features and characteristics [Alosh 2005; Owens 1988, 1990]. This article is based on the traditional Arabic grammatical theory called ʾIʿrab. ʾIʿrab maps the Arabic sentence to a set of dependency relations between words. The words in the sentences are classified into either governor or governed word. The basic concept of this theory lies on linking the governor word to the governed word, since the governor word has the power to affect the governed word and to determine its role in the sentence. Consequently, this naturally fits the dependency-based structure rather than phrase-based structure. It is important to highlight another two important concepts are: (1) ʾIʿrab theory categorizes Arabic sentences into two categories—verbal sentence and nominal sentence—depending on the type of the first word in the sentence, (2) ʾIʿrab theory assures that each verb should have a subject, and this subject should follow the verb and cannot
I3rab: A New Arabic Dependency Treebank Based on Arabic Grammatical Theory

precede it considering all forms of the verb’s subject (nominative noun, independent pronoun, joined pronoun or covert pronoun) [Alosh 2005; Owens 1988]. Therefore, the objective of this article is to construct the I3rab treebank, as new pilot dependency treebank for MSA that is based on ‘I’rab theory.

This article is organized as follows. Section 2 presents the literature survey, while the theoretical concepts of the proposed I3rab dependency treebank are covered in Section 3. The proposed I3rab dependency treebank is covered in Section 3. The implementation and experimental results are discussed in Section 4. Finally, the conclusion and future work are described in Section 5.

2 LITERATURE SURVEY

A Treebank is an annotated corpus with grammatical and semantic analysis for sentences beyond the level of POS tags. The sentences are represented as syntactic structure also called syntactic tree [Frank et al. 2012; Volk et al. 2005]. There are many grammatical formalisms that can be used to present a syntactic tree, such as phrase syntax, dependency syntax, lexical functional grammar, tree-adjoining grammar, and others. Most of these structures are applied to English [Lynn 2016]. Each formalism demonstrates a different way to present the syntactic tree. The most commonly used formalisms are the phrase structure and dependency structure [Lynn 2016].

The first constructed treebank was a phrase structure treebank for English. It was the first large-scale syntactic annotation treebank developed and distributed by the Linguistic Data Consortium (LDC) [Marcus, Santorini, and Marcinkiewicz 2006]. It has primarily been used to develop the statistical parser model for the English language [Collins 2003]. The success of using treebanks in the English language inspired researchers to follow the same approach to develop phrase-structure treebanks for various languages such as Catalan [Civit, Bufí, and Valverde 2004], Spanish [Civit and Martí 2004], Chinese [Xue et al. 2013], Arabic [Maamouri et al. 2004], Hebrew [Sima’an et al. 2001], and Korean [Han et al. 2002].

The first dependency treebank was the Prague Dependency Treebank (PDT) [Böhmová et al. 2003] developed for the Czech language. The dependency syntax of the PDT was strongly influenced by the functional generative description theory, which considers the verb as the main word in the sentence, regardless of its position. In addition to the Czech language, a dependency treebank approach has been adopted for other languages, such as Arabic, Basque, Catalan, Chinese, Czech, English, Greek, Hungarian, Italian, and Turkish. More attention has been given to dependency treebanks through the Conference on Natural Language Learning (CoNLL) 2007 [Nivre et al. 2007a], dedicated to dependency parsing.

Arabic is a low-resource language [Fayyoumi and Idwan, 2021], however, there are a set of treebanks have been developed for Arabic. The most important treebanks are the Penn Arabic Treebank (PAT) [Maamouri et al. 2009, 2004], Prague Arabic Dependency Treebank (PADT) [Smrz, Bielicky, and Hajic 2008], Columbia Arabic Treebanks (CATiB) [Habash and Roth 2009], Classical Arabic Treebank [Dukes and Buckwalter 2010], and U.S. Army Research Laboratory Arabic Dependency Treebank (AADT) [Tratz 2016].

PAT [Maamouri et al. 2009, 2004] was the first treebank developed for Modern Standard Arabic (MSA). It is a phrase structure treebank based on the Penn English Treebank [Marcus, Santorini, and Marcinkiewicz 2006], and it uses the same tagset of the Penn English Treebank to annotate the phrase structure of PAT. Initially, it used more than 400 POS tags, but the number was reduced to 36 POS tags during training and testing parsers [Kulick et al. 2006]. PADT [Smrz et al. 2008] is a dependency treebank for the same text sources in PAT.

1I3rab is derived from the English transliteration (I’rab) of the Arabic word (اةرابة). In general, Arabic speakers replace the Arabic letter (خ) with the number (3) in the Arabization text that is usually used in chats and WhatsApp conversations.
The dependency syntax of PADT is strongly influenced by the PDT. Smrz et al. [2008] have argued that the Arabic and Czech languages are rich in inflection and share the free words order property. The verb is considered the main word in the sentence regardless of its position in the sentence. However, the PADT team manually treated the features of the Arabic language that could not be handled in the same way as the Czech language.

Another dependency treebank for MSA is CATiB [Habash and Roth 2009]. The dependency labels of CATiB were inspired by traditional Arabic grammatical theory, but this treebank uses a small subset of the full traditional syntactic roles. It has only six POS tags and eight relation labels for dependency links. Parsers trained against this treebank can accelerate the development of new treebanks but with limitations in linguistic richness. Although the dependency labels of the CATiB treebank were based on traditional Arabic grammar, it was inspired by PDT in considering the verb as the main word in the sentence regardless of its position in the sentence.

The Classical Arabic Treebank [Dukes and Buckwalter 2010] is an annotated corpus specialized for the text of the Holy Quran. The Holy Quran is a major religious text that is considered to contain unique and challenging language. The Holy Quran is a collection of 114 ordered chapters (سورة, suar), each with a number of ordered verses (آية, ayat). The syntactic tree for a verse is represented in a hybrid dependency-constituency phrase structure model. The syntactic tree depends on the traditional Arabic grammar exposed in the well-known book (إعجاز القرآن الكريم, ‘irabu alquranialkarim). The Dukes and Habash [2011] and Dukes [2015] papers argued that this hybrid representation is sufficiently flexible to represent all aspects of the syntax in the Holy Quran.

The newest dependency treebank for MSA is AADT [Tratz 2016]. It was derived from existing Arabic treebanks distributed by LDC, by using constituent-to-dependency conversion tools. The dependency scheme consists of a total of 35 labels. Many of these are similar to those of Stanford’s basic dependency scheme for English. Similarly, to CATiB, this treebank is based on the idea that the verb is the main word in the sentence.

The Arabic statistical parsers have been developed on the basis of phrase structure, e.g., PAT [Kulick et al. 2006], and dependency-structure, e.g., PADT [Nivre et al. 2007a; Smrz et al. 2008]. Kulick et al. [2006] reported that the accuracy of Arabic parsing obtained by using PAT version 1 to train the Bikel parser [Bikel 2004] had an F1-score of 74%, which was considered a low score relative to the 88% F1-score for English on a comparably sized corpus. CoNLL shared task 2007 was devoted to dependency parsing. It involved developing and evaluating different state-of-the-art parsers for ten languages. The dependency parsers generated for the Arabic language were based on the PADT [Nivre et al. 2007a; Smrz et al. 2008]. The best performance for the dependency parsing of Arabic was found to be 76.52% for the Labeled Attachment Score (LAS). The LAS values were separated into three classes: low, medium and high. The Arabic language received a low score, whereas the Czech language received a medium score of 80.2%, and the English language received a high score of 89.6% [Nivre et al. 2007a].

3 THEORETICAL CONCEPTS OF THE PROPOSED I3RAB DEPENDENCY TREEBANK

‘I’rab is derived from a general well-known Arabic traditional theory called “The Theory of Al Aamil, النازريعة المأملي”. It appeared in the second Hijreecentury (approximately 900 A.D.). The main aim was to aid foreigners in learning the Arabic language in addition to studying and understanding the Holy Quran. The Theory of Al Aamil consists of three main components: governance “العمل، al-‘amalu,” governor “العامل، al-‘amilu” and governed “العملا، al-ma’mulu.”

In this section, the main features of the Arabic language related to ‘I’rab will be covered and followed by the main essential principles of ‘I’rab. Then the main concepts of ‘I’rab and its relationship to dependency grammar will be explained.
3.1 The Sentence in ‘I’râb

‘I’râb considers the sentence as the basic unit of analysis [Alosh 2005; Owens 1988]. Sentences are categorized into two types, verbal and nominal sentences, depending on the class of the first word in the sentence.

A verbal sentence is one starting with a strong verb. For example, in the sentence “يا نام الطفّل، يانام١ّ، الاطلن، يانام١ّ، الطفّل، يانام١ّ، الطفّل، the child sleeps,” the verb is “يا نام، يانام١ّ، eats,” and the agent is “يا نام، يانام١ّ، the child.” A nominal sentence is one starting with a noun. For example, in the sentence “السما، الشمس شرتة، aš-šamsu mušriqatun, the sun is shiny,” the topic is “السما، الشمس، the sun,” and the predicate is “السما، الشمس شرتة، shiny.” In another example, in the sentence “السما، الشيط، muḥammadun yaqra’u al-kitāba, Mohammed reads the book,” the phrase “السما، الشيط، yaqra’u al-kitāba, reads the book” is an entire verbal sentence that has the role of a predicate.

3.2 The ‘I’râb Theory

In Arabic language, the basic unit of analysis is the sentence [Owens 1988], which consists of meaningful words. Each word has a specific role in the sentence determined by syntactical rules, which is used to build a coherent structure [Owens 1988].

3.2.1 The Main Concepts. ‘I’râb considers an Arabic sentence is made up of three main components: governance, governor, and governed. The governor word is linked to the governed word on the basis of the principle that the governor word has the power to affect the governed word in some manner, and to determine its role in the sentence [Alosh 2005; Owens 1988]. This theory maps the Arabic sentence as a set of dependency relations between words. The reason for this is that the governor is associated with the head in dependency grammar, and the governed is associated with the modifier [Alosh 2005; Owens 1988]. In other words, the governor governs (do, operate) the governed and can change its case (for nouns) or mood (for verbs) according to its function. The noun has three cases: nominative, accusative and genitive. For the verb, only imperfect verbs show mood inflection. These mood inflections are indicative, subjunctive and jussive. Linguistically, applying the process of this theory on Arabic sentence is known as (‘irab) [Alosh 2005; Owens 1988]. For example, in the verbal sentence “يا كلو ار-راغل اس-سمك، the man eats the fish,” the verb “يا كلو، ya’kulu, eats” is the main word in the sentence and plays a predominant role in the sentence governing the other two words, “السما، as-samaka, the fish” as an agent and “السما، as-samaka, the fish” as an object [Owens 1988]. Consequently, both the agent and the object are governed by the main word (verb). Table 1 shows the details of the grammatical analysis of the sentence “يا كلو ار-راغل اس-سمك، the man eats the fish.” Below, we will use the term i’rab starting with a small “i” to indicate the grammatical analysis of a sentence.

Preceding the above sentence with a jussive particle “لم، lam, not” that indicates negation would change the mood of the imperfect verb from the inductive to jussive mode. Table 2 shows the ‘I’râb of the sentence “لم، lam ya’kul ar-ragulu as-samaka, the man did not eat the fish.” The existence of the jussive particle makes it the main word that governs the imperfect verb “يا كلو، ya’kulu,

Table 1. The i’râb of the Sentence “يا كلو ار-راغل اس-سمك، the Man Eats the Fish”

| Word | Word role (Arabic) | Word role (English) |
|------|-------------------|---------------------|
| يانام | Inductive imperfect verb | يانام | Inductive imperfect verb |
| الاطلن | Nominative noun in the role of the agent of the verb | الاطلن | Nominative noun in the role of the agent of the verb |
| مفعول به منصب | Successive noun in the role of the object of the verb | مفعول به منصب | Successive noun in the role of the object of the verb |

ACM Trans. Asian Low-Resour. Lang. Inf. Process., Vol. 21, No. 2, Article 23. Publication date: November 2021.
Table 2. The i’rab of the Sentence “لم يأكل الرجل السمك, the Man Did not Eat the Fish”

| Word | Word role (Arabic) | Word role (English) |
|------|--------------------|---------------------|
| لم   | حرف جزم            | Jussive particle    |
| يأكل | فعل مضارع مجزوم   | Jussive imperfect verb |
| الرجل | مؤنث مرفوع          | Nominative noun in the role of the agent of the verb “يأكل” |
| السمك | مفعول به منصوب   | Successive noun in the role of the object of the verb “يأكل” |

Table 3. The i’rab of the Sentence “النسم مشروقة, the Sun is Shiny”

| Word | Word role (Arabic) | Word role (English) |
|------|--------------------|---------------------|
| الشمس | مبتدأ مرفوع          | Nominative noun in the role of the topic “النسم” |
| مشروقة | خبر المبتدأ مرفوع | Nominative noun in the role of the predicate of the topic “النسم” |

Table 4. The i’rab of the Sentence “كانت الشمس مشروقة, The Sun Was Shining”

| Word | Word role (Arabic) | Word role (English) |
|------|--------------------|---------------------|
| كانت | فعل مضام ناقص      | Defective perfect verb |
| الشمس | اسم مبتدأ مرفوع     | Nominative noun in the role of the topic “كانت” |
| مشروقة | خبر كان منصوب | Accusative noun in the role of the predicate of “كانت” |

y’a’kul, eats.” There is no change in the roles of the other two words “الرجل, ar-ragulu, the man” and “السمك, as-samaka, the fish.”

Thus, an imperfect verb is governed by the jussive particle, whereas the agent and the object are still governed by the imperfect verb. Figure 1 shows parsing trees related to two verbal sentences, respectively. The (S) symbol in the tree represents the root of the tree, the main word in the sentence is the unique child of S, and the relations between each parent and its children are labeled by the role (function) of the child word in the sentence.

In the case of the two nominal sentences, the sentence “النسم مشروقة, as-samsu musriqatun, the sun is shiny” and “كانت الشمس مشروقة, kānat as-samsu musriqatan, the sun was shining,” both sentences are similar, but the second sentence starts with the defective verb “كانت, kānat, was.” The i’rab of the two sentences are shown in Tables 3 and 4, respectively.

In the first sentence, the topic “النسم, as-samsu, the sun” is considered the main word that governs the predicate “مشروقة, musriqatun, shiny” [Alotaiby et al. 2010]. In the second sentence, the abolisher “كانت, kānat, was” becomes the main word in the sentence and governs both the topic and the predicate. The tree representations of the above two nominal sentences are illustrated in Figure 2.

3.2.2 The Scope of the Governor. The type of sentence (nominal or verbal) enables understanding of its structure and consequently determines the main word in the sentence, the governor(s), and the governed word(s). Each governor has a scope specifying the words or phrases in the sentences that are affected by the governor. There is no correlation between the simplicity of the sentence and the number of governors. Simple sentences do not necessarily consist of a single
governor, and they may have multiple governors. For example, the sentence ".المساء، يأكل الرجل السمك، the man eats the fish” has one governor “أكل، ya’kulu, eats” but two governed words, “رجل، ar-ragulu, the man” and “سمك، as-samaka, the fish,” for the same governor. The sentence ".المساء، the bird is in the cage” is a nominal sentence with the predicate as an adverbial element “في، fi, in” and one governed “الفص.اف، al-qafasi, the cage,” which is the main word in the sentence [Owens 1988]. This sentence has two governors and two governed words.

3.2.3 Constraints of ʿIrab. This theory is based on the following set of constraints:
(a) The governed word has one and only one governor.
(b) The cardinality of the relation between the governor and the governed is one-to-many. This means that the governor has at least one governed word.
(c) The governors can be verbs, nouns or particles, but there is a set of rules used to determine the governor.
   (i) All verbs are governors, regardless of their tense and type. Each verb should have a single agent that follows the verb.
(ii) Most particles are governors. Particles can be divided into three categories as follows:

1. Particles occurring only with nouns are governors, such as the preposition particles "/charfa/charcd/char40\, ila, to."

2. Particles occurring only with verbs are governors, such as the particles "/char0d\, lam, did not" and "/char0d\, lan, will not," which occur only with the imperfect verb and govern it in the jussive and subjunctive, respectively.

3. Particles occurring with both, such as question particles "/charc9/chareb\, hal, whether" and conjunction particles, "/charf0\, wa, and" are not governors.

(iii) Nouns depend on its role in the sentence. For example, in the sentence "المسور في اللفظ، al-ʿuşfurū fi al-qafaši, the bird is in the cage," the word "المسور، al-ʿuşfurū, the bird" is a noun, and it is a topic, so it is a governor. The word "الفظ، al-qafaši, the cage" is a noun, but it is an object of the prepositional "في, fi, in"; therefore, it is a governed word.

3.3 The Suitability of ʿIrab theory for Developing Dependency Structure for Arabic Text

Mapping ʿIrab theory to dependency grammar is a straightforward task. An Arabic sentence can be represented as a tree of dependency relations between words, because a governor can be associated with the head in dependency grammar, and the governed can be associated with a modifier.

ʿIrab and dependency grammar have several similarities, as follows:

1. The sentence has one and only one independent word that acts as a ROOT, and the main word is a child of the ROOT node.
2. All the items are in dependency relations except the ROOT.
3. A governed word (child) has one and only one governor (parent).
4. A governor has at least one governed word.
5. A dependency relation is a unidirectional relation.

There are two major differences between ʿIrab and dependency grammar. The first difference concerns the determination the main word in the sentence. The dependency grammar always considers the verb as the main word, whereas the main word in ʿIrab depends mostly on the first word in the sentence. Second, the covert element(s) in the syntax representation in ʿIrab should be clearly deduced and depicted in the tree. In contrast, dependency grammar maps only the lexical elements in the sentence.

The previous discussion clearly shows that although there are differences between Arabic grammatical theory and dependency grammar, ʿIrab coincides with the concept of dependency in the dependency structure.

4 THE PROPOSED I3RAB DEPENDENCY TREEBANK

The process of building the proposed new dependency treebank passed five stages: (1) defining the mechanism of the tokenization process, (2) choosing the POS tagset, (3) describing the morphological analysis, (4) determining the I3rab dependency schema, and (5) describing the format of the dependency treebank.

4.1 The Tokenization Process

Tokenization is the task of dividing a sentence into tokens. The token is the smallest unit of syntax [Attia 2007]. In the Arabic language, the word is structured from concatenative morphemes that include stems, affixes and clitics [Attia and Somers 2008, Awajan 2007]. These morphemes appear
Table 5. The Part-of-speech Category and Their Meanings [LDC 2007]

| Part-of-speech category | Description                              |
|-------------------------|------------------------------------------|
| VI VP VC                | imperfect, perfect, and imperative verb forms |
| N- A- D-                | nouns, adjectives, and adverbs          |
| C- P- I-                | conjunctions, prepositions, interjections |
| G- Q- Y-                | graphical symbols, numbers, abbreviations |
| F- FN FI                | particles, especially negative and interrogative |
| S- SD SR                | pronouns, especially demonstrative and relative |
| –                       | isolated definite articles              |
| Z-                      | proper names                            |

Table 6. The Morphological Information and Its Meaning [LDC 2007]

| Morphological feature | Description |
|-----------------------|-------------|
| Mood                  | Indicative, Subjunctive, or Jussive of imperfect verbs, with D if undecided between S and J |
| Voice                 | Active or Passive |
| Person                | 1 speaker, 2 addressee, 3 others |
| Gender                | morphologically overt “gender,” Masculine or Feminine |
| Number                | morphologically overt “number,” Singular, Dual, or Plural |
| Case                  | 1 nominative, 2 genitive, 4 accusative |
| Definiteness          | morphological “definiteness,” Indefinite, Definite, Reduced, or Complex |

sequentially in the word structure as follows [Awajan 2015]:

\[
[\text{proclitic (s)}] + [\text{prefix (es)}] + \text{stem} + [\text{suffix (es)}] + [\text{enclitic}].
\]

The I3rab tokenization process keeps the suffix attached to the word and detaches the clitics. In general, the I3rab tokenization process detaches the following clitics from the word: question particles “؟, َة, whether,” conjunction particles “،, و, and” and “ْ،, fa, so/then,” attached preposition particles “ب, bi, by” and future particles “َلا, sa, will.” In addition, all joined pronouns are considered clitics and should be detached from the words. Furthermore, the covert pronouns should be surmised and explicitly represented as individual tokens.

4.2 POS Tagset

I3rab uses the same POS tagset used by PADT [Hajic et al. 2004]. There are 20 tags, as listed in Table 5. The POS tags are for verbs, nouns, adjectives, adverbs, prepositions, and proper nouns. In contrast to PADT, I3rab tags all pronouns, including covert pronouns, with the (S-) POS tag. Further examples of these tags can be found in Appendix A.

4.3 Morphological Analysis

I3rab uses the same approach as PADT, which is based on morphological information generated by MorphoTrees [Smrz and Pajas 2004] and the Lemmas and Glosses based on the Buckwalter lexicon [LDC 2004a]. The most important morphological information is the mood, voice, person, gender, number, case, and definiteness. The complete list of morphological information is shown in Table 6. As mentioned above, the tokens of all joined and covert pronouns should be processed...
### 4.4 I3rab Dependency Schema

The current I3rab dependency schema has 34 dependency relation labels listed in Table 7. All these labels are selected and derived from I‘rab theory. The process is performed by working on a set of Arabic sentences and performing their grammatical analysis through application of the I‘rab

| #  | Dependency Relation | Description (English) | Description (Arabic) | #  | Dependency Relation | Description (English) | Description (Arabic) |
|----|---------------------|-----------------------|----------------------|----|---------------------|-----------------------|----------------------|
| 1  | ADJ                 | Adjective             | صفعة                 | 18 | PRED-ADVP           | Predicate-Adverbial phrase | خبر - شبه جملة - طرفية |
| 2  | ADVP                | Adverb                | ضرف زمان/مكان        | 19 | PRED-NOUN           | Predicate-Nominaive noun | خبر - مفرد |
| 3  | AGENT               | Agent                 | فاعل                  | 20 | PRED-NP             | Predicate-Nominaive phrase | جملة - اسمية |
| 4  | ALTER               | Alternate             | بال                  | 21 | PRED-PP             | Predicate-Prepositional phrase | خبر - جملة - مفرد ومحور |
| 5  | COMMA               | Comma (punctuation)  | ترقيم - فصلية       | 22 | PRED-VP             | Predicate-Verbal phrase | جملة - فعلية |
| 6  | COND                | Condition             | شرط                  | 23 | PREDX-ADVP          | Predicate of P-ACC or VBX-Adverbial phrase | خبر كان/إن - شبه جملة - طرفية |
| 7  | COORD               | Coordinating particle | أداة ربط              | 24 | PREDX-NOUN          | Predicate of P-ACC or VBX-Nominaive noun | خبر - مفرد - جملة اسمية |
| 8  | END                 | End (punctuation)    | ترقيم - نقطة         | 25 | PREDX-NP            | Predicate of P-ACC or VBX-Nominaive phrase | خبر كان/إن - جملة اسمية |
| 9  | EXCEPT              | Exception             | استثناء              | 26 | PREDX-PP            | Predicate of P-ACC or VBX-Prepositional phrase | خبر كان/إن - شبه جملة جار ومجرور |
| 10 | GEN                 | Genitive              | مجرور                | 27 | PREDX-VP            | Predicate of P-ACC or VBX-Verbal phrase | خبر كان/إن - جملة فعلية |
| 11 | HAAL                | Adverb of manner      | حال                  | 28 | PUNCT               | Punctuation            | ترقيم |
| 12 | MA3TOUF             | The coordinate modifier | معطوف             | 29 | TAMYEEZ             | The specifier         | تمييز |
| 13 | NEG (Negation)      | Negation particle     | حرف نفي             | 30 | TAWKEED             | Emphasis              | توكيد |
| 14 | OBJ                 | Object                | معول به (الفاعل)     | 31 | TOPIC               | Topic                  | مبدأ |
| 15 | P                   | Particle              | حرف                  | 32 | TOPICX              | Topic of P-ACC or VBX | اسم كان/إن |
| 16 | P-ACC               | Accusative particle   | حرف نصب             | 33 | VB (Verb)           | Verb (strong)          | فعل نام |
| 17 | PART                | Part particle         | حرف تابع             | 34 | VBX (copula)        | Defective verb (copula) | فعل نابض |

by the morphological analyzer to generate their morphological features. Further examples of this morphological information can be found in Appendix A.
Fig. 3. The sentence “وصول وزير الخارجية الأمريكي إلى بيروت,” U.S. Secretary of State arrives in Beirut,” represented in CoNLL-X format.

process by linguistic experts² [Halabi et al 2020, 2021]. All dependency relations are extracted from the grammatical analysis. Further examples related to dependency schema can be found in Appendix B.

4.5 Dependency Treebank format

The I3rab dependency treebank will be presented in the CoNLL-X format in a tab separated file in which sentences are separated by a blank line [Buchholz and Marsi 2006]. Each sentence has one or more tokens. Each token has ten attributes separated by a tab. The ten attributes are ID (token index in the sentence), FORM (surface form of token as appears in the sentence), LEMMA (typically the lemma of the token), CPOSTAG (coarse POS), POSTAG (grain POS), FEATS (set of optional morphological features), HEAD (ID of the head of the token), DEPREL (dependency label between the HEAD and the token), PHEAD (projective head of the current token), and PDEPREL (dependency relation to the PHEAD). The columns ID, Form, HEAD, and DEPREL are mandatory, and the others are optional. For example, the sentence “وصول وزير الخارجية الأمريكي إلى بيروت,” wasulu waziri al-hariqiyati al-ʿamriki ʿilâ bayrunû, U.S. Secretary of State arrives in Beirut” is represented in CoNLL-X format, as shown in Figure 3.

The dependency labels in the example were taken from the dependency labels for the I3rab Dependency Treebank.

5 IMPLEMENTATION

To validate our approach and demonstrate the quality of the predicted grammatical structure [Solberg et al. 2014], we evaluated the I3rab treebank against the PADT treebank in a dependency parsing task. The evaluation was limited to only the PADT treebank, because the PADT [Hajic et al. 2004] treebank was involved in the CoNLL shared task 2007 [Nivre et al. 2007a; LDC 2007]. In addition, the PADT treebank is available for free download from the LDC site under the LDC user agreement for non-members [LDC 2018], whereas the other MSA dependency treebanks are not available for free download.

The validation process included four important components: datasets, a parser generator, accuracy metrics, and a parser evaluation tool.

(1) Datasets. Two datasets were used in the evaluation throughout the experiments: the part-PADT dataset and the I3rab dataset. The part-PADT dataset is a subset of PADT [Hajic et al. 2004]. The PADT was mainly collected from six news agencies [Smrz, Snaidauf, and Zemánek 2002; LDC 2004b, 2007]. For the CoNLL shared task 2007, the available PADT dataset included 3,043 sentences with a total of 116,800 tokens [LDC 2007] and was represented in CoNLL-X format. The original

²Two linguistic experts were involved in this work. One has a PhD in Arabic language, and the other is in the 2nd year of a Master’s program in Arabic language.
PADT dataset was divided into two sets: the training dataset contained 2912 sentences representing 95.7% of the whole dataset (3,043 sentences) and the testing dataset contained 131 sentences representing 4.3% of the whole dataset [LDC 2018]. The morphological features of tokens had been annotated with MorphoTrees [Smrz and Pajas 2004; LDC 2007], and Lemmas and Glosses were generated according to the Buckwalter lexicon [LDC 2007, 2004a].

The part-PADT dataset contains 1,131 sentences. It was sampled from the original PADT and was divided into two parts: a training dataset including 1,000 sentences and a testing dataset containing 131 sentences. The training dataset of part-PADT was sampled from the PADT training dataset on the basis of sentence length. While, the testing dataset of part-PADT had the whole sentences of PADT testing dataset. Below, we use the term PADT instead of part-PADT. Figure 4 shows the distribution of the sentence lengths for the PADT training dataset.

The I3rab dataset was constructed by using the same 1,131 sentences from the PADT dataset. Each sentence was tokenized, POS tagged, morphological analyzed, and finally, the dependency structure was constructed. We applied the I3rab approach in each step. Initially the I3rab dataset had the same tokenization, the same POS tags, and the same morphological features as the PADT dataset. The fields related to dependency relations were reinitialized to prepare them for I3rab approach. The tokenization of the sentences was revised and modified as required according to the tokenization approach in I3rab, particularly for verbs. In the tokenization process, the new tokens were generated through perform the following steps:

1. The first step was based on the theory utilized by I3rab. Each verb should have an agent, and this agent should come after the verb. This agent should be considered a separate token. According to I3rab, the agent might be a single nominative noun, independent pronoun, joined pronoun, or covert pronoun.
   a. The single nominative noun and independent pronoun: I3rab is consistent with PADT in considering them as separate tokens from the verb. For example, in the sentence “وَقَّعَ التَّفْجِيرُ فِي قَاعَ أَحَدَ النَّاَقِمِينَ جَعَلَ النَّاَجِمَ النَّاجِمًا” (التفجير), the noun

Fig. 4. The distribution of sentence length for the PADT training dataset.
I3rab: A New Arabic Dependency Treebank Based on Arabic Grammatical Theory

al-i-'nfiğaru, the explosion” is the agent of the verb “vue, waqa’a, a, occurred” and is already considered a separate token according to the white-space segmentation method.

(b) The joined pronoun: I3rab behaves differently from the PADT in considering the joined pronoun as a separate token. For example, the word “jumūman, yağtami’āni, they (dual) meet” was segmented into two separate tokens: the verb “jumūman, yağtami’a, meets” and the joined nominative pronoun “ān, ani, they (dual)” that acts as the agent to the verb “jumūman, yağtami’a, meets.” Another example is the word “bendax, yadḫulūna, they enter,” which was segmented into two separate tokens: the verb “bendax, yadḫulū, enters” and the joined nominative pronoun “ān, unna, they (plural)” that acts as an agent to the verb “bendax, yadḫulū, enters.”

(c) The covert pronoun: I3rab behaves differently from the PADT by allowing the agent to be surmised, despite its not being explicitly stated in the sentence. A new token was generated and added immediately after the verb. For example, in the sentence “ān liyan lubnāna yadinu al-irḥāba Lebanon condemns terrorism,” the verb “yadinu, condemn” has no explicit agent. The word “ān, al-irḥāba, terrorism” is a direct object of the verb, whereas the covert pronoun is considered the agent of the verb. This agent is surmised as “huwa, he,” which refers to the noun “liyan, Lebanon.”

(2) The second step was related to the tokenization process of PADT, which tokenizes some words such as “hashaba, according to” as one token, whereas I3rab tokenizes them into two separate tokens: the words “hashaba, according to” and “ma, ma, which.” This segmentation is based on the i’rab of sentences.

(3) The third step was related to the presence of errors in the tokenization process in the PADT. For example, the word “al-sars, bias-sârs” should be segmented into two separate tokens: the preposition particle “bi, with” and the genitive proper name “al-sars, as-sârs, SARS.” In another example, the word “Huatian, maHuatian” should be segmented into two separate tokens: the coordinative particle “wa, and” and the noun “Huatian.” This process complies with the PADT philosophy. Another error is related to the definite article in Arabic “al, the.” It is well known to consider the definite tool “al, al, the” as a morphological feature of the noun and not to tokenize it as a separate token. The last error is related to the presence of redundancy for some words in the text, although this redundancy did not exist in the original text. In I3rab, we deleted the redundant words.

Table 8 shows descriptive statistics of tokens throughout the tokenization process in PADT and I3rab. The number of tokens in the I3rab dataset exceeded the number of tokens in the PADT.
dataset by 1,031, owing to the previously described tokenization process used to construct the I3rab dataset.

After revising the tokenization process, for each newly created or modified token, we assigned the morphological features to them on the basis of MorphoTrees. For the verbs with a joined nominative pronoun, the number feature of the verb was changed into singular, and the number feature of the pronoun was either dual or plural, depending on the joined pronoun.

After the tokenization was completed and the morphological features for the tokens were reassigned as required, the dependency structure for each sentence was constructed [Halabi et al. 2020, 2021]. The syntactic analysis process starts by prefixing the tokens of sentence with an artificial node called ROOT. Next, we created the Pair_Governor_Governed_List, which consists of the order pair (governor, governed) in the sentence. This is achieved by first finding the modifiers of ROOT nodes or in other words determining the main word(s) of the sentence. Then for each ROOT’s modifier, find its modifiers until ends all words in the sentence. The formal description on the whole process is illustrated in Algorithm: I3rab_Dependency_Structure. As mentioned above, there are two main differences between the I3rab approach and PADT approach. The first is determining the main word of the sentence. The second is the explicit representation of all pronoun types (independent, joined and covert).

For example, the sentence "وظفو اليونيسف يدانون العودة إلى بغداد" Sentence in PADT and I3rab was separated into two tokens: the first token was the verb "يبدأون, yabda ′una, are starting" and the second token was the joined nominative pronoun "يبدأون, yabda ′una, are starting" in the I3rab dataset separated into two tokens: the first token was the verb "يبدأون, yabda ′una, are starting" and the second token was the joined nominative pronoun "يبدأون, yabda ′una, are starting" in the I3rab dataset. Moreover, according to the PADT approach, this sentence was considered a verbal sentence, and the main word of the sentence (modifier of ROOT) was the verb "يبدأون, yabda ′una, are starting," where the ID of head was zero. In contrast, according to I3rab, this sentence was considered a nominal sentence, and because it was not preceded by any of the abolishers, the main word of the sentence was the topic "وظفو اليونيسف, staff". The CoNLL-X format of the sentence in PADT and I3rab is presented in Figures 5. The dependency trees for the sentence in PADT and I3rab are shown in Figures 6.

(2) Parser generator. The MaltParser parser generator was used in our experiments to evaluate the quality of the predicted dependency structure obtained from the I3rab treebank. MaltParser is a state-of-the-art independent-language dependency parser generator. It is a shift-reduce transition-based dependency parser [Nivre et al. 2006; Nivre et al. 2007b]. In this article, we used the freely available MaltParser version 1.9.2 [Nivre 2018].

(3) Accuracy metrics. There is a set of metrics used to measure the quality of a dependency parser. The two most important metrics are the Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS). UAS (head right) is the percentage of tokens that correctly linked to its head and LAS (both right) is the percentage of tokens that correctly linked to its head with right dependency relation comparing gold test data.

(4) The Parser evaluation tool. In this article, we used the free available evaluation tool MaltEval [Nilsson and Nivre 2008]. It provides quantitative evaluation for the accuracy metrics (UAS, LAS) for the predicted dependency trees. The version used in this article was released on 05/10/2014 [Hall et al 2013].
Fig. 5. The CoNLL-X format of the sentence "موظفو اليونيسيف يبدأون العودة إلى بغداد،" UNICEF staff are starting to return to Baghdad.

Fig. 6. The dependency tree of the sentence "موظفو اليونيسيف يبدأون العودة إلى بغداد،" UNICEF staff are starting to return to Baghdad.

6 EXPERIMENTAL RESULTS AND DISCUSSION

The evaluation process involved three main steps: training, testing, and computation of evaluation metrics. In the first step, training, the MaltParser was elaborated (in learn mode) to produce two trained parser models: one involving the PADT training dataset and the other involving the I3rab dataset. Of note, the PADT dataset was used as is without any modification, and then the PADT parser was considered as the base parser. The second step was the testing step, in which the MaltParser was elaborated (in parse mode) with the trained parser model against a blind dataset; a set of predicted grammatical structures was produced for each sentence in the testing datasets of PADT and I3rab. The blind testing dataset included the sentences for testing in CONLL-X format.

3In all dependency structure figures, the words are annotated with coarse POS tags. These tags are (N: noun), (V: verb), (S: pronoun), (P: particle) and (A: adjective).
Table 10. The Percentage of UAS and LAS for Experiments for PADT and I3rab

|                | PADT | I3RAB | The percentage of improvement |
|----------------|------|-------|------------------------------|
| UAS (Mean)     | 78.1 | 86.1  | 10.24                        |
| LAS (Mean)     | 67.3 | 79.7  | 18.42                        |
| Row count      | 5124 | 5379  | ——                           |

Algorithm 1: I3rab_Dependency_Structure(S)

Input: S = w₁ w₂ ... wₙ: Arabic sentence.
Output: DSA: Dependency structure of Arabic sentence in CoNLL-X format.

Note:
- Pair_Governor_Governed_List: list contains all pairs of (governor, governed) in the Arabic sentence S.
- ROOT_Modifiers: list contains the main word(s) of sentence S.
- All_Modifiers: list contains the modifiers of other word in the sentence (not w₀)

Method:
1. S = I3rab_Tokenize(S)
2. S = I3rab_POS_Tagging(S)
3. S = I3rab_Morphological_Analysis(S)
4. Pair_Governor_Governed_List = NULL
5. Prefix S with an artificial node called ROOT w₀ to be S = w₀ w₁ w₂ & wₙ
6. Find all modifiers for w₀ and add them to ROOT_Modifiers
7. For each elementx in ROOT_Modifiers Do
   8. Add (w₀, elementx) to Pair_Governor_Governed_List
   9. Find all modifiers for the elementx and add them to All_Modifiers.
10. For each elementy in All_Modifiers Do
    11. Add (elementx, elementy) to Pair_Governor_Governed_List
    12. End For
13. All_Modifiers = NULL
14. End For
15. DSA = Link_Pairs (Pair_Governor_Governed_List)
16. Return DSA
17. END Algorithm I3rab_Dependency_Structure(S)

The produced annotation from I3rab parser has been checked by a linguistic expert in Arabic language, to validate the notation.

The results clearly indicated that the highest UAS and LAS scores for dependency parsing for Arabic were achieved by using the I3rab dataset. The UAS reached 86.1%, and the LAS reached 79.7%. The percentage improvement achieved with the I3rab strategy against PADT was 10.24%

---

4The produced annotation from I3rab parser has been checked by a linguistic expert in Arabic language, to validate the notation.
and 18.42% for UAS and LAS, respectively. Moreover, the differences in UAS and LAS between two approaches were extremely statistically significant ($p < 0.0001$, two tail paired t-test).

6.1 Analysis

(1) Analysis—UAS. The UAS metric relates to identifying the head node of the dependent node correctly. I3rab has a higher UAS than PADT, thus indicating that the new developed proposed I3rab parser annotation system produces better results relative to its annotation system. Moreover, the UAS results indicated that I3rab has lower syntactic complexity than PADT [Nivre 2009].

The unlabeled dependency relation has two main attributes: the direction of relation and the distance\(^5\) between the head and modifier. These two attributes directly affect the value of UAS. For the first attribute, the direction of the dependency relation is the direction of the arc from the head node toward the dependent node. If the head node precedes the dependent node (Index\(_{head} <\) Index\(_{dependent}\)), then the direction of the dependency relation is RIGHT. If the dependent node precedes the head node, then the direction of dependency relation is LEFT (Index\(_{head} >\) Index\(_{dependent}\)). For example, the sentence “قبل فترة وقع حادث مماثل، a similar incident happened a while ago” has the same dependency structure in PADT and I3rab. The token “وقت، waqa’a, happened” is the head for the token “حادث، haḍīṭun, incident,” and the direction of the dependency relation is RIGHT. In contrast, the token “قبل، قبل، qabla, while ago,” and the direction of the dependency relation is LEFT. The unlabeled dependency structure of the sentence is shown in Figure 7.

The percentages of both directions for the dependency relations in the PADT and I3rab datasets are shown in Table 11. Both datasets clearly had a high percentage of the RIGHT direction dependency relation. In other words, the governors tended to precede the items they governed.

\(^5\)Dependecy distance $= \text{absolute (Index}_{head} - \text{Index}_{dependent}) - 1$. 
The high percentage in both datasets showed that the Arabic language tends to support the RIGHT dependency relation more than the LEFT dependency relation.

I3rab had a higher percentage of RIGHT dependency relations than PADT. From the perspective of a supervised data-driven dependency parser, this aspect makes learning and predicting the direction of dependency relation more appropriate. For the second attribute, the distance between the head and its modifiers is related to the long dependency distance problem that the dependency parsing task faces. In general, if this distance is increased, then the task of linking the head with its modifier(s) becomes more difficult, and the UAS value may be decreased. The long dependency distance problem can be divided into two sub-problems. The first problem is the long dependency distance between the ROOT node and main word(s) in the sentence. The second problem is the long dependency distance between the head and its modifier(s) (where the head is not the ROOT). In Figure 8(a), I3rab shows 33% of cases in which the distance between the ROOT and main word is one, whereas PADT has 24%. From Figure 8(b), I3rab and PADT show similar distributions for the distance between the head and its modifier(s). From the results, we conclude that the effect of a long dependency distance between the ROOT and main word(s) affects the value of UAS. That is, as the ratio of one distance increases, the UAS accuracy increases. This conclusion implies that the main concept of I3rab in considering the first word in the sentence as the main word will absolutely increase the ratio of one distance between the ROOT and main word(s). Moreover, because I3rab has a higher UAS ratio than PADT, the parser task is expected to be easier with I3rab than PADT. Consequently, the annotation approach in I3rab in general appears to decrease the complexity of the syntax structure of Arabic sentences [Nivre 2009].

(2) Analysis—LAS. The LAS metric relates to determining the type of dependency relation. In this early stage in developing I3rab treebank, we used small datasets, and thus many labels were sparsely represented in both datasets. The cardinality of dependency relations in both datasets was classified into five categories: very high (30–36%), high (10–15%), moderate (5–9%), low (1–4%), and rare (<1%). The cardinality distribution is shown in Figure 9.

In general, sparsity has a negative effect on LAS. Although, the distribution in I3rab was worse than that in PADT, I3rab has a higher LAS than PADT. This higher value was mainly related to the higher value of UAS. We expect to achieve an improvement in LAS as the size of datasets is increased.

To check the quality of the generated dependency structure of Arabic sentences produced by I3RAB parser, three Arabic linguistics were contacted as human tester to evaluate the testing dataset. The resultant output of the parser was sent to the three Arabic linguistics. Each Arabic linguistic marked each relation in each sentence either as “True” or “False.” If the relation was annotated correctly by the parser from the linguistic point of view, then it was labeled as “True”; otherwise, it was labeled as “False.” Finally, the evaluation results from the Arabic linguistics were compared with each other. When the three linguistics agreed on labeling the relation (either as true or false), then the relation was labeled as agreed. Otherwise, if the three linguistics were not agreed on a relation label, then the relation was labeled with the majority label. Table 12 shows the ID of the 131 testing sentences with the number of relations in each sentence, the number of correctly annotated relations and the percentage of the accuracy of I3rab parser for each sentence. The accuracy is calculated as the number of correctly annotated relations divided by the total

---

6To evaluate this problem in a better way, we ignore the dependency relation between the ROOT and dot (.) at the end of sentences. This relation is ignored, because both I3rab and PADT agree in considering the dot (.) at the end of the sentence as a mandatory modifier for the ROOT node, so it has the highest dependency distance compared with other tokens in the sentence, and its value is the length of the sentence.
Table 12. The Accuracy Values for 131 Unseen Arabic Sentences

| Sentence ID | No of relations per sentence | No of corrected annotated relations | Accuracy % | Sentence ID | No of relations per sentence | No of corrected annotated relations | Accuracy % |
|-------------|------------------------------|-----------------------------------|------------|-------------|------------------------------|-----------------------------------|------------|
| 1           | 4                            | 2                                 | 50.0%      | 67          | 19                           | 17                                 | 89.5%      |
| 2           | 66                           | 49                                | 74.2%      | 68          | 2                            | 1                                  | 50.0%      |
| 3           | 44                           | 32                                | 72.7%      | 69          | 70                           | 56                                 | 80.0%      |
| 4           | 9                            | 7                                 | 77.8%      | 70          | 105                          | 80                                 | 76.2%      |
| 5           | 35                           | 30                                | 85.7%      | 71          | 65                           | 54                                 | 83.1%      |
| 6           | 24                           | 22                                | 91.7%      | 72          | 104                          | 84                                 | 80.8%      |
| 7           | 25                           | 19                                | 76.0%      | 73          | 1                            | 1                                  | 100%       |
| 8           | 41                           | 26                                | 63.4%      | 74          | 106                          | 86                                 | 81.1%      |
| 9           | 43                           | 33                                | 76.7%      | 75          | 100                          | 82                                 | 82.0%      |
| 10          | 26                           | 26                                | 100%       | 76          | 46                           | 41                                 | 89.1%      |
| 11          | 31                           | 28                                | 90.3%      | 77          | 68                           | 55                                 | 80.9%      |
| 12          | 11                           | 9                                 | 81.8%      | 78          | 9                            | 7                                  | 77.8%      |
| 13          | 63                           | 49                                | 77.8%      | 79          | 66                           | 58                                 | 87.9%      |
| 14          | 63                           | 45                                | 71.4%      | 80          | 107                          | 97                                 | 90.7%      |
| 15          | 96                           | 69                                | 71.9%      | 81          | 32                           | 23                                 | 71.9%      |
| 16          | 95                           | 77                                | 81.1%      | 82          | 71                           | 66                                 | 93.0%      |
| 17          | 34                           | 30                                | 88.2%      | 83          | 105                          | 92                                 | 87.6%      |
| 18          | 11                           | 11                                | 100%       | 84          | 126                          | 104                                | 82.5%      |
| 19          | 85                           | 69                                | 81.2%      | 85          | 15                           | 13                                 | 86.7%      |
| 20          | 75                           | 59                                | 78.7%      | 86          | 56                           | 48                                 | 85.7%      |
| 21          | 7                            | 6                                 | 85.7%      | 87          | 66                           | 52                                 | 78.8%      |
| 22          | 33                           | 26                                | 78.8%      | 88          | 64                           | 54                                 | 84.4%      |
| 23          | 4                            | 3                                 | 75.0%      | 89          | 39                           | 33                                 | 84.6%      |
| 24          | 25                           | 21                                | 84.0%      | 90          | 40                           | 32                                 | 80.0%      |
| 25          | 30                           | 23                                | 76.7%      | 91          | 3                            | 3                                  | 100%       |
| 26          | 11                           | 10                                | 90.9%      | 92          | 23                           | 21                                 | 91.3%      |
| 27          | 36                           | 24                                | 66.7%      | 93          | 45                           | 34                                 | 75.6%      |
| 28          | 42                           | 28                                | 66.7%      | 94          | 52                           | 39                                 | 75.0%      |
| 29          | 16                           | 15                                | 93.8%      | 95          | 31                           | 24                                 | 77.4%      |
| 30          | 51                           | 39                                | 76.5%      | 96          | 31                           | 27                                 | 87.1%      |
| 31          | 133                          | 111                               | 83.5%      | 97          | 10                           | 6                                  | 60.0%      |
| 32          | 8                            | 8                                 | 100%       | 98          | 3                            | 3                                  | 100%       |
| 33          | 26                           | 20                                | 76.9%      | 99          | 31                           | 25                                 | 80.6%      |
| 34          | 21                           | 16                                | 76.2%      | 100         | 16                           | 11                                 | 68.8%      |
| 35          | 29                           | 24                                | 82.8%      | 101         | 36                           | 25                                 | 69.4%      |
| 36          | 44                           | 34                                | 77.3%      | 102         | 25                           | 16                                 | 64.0%      |
| 37          | 22                           | 18                                | 81.8%      | 103         | 41                           | 35                                 | 85.4%      |
| 38          | 21                           | 18                                | 85.7%      | 104         | 9                            | 9                                  | 100%       |
| 39          | 49                           | 39                                | 79.6%      | 105         | 24                           | 19                                 | 79.2%      |
| 40          | 29                           | 22                                | 75.9%      | 106         | 42                           | 30                                 | 71.4%      |

(Continued)
| Sentence ID | No of relations per sentence | No of corrected annotated relations | Accuracy % |
|-------------|-----------------------------|-------------------------------------|------------|
| 41          | 32                          | 25                                  | 78.1%      |
| 42          | 16                          | 15                                  | 93.8%      |
| 43          | 36                          | 33                                  | 91.7%      |
| 44          | 20                          | 18                                  | 90.0%      |
| 45          | 26                          | 21                                  | 80.8%      |
| 46          | 26                          | 20                                  | 76.9%      |
| 47          | 40                          | 27                                  | 67.5%      |
| 48          | 44                          | 37                                  | 84.1%      |
| 49          | 41                          | 23                                  | 56.1%      |
| 50          | 39                          | 29                                  | 74.4%      |
| 51          | 30                          | 19                                  | 63.3%      |
| 52          | 35                          | 23                                  | 65.7%      |
| 53          | 10                          | 9                                   | 90.0%      |
| 54          | 36                          | 32                                  | 88.9%      |
| 55          | 60                          | 44                                  | 73.3%      |
| 56          | 69                          | 44                                  | 63.8%      |
| 57          | 22                          | 14                                  | 63.6%      |
| 58          | 171                         | 134                                 | 78.4%      |
| 59          | 12                          | 12                                  | 100%       |
| 60          | 40                          | 34                                  | 85.0%      |
| 61          | 62                          | 50                                  | 80.6%      |
| 62          | 115                         | 90                                  | 78.3%      |
| 63          | 37                          | 32                                  | 86.5%      |
| 64          | 61                          | 48                                  | 78.7%      |
| 65          | 14                          | 9                                   | 64.3%      |
| 66          | 79                          | 60                                  | 75.9%      |

| Sentence ID | No of relations per sentence | No of corrected annotated relations | Accuracy % |
|-------------|-----------------------------|-------------------------------------|------------|
| 107         | 40                          | 34                                  | 85.0%      |
| 108         | 50                          | 38                                  | 76.0%      |
| 109         | 24                          | 18                                  | 75.0%      |
| 110         | 45                          | 39                                  | 86.7%      |
| 111         | 4                           | 4                                   | 100%       |
| 112         | 12                          | 11                                  | 91.7%      |
| 113         | 6                           | 4                                   | 66.7%      |
| 114         | 116                         | 94                                 | 81.0%      |
| 115         | 23                          | 16                                  | 69.6%      |
| 116         | 40                          | 36                                  | 90.0%      |
| 117         | 14                          | 12                                  | 85.7%      |
| 118         | 6                           | 4                                   | 66.7%      |
| 119         | 22                          | 18                                  | 81.8%      |
| 120         | 48                          | 36                                  | 75.0%      |
| 121         | 75                          | 64                                  | 85.3%      |
| 122         | 3                           | 3                                   | 100%       |
| 123         | 40                          | 33                                  | 82.5%      |
| 124         | 30                          | 21                                  | 70.0%      |
| 125         | 33                          | 26                                  | 78.8%      |
| 126         | 33                          | 26                                  | 78.8%      |
| 127         | 17                          | 11                                  | 64.7%      |
| 128         | 28                          | 23                                  | 82.1%      |
| 129         | 13                          | 8                                   | 61.5%      |
| 130         | 47                          | 36                                  | 76.6%      |
| 131         | 15                          | 10                                  | 66.7%      |

Average accuracy 80%

number of relations per sentence. The average value of the percentage of the accuracy among all sentences is equal to 80%, which is considered to be positive results.

6.2 Discussion

This article argues that the main reasons behind the higher value of UAS and LAS for I3rab than PADT are related to the concepts of determining the main word of the sentence and the explicit presentation for all pronouns: independent, joined, or covert. In this section, we discuss the common linguistic structure of Arabic sentences:

(1) The nominal sentence.
   (a) The pure nominal sentence.

The pure nominal sentence has no verb at all. It is a common and frequent linguistic structure in Arabic sentences. In the case of a pure nominal sentence that is not introduced by an abolisher, I3rab and PADT use a similar approach in constructing the dependency structure. For example, in the sentence “وصول وزير الخارجية الأمريكي إلى بيروت”, wuṣūlu wazir al-ḥārifiyati al-ʾamrikiyi
In this case of pure nominal sentence I3rab and PADT follow the approach of considering the topic of the sentence as the main word in the sentence. This approach reduces the distance between the ROOT and its modifiers that implies to avoid the long distance between the ROOT and its modifiers. It is worth mentioning that in nominal sentences, mostly the topic is the first word in the sentence. It often comes before the predicate although there are some cases that the predicate comes before the topic.

7 It is worth mentioning that in the original paper of PADT, PADT followed the approach of considering the predicate (لا) as the main word in the sentence, because they argued that the predicate is not usually omitted in the sentence, but the topic could be omitted in some cases. This sentence is extracted from the PADT freely available from LDC.
In the case of a pure nominal sentence introduced by an abolisher of the Inna-its-sister type, PADT and I3rab do not use the same approach in constructing the dependency structure. In the sentence “ان العراقيين قادرون على تقرير مصيرهم بأنفسهم” the PADT approach links the predicate “قادرون,” qādirūna, are capable” as a modifier to the abolisher “ان,” ‘inna, that” and links the topic “ال العراقيين,” al-‘irāqiyya, Iraqis” as a modifier to the predicate “قادرون, qādirūna, are capable.” In contrast, the I3rab approach links the topic “ان,” ‘inna, that” as its modifiers. The dependency structure of the sentence is illustrated in Figure 11. By comparing PADT with I3rab, we find that PADT increases the dependency distance, whereas I3rab has minimal dependency distance. Moreover, PADT adopts the LEFT direction for the dependency relation, but I3rab keeps the RIGHT direction for the dependency relation.

For abolishers of the Kana-its-sister type, PADT and I3rab use a similar approach in constructing the dependency structure. As shown in the sentence “كان محمد يقرأ الكتاب الجديد,” kāna muḥammadun
Fig. 11. The dependency tree for the sentence “ان العراقيين قادرون على تقرير مصيرهم بأنفسهم.”, Iraqis are capable of self-determination.”

Fig. 12. The dependency tree for the sentence “”محمد يقرأ الكتاب الجديد.”, Mohammad was reading the new book.”

yaqra’u al-kitāba al-ğādiḍa, Mohammad was reading the new book,” the dependency structure of the sentence in both I3rab and PADT links the topic “محمد, muḥammadun, Mohammed” and the predicate ”يرأى, yaqra’u, reads” to abolisher “كان, kāna, was” as modifiers. In this case, both PADT and I3rab have minimal dependency distance and keep the RIGHT direction for the dependency relation. The PADT and I3rab dependency structure of the sentence is illustrated in Figure 12.

(b) The nominal sentence with a predicate as a verbal sentence.

Constructing the dependency structure of this type of sentence is one of the major differences between the PADT and I3rab approaches. For example, in the sentence “وزير الخارجية الأمريكي يرحّب بمبادرة السلام الهندية,” waziru al-ḥāriḍiyati al-‘amrikiyi yuraḥḥibu bi mubādarati assalāmi al-hindiyati, U.S. Secretary of State Welcomes Indian peace initiative,” PADT considers the verb ”يرحب, yuraḥḥibu, welcomes” as the main word in the sentence and associates it with the ROOT as modifier. In addition, it considers the word ”وزير, waziru, minister” as a subject that precedes the
verb and links it to the verb "يرحب, yurahhibu, welcomes" as a modifier with the LEFT direction for the dependency relation. However, I3rab considers this sentence as a nominal sentence that is not introduced by any of abolishers. I3rab considers the topic "وزير, waziru, minister" as the main word in the sentence and links it with the ROOT node as a modifier. I3rab also addresses the concept that each verb should have a subject, and this subject must come after the verb. In this sentence, the verb "يرحب, yurahhibu, welcomes" has a covert subject pronoun that is surmised to be "هو, huwa, he" and is explicitly represented as an individual token in the sentence. This covert pronoun is linked to the verb "يرحب, yurahhibu, welcomes" as an agent with the RIGHT direction for the dependency relation. The dependency structure of the sentence following PADT and I3rab approaches is shown in Figure 13.

By comparing PADT with I3rab in the case of sentences with a verb, we find that PADT increases the dependency distance while I3rab has minimal dependency distance. Moreover, PADT adopts the LEFT direction for the dependency relation, but I3rab keeps the RIGHT direction for the dependency relation.

2. The verbal sentence. In a sentence starting with a verb, both PADT and I3rab use a similar approach. For example, in the sentence "يا تحق متسبو الشرطة العليا بغداد، جنرال ضابط، Local police officers join Baghdad at their former headquarters," the subject "متسبو, muntasib, officers" is linked as a modifier to the verb "يتحق, yaltahiq, join." The dependency structure of the sentence following the PADT and I3rab approaches is shown in Figure 14. In this case, both PADT and I3rab have minimal dependency distance and keep the RIGHT direction for the dependency relation.

However, if the sentence starts with an accusative or a jussive particle, then the behavior in PADT and I3rab differs. For example, in the sentence "لن يقرأ محمد الكتاب ليلا, lan yaqra’a muhammadun al-kitaba laylan, Muhammad will not read the book at night," PADT considers the accusative particle “لا, lan, will not” as a modifier of the verb “يقرأ, yaqra’a, reads,” and the direction of dependency relation is LEFT. I3rab considers the accusative particle “لا, lan, will not” as a head to the verb “يقرأ, yaqra’a, reads,” and the direction of dependency relation is RIGHT. The dependency structure of the sentence following PADT and I3rab approaches is shown in Figure 15.
Fig. 14. The dependency tree for the sentence 

بلتحق متسوس الشرطة المحلية ببغداد عميات عملهم السابعة

Local police officers join Baghdad at their former headquarters.”

Fig. 15. The dependency tree for the sentence “

لن يقرأ محمد الكتاب ليلًا

Muhammad will not read the book at night.”

The same issue is repeated with a jussive particle. In the sentence

لم يقرأ محمد الكتاب في الكتبة

PADT considers the jussive particle “ل، lam, did not” as a modifier of the verb “قرأ, yaqra’, reads,” and the direction of the dependency relation is LEFT. I3rab considers the jussive particle “لم, lam, did not” as the head to verb, and the direction of dependency relation is RIGHT. The dependency structure of the sentence following PADT and I3rab approaches is shown in Figure 16.

By comparing PADT with I3rab in the case of verbal sentences that start with an accusative or a jussive particle, PADT increases the dependency distance, whereas I3rab has minimal dependency distance. In addition, PADT adopts the LEFT direction for the dependency relation, but I3rab keeps the RIGHT direction for the dependency relation.

ACM Trans. Asian Low-Resour. Lang. Inf. Process., Vol. 21, No. 2, Article 23. Publication date: November 2021.
7 CONCLUSION AND FUTURE WORK

In this article, we have presented a new dependency treebank for Arabic language that allows the syntactic tree to reflect the real characteristics of Arabic sentences. The performance of the dependency parser using the newly constructed I3rab treebank was compared with its performance using the PADT, which is the competitive dependency treebank for Arabic. We have demonstrated the effectiveness of our approach of constructing dependency treebank based on the concepts and theories identifying the linguistic structure of Arabic. In general, the I3rab approach tends to provide minimal dependency distance and to keep the direction of dependency relation to the RIGHT direction as much as possible. The minimal dependency distance simplifies the parser task, because dealing with small distances is easier than dealing with long distances. However, keeping one direction for the most dependency relations increases the stability in training the parser model. The results showed that we gained an improvement of 10.24% and 18.42% in UAS and LAS, respectively, by using I3rab rather than PADT.

We foresee numerous venues for future work. First, we plan to perform Inter-annotator agreement to improve the current annotation rules. Another avenue will be the generalization of I3rab treebank to cover the phrase level of analysis. Another avenue for future work is to develop a universal version of the I3rab dependency treebank to facilitate cross-lingual studies, and to compare the new Arabic dependency treebank with the existing resources for Arabic and other languages.
## APPENDICES

### A SAMPLE CONLL FILE

| id  | form     | lemma   | cpos  | postag | feats | head | deprel | phead | deprel |
|-----|----------|---------|-------|--------|-------|------|--------|-------|--------|
| 1   | mane_ة  | mane_ة  | C     | C-     |       | 0    | COND   | _     | _      |
| 2   | ya3ahal_wa | ya3ahal_wa | V     | Vl    | Mood=| 1    | VB     | _     | _      |
| 3   | haww_ة   | haww_ة  | S     | S-    | Gender=M| 2    | AGENT  | _     | _      |
| 4   | ya3ahal_wa | ya3ahal_wa | V     | Vl    | Mood=| 4    | VB     | _     | _      |
| 5   | haww_ة   | haww_ة  | S     | S-    | Gender=M| 4    | AGENT  | _     | _      |
| 6   | ba3ira-r_ة | ba3ira-r_ة | V     | Vp    | _     | 0    | VB     | _     | _      |
| 7   | al-ya3adsi _ة | al-ya3adsi | N    | N-    | Gender=M| 1    | AGENT  | _     | _      |
| 8   | il7la _ة | il7la _ة | C     | C-    |       | 1    | EXCEPT | _     | _      |
| 9   | slyayn_ة | slyayn_ة | N     | N-    | Gender=M| 3    | OBJ    | _     | _      |

**Note:**
- ID: Unique identifier for each word.
- Form: The original word form.
-Lemma: The base form of the word.
-CPOS: The constituent position.
-POSTAG: The part of speech.
- FEATS: Morphological features.
- HEAD: The head of the word.
- DEPREL: The dependency relation.

---

ACM Trans. Asian Low-Resour. Lang. Inf. Process., Vol. 21, No. 2, Article 23. Publication date: November 2021.
| id | form | lemma | pos | mpos | tense | head | deprel | head | deprel | phs | deprel | phs | deprel |
|----|------|-------|-----|------|-------|------|--------|------|--------|-----|--------|-----|--------|
| 2 | al-qafal | al-qafal | N | N | Gender=M | Number=S | case=1 | Def=0 | 1 | PREP | PP | - | - | - |
| 3 | mubammadun | mubammadun | N | N | Gender=M | Number=S | case=1 | Def=0 | 0 | TOPIC | - | - | - | - |
| 2 | ya'ara | ya'ara | V | V | Mood=A | Voice=A | Person=3 | Gender=M | Number=S | 1 | PREP | VP | - | - | - |
| 3 | huwa | huwa | S | S | Gender=M | Number=S | case=1 | Def=0 | 2 | AGENT | - | - | - | - |
| 4 | al-kitab | al-kitab | N | N | Gender=M | Number=S | case=4 | Def=0 | 2 | OBJ | - | - | - | - |
| 5 | fawqa | fawqa | A | A | Gender=M | Number=S | case=4 | Def=0 | 1 | PREP | VP | - | - | - |
| 6 | al-maktab | al-maktab | N | N | Gender=F | Number=S | case=1 | Def=0 | 5 | GEN | - | - | - | - |
| 1 | al- | al- | C | C | - | - | - | 0 | P-ACC | - | - | - | - |
| 2 | al-kitab | al-kitab | N | N | Gender=M | Number=S | case=4 | Def=0 | 1 | TOPICX | - | - | - | - |
| 3 | fawqa | fawqa | A | A | Gender=M | Number=S | case=4 | Def=0 | 1 | PREP | VP | - | - | - |
| 4 | al-maktab | al-maktab | N | N | Gender=F | Number=S | case=1 | Def=0 | 3 | GEN | - | - | - | - |
| 1 | al- | al- | C | C | - | - | - | 0 | P-ACC | - | - | - | - |
| 2 | al-fata'a | al-fata'a | N | N | Gender=F | Number=S | case=4 | Def=0 | 1 | TOPICX | - | - | - | - |
| 3 | fawqa | fawqa | A | A | Gender=M | Number=S | case=4 | Def=0 | 3 | GEN | - | - | - | - |
| 4 | huwa | huwa | S | S | Gender=M | Number=S | case=1 | Def=0 | 3 | OBJ | - | - | - | - |
| 5 | al-kitab | al-kitab | N | N | Gender=M | Number=S | case=4 | Def=0 | 3 | OBJ | - | - | - | - |
| 6 | fawqa | fawqa | A | A | Gender=M | Number=S | case=4 | Def=0 | 1 | PREP | PP | - | - | - |
| 7 | al-maktab | al-maktab | N | N | Gender=F | Number=S | case=2 | Def=0 | 3 | GEN | - | - | - | - |
B  EXAMPLE OF DEPENDENCY PARSING FOR ARABIC STATEMENT BASED ON I3rab

Example 1: “He who strives will succeed”

Example 2: “Only friends attended me, except Ali”

Example 3: “Commander Saeed won”

Example 4: “I bought a horse fast”

Example 5: “We’ll go on a trip tomorrow”

Example 6: “Mohamed does not drive fast”

Example 7: “Muhammed and Salem, the Taliban, are diligent”

Example 8: “The sun is shining”

Example 9: “The book is on the table”

Example 10: “The girl has long hair”
REFERENCES

M. Alosh. 2005. Using Arabic: A Guide to Contemporary Usage. Cambridge University Press, Cambridge, UK.

F. Alotaiby, S. Foda, and I. Alkharashi. 2010. Clitics in Arabic language: A statistical study. In Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation. 595–601.

I. A. Al-Sughaiyer and I. A. Al-Kharashi. 2004. Arabic morphological analysis techniques: A comprehensive survey. J. Amer. Soc. Info. Sci. Technol. 55, 3 (2004), 189–213.

R. G. Athreya, S. K. Bansal, A. C. N. Ngomo, and R. Usbeck. 2021. Template-based question answering using recursive neural networks. In Proceedings of the IEEE 15th International Conference on Semantic Computing (ICSC’21). IEEE, 195–198.

M. Attia. 2007. Arabic tokenization system. In Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources. 65–72.

M. A. Attia. 2008. Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation. The University of Manchester, UK.

A. Awajan. 2007. Arabic text preprocessing for the natural language processing applications. Arab Gulf J. Sci. Res. 25, 4 (2007), 179–189.

A. Awajan. 2015. Keyword extraction from Arabic documents using term equivalence classes. ACM Trans. Asian Low-Res. Lang. Info. Process. 14, 2 (2015), 1–18.

D. M. Bikel. 2004. On the Parameter Space of Generative Lexicalized Statistical Parsing Models. University of Pennsylvania.

A. Böhmová, J. Hajič, E. Hajičová, and B. Hladká. 2003. The Prague dependency treebank. In Treebanks. Springer, Dordrecht, 103–127.

S. Buchholz and E. Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X’06). 149–164.
I3rab: A New Arabic Dependency Treebank Based on Arabic Grammatical Theory

N. Chomsky. 2009. Syntactic Structures. De Gruyter Mouton, Berlin.
M. Civit and M. Marts. 2005. Building Cast3LB: A Spanish treebank. Res. Lang. Comput. 2, 4 (2005), 549–574.
M. Civit, N. Buñi, and P. Valverde. 2004. Cat3LB: A treebank for Catalan with word sense annotation. In Proceedings of the 3rd Workshop on Treebanks and Linguistic Theories. Tuebingen, Germany.
M. Collins. 2003. Head-driven statistical models for natural language parsing. Comput. Linguist. 29, 4 (2003), 589–637.
P. R. Comas, J. Turmo, and L. Márquez. 2010. Using dependency parsing and machine learning for factoid question answering on spoken documents. In Proceedings of the 11th Annual Conference of the International Speech Communication Association.
K. Dukes and T. Buckwalter. 2010. A dependency treebank of the Quran using traditional Arabic grammar. In Proceedings of the 7th International Conference on Informatics and Systems (INPOS’10). IEEE, 1–7.
K. Dukes and N. Habash. 2011. One-step statistical parsing of hybrid dependency-constituency syntactic representations. In Proceedings of the 12th International Conference on Parsing Technologies. 92–103.
K. Dukes. 2015. Statistical parsing by machine learning from a classical Arabic treebank. Retrieved from https://arXiv:1510.07193.
E. Fayyouni and S. Idwan. 2021. Semantic partitioning and machine learning in sentiment analysis. Data 6 (2021), 67. https://doi.org/10.3390/data6060067
A. Frank, A. Zaenen, and E. Hinrichs. 2012. Treebanks: Linking linguistic theory to computational linguistics. Linguis. Iss. Lang. Technol. 7, 1 (2012).
M. Galley and C. D. Manning. 2009. Quadratic-time dependency parsing for machine translation. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. 773–781.
J. Gillenwater, X. He, J. Gao, and L. Deng. 2013. End-to-end learning of parsing models for information retrieval. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 3312–3316.
A. Bharati, V. Chaitanya, and R. Sangal. 1995. Natural language processing: A Paninian perspective. Indian Institute of Technology, Kanpur. New Delhi: Prentice-Hall of India.
N. Habash and R. Roth. 2009. Catib: The Columbia Arabic treebank. In Proceedings of the ACL-IJCNLP Conference Short Papers. 221–224.
J. Hajic, O. Smrž, P. Zemánek, J. Snáidauf, and E. Beška. 2004. Prague Arabic dependency treebank: Development in data and tools. In Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools. 110–117.
D. Halabi, A. Awajen, and E. Fayyouni. 2017. Arabic LFG-inspired dependency treebank. In Proceedings of the International Conference on New Trends in Computing Sciences (ICTCS’17). IEEE, 207–215.
D. Halabi, A. Awajen, and E. Fayyouni. 2020. Improving Arabic dependency parsers by using dependency relations. In Proceedings of the 21st International Arab Conference on Information Technology (ACT’20). IEEE, 1–7.
D. Halabi, A. Awajen, and E. Fayyouni. 2021. Syntactic annotation in the I3rab dependency treebank. Int. Arab J. Inf. Technol 18, 1 (2021).
J. Hall, J. Nilsson, and J. Nivre. 2013. Malteval. Retrieved from http://www.maltparser.org/malteval.html.
C. H. Han, N. R. Han, E. S. Ko, P. Martha, and Y. Heejong. 2002. Penn Korean treebank: Development and evaluation. In Proceedings of the Korean Society for Language and Information Conference. Korean Society for Language and Information, 69–78.
J. Katz-Brown, S. Petrov, R. McDonald, F. J. Och, D. Talbot, H. Ichikawa, M. Seno, and H. Kazawa. 2011. Training a parser for machine translation reordering. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 183–192.
S. Kübler, R. McDonald, and J. Nivre. 2009. Dependency parsing. Synth. Lect. Hum. Lang. Technol. 1, 1 (2009), 1–127.
S. Kulick, R. Gabbard, and M. Marcus. 2006. Parsing the Arabic treebank: Analysis and improvements. In Proceedings of the Treebanks and Linguistic Theories Conference. 31–42.
LDC. 2004a. Buckwalter Arabic morphological analyzer version. Retrieved from https://catalog.ldc.upenn.edu/LDC2004L02.
LDC. 2004b. Prague Arabic dependency treebank 1.0. Retrieved from https://catalog.ldc.upenn.edu/docs/LDC2004T23/.
LDC. 2007. Prague Arabic dependency treebank ++. Retrieved from http://padt-online.blogspot.com/2007/01/conll-shared-task-2007.html.
LDC. 2018. 2007 CoNLL shared task—Arabic and English. Retrieved from https://catalog.ldc.upenn.edu/LDC2018T08.
H. Li and F. Xu. 2016. Question answering with DBpedia based on the dependency parser and entity-centric index. In Proceedings of the International Conference on Computational Intelligence and Applications (ICCIA’16). IEEE, 41–45.
T. Lynn. 2016. Irish Dependency Treebanking and Parsing. PhD Thesis, Dublin City University and Macquarie University Sydney, 2016.
M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki. 2004. The Penn Arabic treebank: Building a large-scale annotated Arabic corpus. In Proceedings of the NEMLAR Conference on Arabic Language Resources and Tools. 466–467.
