Vol.3, No.5, Page 1178- 1186, 2025 https://doi.org/10.59628/jast.v3i5.1468 ISSN 2958-9568 JAST

SUN PUBLISHER

مجلة جامعة صنعاء للعلوم التطبيقية والتكنولوجيا

Sana'a University Journal of Applied Sciences and Technology

https://journals.su.edu.ye/index.php/jast/

Annotation of Inchoative and Predicate in Arabic Nominal Sentence

Saeed AL-Dobai[1] *, Bakeel Azman[2], Ghalib AL-Gaphari[2] and Monier Ana'am[3]

[1]**Department of Computer & Mathematics, Faculty of Sciences, Sana'a University, Sana'a, Yemen,**
[2]**Department of Computer Science, Faculty of Computer and Information Technology, Sana'a University, Sana'a, Yemen,**
[3]**Department of Arabic Language, Faculty of Education, Sana'a University, Sana'a, Yemen**

*Corresponding author: Bakeel.Azman@su.edu.ye

Abstract

Linguistic annotation provides additional information associated with a particular purpose in a document or another piece of information. It is widely used in various fields, ranging from computing and bioinformatics to psychology, law, and linguistics. This study aimed to develop a linguistic annotator for nominal Arabic sentences. The annotator detects both the Inchoative (I) and Predicate (P) phrases within nominal sentences in Arabic text. This is based on a shallow parsing method for chunking sentence constituents. The motivation behind this study is to produce a tool that enables further analysis, such as rhetorical parsing. This annotator can help improve the accuracy of linguistic parsers and enhance natural language understanding. The experiments were conducted on a standard dataset supported by a competitive case study. These results are promising and encouraging.

Article Info

Keywords:
Inchoative, parsing, predicate, identification, subject, annotation, nominal.

Article History:
Received: 1-March-2025,
Revised: 11-July-2025,
Accepted: 14-August-2025,
Available online: 28 October 2025.

1. INTRODUCTION

A nominal sentence is one of the two types of Arabic sentences that do not start with a verb. Instead, it starts with a verbless phrase called the inchoative (subject) phrase, after which a phrase that completes the sentence structure and meaning is called the predicate phrase [1]. In English, the term "nominal sentence" can refer to two types of sentences. The first type of nominal sentence is a sentence in which the predicate is not a verb but is joined to the subject (inchoative in Arabic) by a copula containing a verb. The second type of nominal sentence did not contain a verb. The first and most common type of nominal sentence (in English) is a sentence in which the subject is followed by a predicate that contains a copula, connection, and predicative. The copula is a form of the verb "to be." For example, the sentence جون دكتور. "June is a doctor" is a nominal sentence of this type. The predicative in this case is called a nominative predicative because it centers on the noun "doctor" [2]. The second, rarer type of English nominal sentence is one in which the verb "to be" is absent but implied by the structure of the sentence. The missing verb "to be" is implied. Therefore, verbless sentences in Arabic do not consist of a subject but rather a topic (which is here the inchoative) followed by a predicate, and are only possible in present tense sentences [3].

Nominal sentences are relatively uncommon in English, but are much more frequent in Arabic language and other languages. For example, in Arabic, the nominal sentence "أحمد مجتهد." consists only of the name "أحمد" and the adjective "مجتهد." Translating the sentence into English requires the translator to insert the correct form of "to be." This is true not only in Arabic but also in other languages such as Hebrew, Russian, and Latin [4]. The practice of connecting the subject and predicative without a copula, as in these languages, is known as "zero copula," so Arabic is a zero copula language [5].

On the other hand, the two Arabic linguistic categories المسند إليه 'al-musnad-ilaihi' (the predicative-to or the subject) and المسند 'al-musnad' (the predicate) are the central

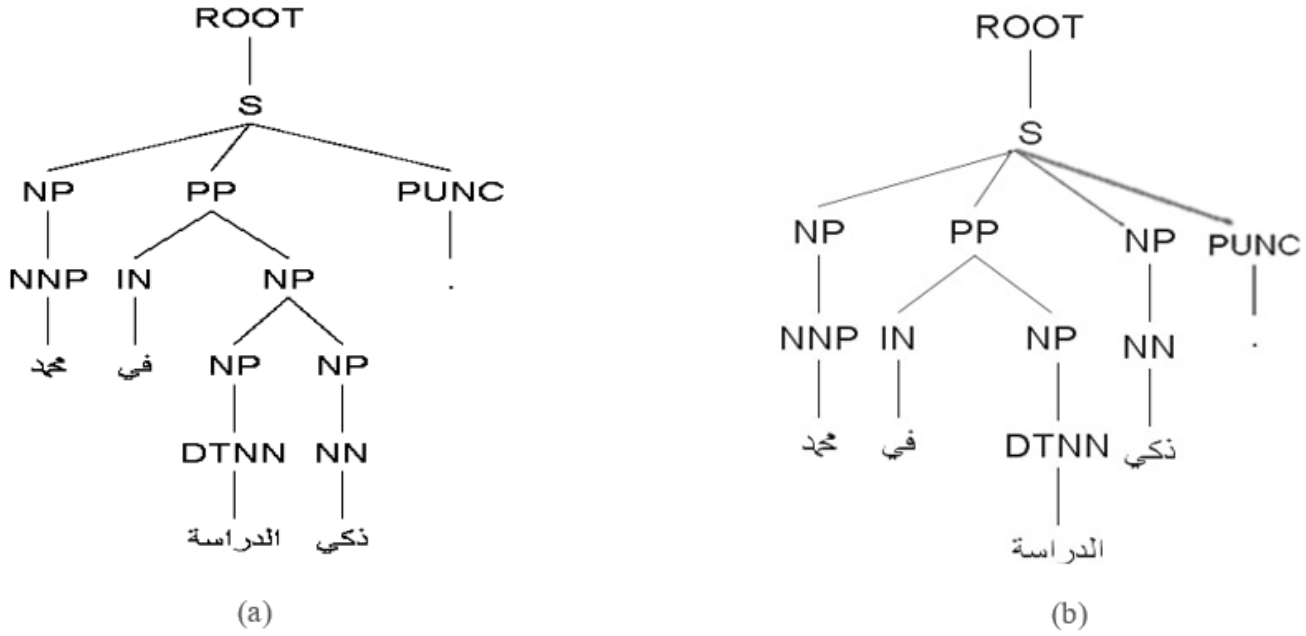

Figure 1. Stanford parse tree example

elements in any type of Arabic sentence, whether nominal or verbal, as well as in statements, interrogatives, imperatives, or any type of fully meaningful Arabic sentence [6]. A meaningful sentence is not complete until it has at least these two elements: 'al-musnad-ilaihi' and 'al-musnad'. By saying (زيد شجاع. – Zaid is a hero), we have attributed (الشجاعة – heroism), which is al-musnad, to (زيد – Zaid), which is al-musnad-ilaihi. Thus, I and P structure is the dominant forms of Arabic sentence composition, commonly called *attributed composition*.

The parsing process of written text, in which the sentence is the smallest unit for parsing, is the linguistic syntactic analysis of the text according to the language's formal grammar. In Arabic, parsing a sentence should begin by identifying the sentence type, whether nominal or a verbal. This enables parsing of a sentence according to the characteristics of its sentence type. Therefore, parsing a nominal sentence requires knowing two main phrases, I and P, whereas parsing a verbal sentence requires knowing the subject, verb, and object(s) of that sentence. Nominal sentence analysis, based on its main phrases (I and P), is essential for determining the main branches of the parse tree at its first level. Unfortunately, most current Arabic parsers do not pay attention to the head phrases of nominal sentences, despite the importance of both components in linguistic parsing and semantic parsing [7]. One of the dominant parsers is the Stanford parser [8], which does not focus on these phrases in the parse tree. For instance, parsing the sentence ("محمد في الدراسة ذكي", "Mohammed is intelligent in the study.") produces the sparse tree shown in Fig. [1]a. Such a tree should have considered the last NP (ذكي intelligent) in parallel with the first NP (محمد, Mohammed) from the first level of the tree, as shown in Fig. 1b. This means that the representation does not account for the particularities of nominal sentences such as the I and P constituents. Little attention has been paid to characterizing the nominal sentence itself, ignoring the mechanism that produces or deciphers it. No previous study has independently assumed detection these two main phrases in the sentence or annotating them. Researchers have focused on parsing texts and developing different approaches to achieve correct parsing. The concerned Arabic parsing models generally conduct the parsing process according to sentence phrases and then place them in the right node in the parse tree regardless of the importance of the phrase type in the sentence. This, in turn, produces incorrect parses for I and P at the correct node in the parse tree. Ignoring the importance of these phrases in the parsing process is considered a problem that leads, at least from our view, to an immature Arabic parser. **The purpose of this study is to address that problem.**

Therefore, this study presents a proposed annotator tool to enhance Arabic parsers by detecting I and P in nominal sentences, and then annotating them with two tags that can be included in a standard tagset such as the Penn Arabic Treebank (PATB) tagset. To achieve the detection task, a method such as shallow parsing or chunking should be adopted to segment a sentence into a sequence of syntactic constituents or chunks, that is sequences of adjacent words grouped based on linguistic properties [9]. Chunking is the basic task in partial parsing. Partial parsing was introduced as a response to the difficulties of full traditional parsing and is described as a technique to recover syntactic information efficiently and reliably from unrestricted text by sacrificing complete-

ness and depth of analysis [10]. Among the critiques of full parsing (and in favor of partial parsing), the most important is that full parsers are not sufficiently robust for many NLP applications, and that full parsing fails to identify a good parse tree in noisy environments [11, 12]. Recent progress in full statistical parsing shows that a full parser is not robust and can not produce reliable results when analyzing many different languages [13]. See, for instance, the CoNLL 2018 Shared Task on multilingual dependency parsing [10, 14].

This study offers a different perspective on sentence parsing using the shallow parsing method, which relies on a conceptual principle that serves as a foundation for semantic parsing. Any well-formed nominal sentence predominantly consists of an inchoative or predicate. These two constituents or phrases convey the core concepts of the utterance. Detecting I and P requires parsing a nominal sentence (S) into two tree branches according to its constituents: the left branch as I and the right branch as P. The complementary phrases of S should be associated with the respective branch, either I or P.

The I and P modes are syntactically represented by certain rules that define them. For example, if I is placed at the beginning of the sentence, so the first term determines the I mode type. Inchoative modes are limited and may include a singular noun, demonstrative pronoun, conditional, expression equivalent to an infinitive, separated pronoun, or interrogative particle [15]. Predicate modes, vary and may include an indefinite noun or singular noun, quasi-sentences (a prepositional phrase or adverbial phrase), nominal sub-sentence, or verbal subsentences [1]. These are all the I and P modes that may appear in a sentence.

Based on the premise that I and P annotations are central linguistic axes, they serve as the foundation for full parsing. Therefore, this study aims to develop an annotation tool that can categorize Arabic sentences into verbal and nominal types to enable nominal sentence identification. It then captures these two elements and annotates them as I and P, respectively. Such a tool may further enrich the performance of Arabic parsers in terms of structural and functional features, and serve as a foundation for a syntactic-semantic parser.

2. LITERATURE REVIEW

The Arabic language is well known for its morphological richness and syntactic complexity [16, 17]. Furthermore, it exhibits characteristics such as relatively free word order, long sentence structures, and omission of punctuation in written text. These features, along with other aspects such as writing without diacritics, make Arabic challenging to perform text parsing. In addition, this often leads to considerable ambiguity, as several words with different diacritic patterns may appear identical in a diacritic-less setting. In fact, text without diacritics can be difficult even for Arabic-speaking humans to read, let alone for computational processing applications [18].

In this section, we discuss some annotators who focus on Arabic sentence structures, such as the Treebank project [19]. The PATB is one of the most popular and extensively annotated corpora for Arabic text, modeled after the Penn English Treebank. It used the Buckwalter Arabic morphological analyzer [20] for transliterating, POS tagging, and morphological analysis of input text, with the output manually revised by trained annotators. The syntactic annotation was based on a rigorous understanding of, and adherence to, traditional Arabic grammar principles, and was performed manually [13, 21].

Two main approaches are used to annotate text: the manual approach, which depends on human labor, and the automatic approach, which uses annotation tools [22]. Owing to the complexity of Arabic text processing, most prior work on Arabic annotation have relied on manual annotation [19, 23, 24]. The annotation process involves multiple layers, as represented in the Quranic Arabic Corpus (QAC) [25]. This is a collaboratively constructed linguistic resource initiated at the University of Leeds, with multiple annotation layers including part-of-speech tagging, morphological segmentation, syntactic analysis, and semantic ontology.

Marton et al. [26] sought to enrich inflectional and morphological features to increase parser accuracy. They investigated sentence structure at the lexical level, including the forms of ending terms, whether nouns or verbs. This is highly relevant to discussions such as the identification of I and P. Reducing the complexity while maintaining core set tags has been a major focus of researchers.

Green et al. [27] stated that one baseline for improving Arabic parser precision is annotating sentences based on their type, either nominal or verbal, and recognizing phrases such as iDafa. Their results suggested that many current parsers could benefit from annotation consistency and syntactic enrichment in key configurations.

Although some efforts have been devoted to Arabic parsing, relatively few studies have focused on syntactic analysis. This is largely due to challenges such as the high degree of ambiguity, complex syntax, and lack of fixed grammar rules, in addition to the issues already mentioned [28]. Some progress has been made in recent years [27, 29–38], but there is still no general-purpose Arabic parser with a wide and robust coverage. At present, no analyzer appears capable of fully processing real-world Arabic text. Most systems focus on limited syntactic phenomena, often with significant lexical limitations. However, real-world texts, such as news articles, scientific abstracts, and web pages, typically contain diverse sentence structures that pose serious challenges for parsers [39].

Several parsing approaches have been proposed. The rule-based approach uses well-defined formal gram-

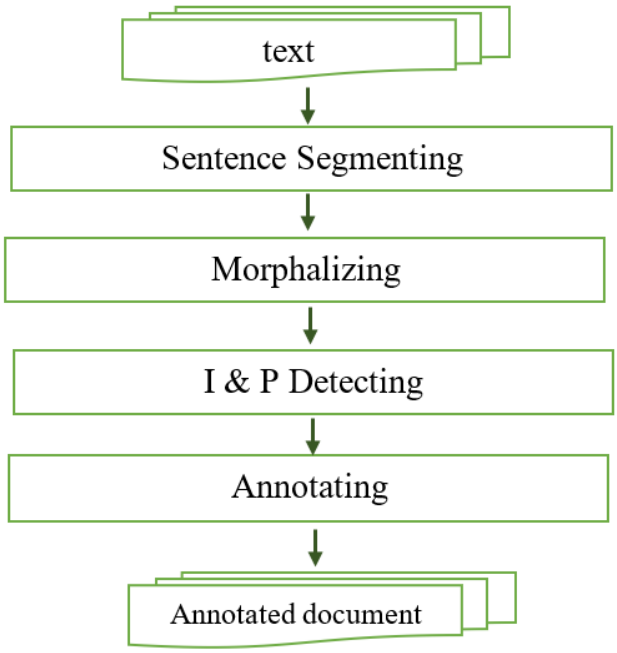

Figure 2. System architecture

mar [29, 40]. The statistical approach incorporates machine learning to automatically derive grammar rules [41]. The hybrid approach combines a predefined grammar with statistical models to resolve issues and improve parsing results [42]. The hybrid model was adopted by Green et al. [27], who developed a syntactic parser for Modern Standard Arabic (MSA) based on the Stanford parser. They compared their results to those of other systems such as Berkeley [43] and Bikel [30]. After identifying sources of ambiguity, they built a lexicalized PCFG with second-order Markovization and created their own manually annotated grammar. Thes parsers demonstrated good accuracy. **They concluded—relevantly to this paper**—that annotation consistency and syntactic enrichment (e.g., identifying I and P) lead to better parsing results. This supports our investigation of sentence syntactic structures.

Mona Diab [29] worked on shallow parsing through Base Phrase Chunking (BPC) for Arabic, using a rule-based approach. BPC involves grouping adjacent words into non-recursive chunks. To improve BPC for Arabic, Diab proposed a solution that extended her earlier work using Support Vector Machines (SVM). Her approach involved modification of POS tagging and BPC. She adopted IOB annotation for ten identified base phrase chunks. Training and testing were performed on various PATB versions using different tag sets and morphological features. However, she did not address a specific case of nominal sentence structures, particularly the I and P components.

Some efforts have been made to utilize chart parsing based on traditional grammar [44]. However, no specific treatment of the I and P phrases was included. Most recently, Sawalha et al. [45] presented morphological and syntactic parsing of Arabic. This study tackled vocabulary analysis and developed a model for certain structures in the Holy Quran. While it focused on annotating complex grammatical formulations, it did not include a comprehensive approach to sentence construction, including nominal sentence structure.

Most current linguistic parsers focus on sentence constituents simply as phrases without highlighting the most important phrases, particularly in nominal sentences. These studies did not parse sentences according to their type (nominal or verbal), although they are most closely related to the subject of this paper. **This gap highlights the novelty and originality of our research**.

3. PROPOSED WORK

This section focuses on the proposed model, which includes the design and implementation of an automatic detector for I and P in nominal Arabic sentences. This work represents one of the four phases intended to build an expanded annotator for identify and annotate Arabic rhetorical relations in a text. It begins with the development of an I and P annotation tool, capable of detecting the two components within Arabic sentences.

The basic approach in chunking is to exploit the work already performed by POS taggers to identify simple phrases by recognizing the sequences of POS tags. The MADAMIRA morphology system [46] is adopted for tagging Arabic sentences as a preprocessing step, including providing the text's morphological features. MADAMIRA is one of the best Arabic taggers, in addition to offering morphological word analysis. To study the words of a sentence more precisely and clearly during the syntactic parsing stage, it is necessary to rely on the morphological analysis of the words. This is provided by the morphological system MADAMIRA in addition to its ability to label words. This process, along with the remaining components, is illustrated in the system architecture shown in Fig. 2.

This work serves as a basic building block for six Arabic sentence constructions: attributed, additional, statement, conjunctive, intermingled (admixture), and numerical composition. The focus here is on the most active composition in utterance —attributed composition— noting that a composition is a phrase of two or more words that conveys a benefit.

For instance, a composition may express a complete meaning (e.g.,النجاة في الصدق – Survival is in the honesty), or an incomplete meaning (e.g., ضوء الشمس – Sunlight), which is called a fragment sentence. A sentence with the attributed composition reflects a judgment made about something. As in the example سمير ذكي – Samir is intelligent, it expresses a ruling about Samir's attribute of intelligence. Here, the I (al-musnad-ilaihi) is "سمير – Samir," and the P (al-musnad) is "ذكي – intelligent."

Our study of these compositions is useful for parsing the sentence to determine whether it is complete, in addition to contributing to semantic interpretation, which evaluates whether the sentence conveys full meaning.

The main factor in forming constituent boundaries is the orthographic and linguistic properties of the words. Therefore, the decision tree technique is suitable for building such models. Most text-parsing algorithms, such as Shift-Reduce, rely on this factor during chunking. For example, a nominal phrase is a group of words not separated by a delimiter (e.g., a preposition, verb, relative noun, etc). In other words, a cascade of nouns, definite or indefinite (undefined nouns typically placed at the beginning of the phrase), constitutes nominal phrases. The following rule, one of the adopted grammatical rules, illustrates this:

S → NN * [[DTNN | DTNNS] | [NNP | NNPS]]

3.1. I and P detection

The grammatical and lexical characteristics of each noun determine whether it belongs to one phrase or should be separated from the other. In general, the nominal analytical unit is a sequence of nouns separated by a reductive term, which may be a verb, preposition, relative pronoun, jointed pronoun, an indefinite noun following a definite noun, or a proper or indefinite adjective. This task falls under the scope of NLP fields, particularly Base Phrase Chunking (BPC) [29], which separates and segments a sentence into its sub-constituents, such as nouns, verbs, and prepositional phrases.

Generally, the inchoative (I) is a definite noun, and the predicate (P) is an indefinite noun. Several points regarding this structure that should be noted:

First, when an indefinite noun follows a definite noun, it is typically a predicate that follows an inchoative.

Second, if a definite noun is followed by an indefinite noun, the sentence is composed of an inchoative followed by a predicate.

Third, the inchoative is always definite in the following three cases. The first bold phrase represents I, whereas the second bold phrase represents P:

(1) If it is a proper noun, then: Example:

“Ahmed مُجد طالبٌ” (Ahmed talib mujed)

"Ahmed is a hardworking student."

“مصر واحة أمن” (Masr Wahatul Amn)

"Egypt is a land of peace."

(2) If it is defined by “Al-ال” (the definite article):

Example:

“القناعة كنزٌ لا يفنى” (Al-Qana’a Kanzon La Yafna)

"Contentment is a sustainable treasure."

“القليل النافع خيرٌ من الكثير الضار” (Al-Qalil Annafe’ Khairon men Al-Kathir Al-Dhar)

"A little useful is better than a lot harmful."

(3) If it is added to a definite noun (i.e., the iDafa construction):

Example:

“شهر رمضان شهرٌ كريم” (Shahru Ramadan Shahron Kareem) "Ramadan is a holy month."

“نصرةُ المؤمن حق” (Nusratul Mo’men Haq)

"Helping the believer is a duty."

Fourth If the inchoative is a descriptive indefinite noun, the predicate will often be a clause or a quasi-phrase:

Example:

“رجلٌ كريمٌ زارنا” (Rajulon Kareemon Zarana)

"A generous man visited us."

Here, the inchoative is Rajul, and the predicate is Zarana, which in this case is a verbal sentence.

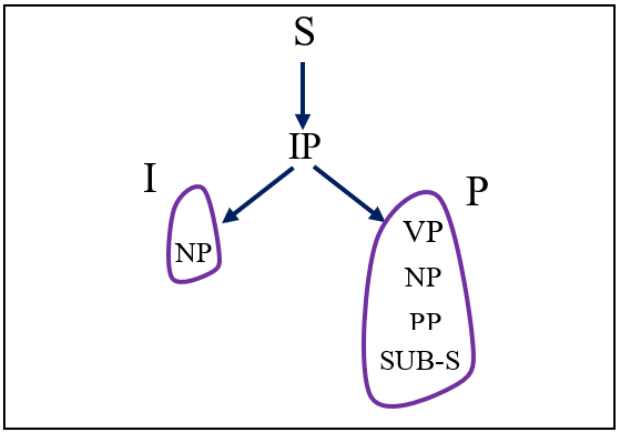

Figure 3. Syntactic structure of I and P

Our system processes at nominal sentence, which consists of two parts, as previously mentioned: P_1 (Inchoative phrase) and P_k(Predicate phrase). these two components can be formalized as Equations (1) and (2).

$$I = p_1 \tag{1}$$

$$P{=}p_2 \mid p_3 \mid p_4 \mid .. . \mid p_k \tag{2}$$

In general, p_1 constitutes an inchoative as a NP phrase. Predicate *P* is formed by one of the remaining sentence phrases, $p_2 \mid p_3 \mid p_4 \mid .. . \mid p_k$ which may be an NP or another functional category (PP, SBAR, etc.)(Fig. 3). p_1 was selected for individual analysis and separated from the rest of the sentence phrases. NP boundary determination plays a crucial role here. Typically, the parser relies on lexical analysis to identify phrases. The element that splits at phrase may be an indefinite noun, an adjective placed after certain nouns, verbs, prepositions, relative pronouns, jointed pronouns, or any syntactic category described as a splitter.

In Arabic grammar, NP is defined as a sequence of definite or indefinite nouns at the beginning of a phrase, optionally accompanied by pre-adjectives or post-adjectives.

Example:

قرارات الرئيس الأمريكي دونالد ترامب حمقاء *qrArt Alr}ys Al>mryky dwnAld trAmb HmqA*

"**Decisions** of the American president Donald Trump are **stupid**." Here, the long phrase قرارات الرئيس الأمريكي دونالد ترامب "Decisions of the American president Donald trump" represents p_1, while the term حمقاء "stupid" represents p_2.

The inchoative is identified by a term in p_1, which serves as the head of that NP. The system module performs this task based on syntactic features, incorporating Chomsky's syntax theory [47] and X-head (X-bar) theory [48].

Once the inchoative is detected, the predicate analysis begins to default except in special cases where the predicate precedes the inchoative [1], which are not addressed in this work.

The remaining phrases ($p_2 \mid p_3 \mid p_4 \mid .. . \mid p_k$) are evaluated and weighted to select the most likely candidate that represents the predicate P_p. Several syntactic rules are adopted for this purpose: The first term in $P_p(t_0)$ identifies the predicate type,

- If t_0 in p_2 is a verb, then the predicate is classified as a *verbal sentence*.
- If t_0 is a preposition, the predicate is a *semi-sentence*.
- If t_0 is a noun, the predicate is a *single noun phrase* or *nominal sentence*, and so on.

The Predicate may also be a *relative sentence,* when t_0 is a relative pronoun, or may be delayed across multiple clauses. This type of predicate is considered to be one of the most difficult to detect because of its complex structure. However, detecting a single predicate is a trivial task. The predicate that is a nominal sub-sentence, requires extracting the phrase head, similar to how the inchoative is detected. **Eventually, the extracted term that has been identified is considered the core output of this module - what is called the predicate**.

Fig. 3 indicates that the predicate is placed directly after the inchoative, with no separation between them. However, the varied structures of Arabic sentence may change this sequence and insert a complementary phrase between the inchoative and the predicate (a phrasal separator between I and P). The presence of a complementary phrase increases the complexity of the system when it precedes a predicate. This is manageable when only one complementary phrase separates the two. However, certain sentence structures allow for multiple complementary phrases, which in turn increase the system's exhaustion in tracking all phrases until reaching the one that includes the predicate P_p. See the example illustrated in Fig. 4.

"Dr. **Ghaleb,** who taught us the artificial intelligence course last year at Sana'a University, is a **sea**.".

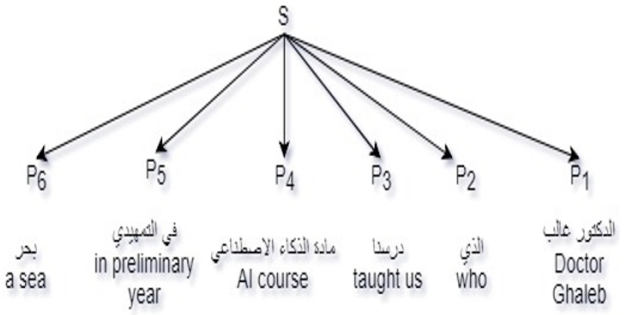

Figure 4. Phraser output example

In this case, the system must trace all intervening phrases to reach the predicate, phrase p_6 بحر "sea". Multiple distinct phrases types separate the inchoative p_1 of predicate p_6, including *relative phrase* p_2*, verbal phrase* p_3*, nominal phrase* p_4*, and prepositional phrase* p_5.

The system also includes a dedicated unit to process the pseudo-verbs and abolished particles, such as إنّ وأخواتها وكان وأخواتها *Inn wa Akhawteha, Kan wa Akhawteha* among others. These exceptions are treated as though they do not exist in the structural analysis. This is because their effect is limited to changing nominative and accusative cases of nouns [1].

For example: كانت الشمس مشرقة

kAnat Al$amos mu$oriqap

"The sun was shining"

The pseudo-verb *kAnat* is excluded from the sentence and only the phrase of الشمس مشرقة is considered for parsing.

3.2. I and P annotation

Some tagsets are considered standard sets in syntactic annotation; the most common syntactic tagset is PATB. In this study, two additional tags were proposed to annotate components I and P.

- **I** for *inchoative*
- **P** for *predicate*

These tags are attached to the corresponding sentence components. The output document retains the original sentence structure, but with the annotated information appended. The following example illustrates how a sentence appears in the annotated output.

Arabic: نجاح الطالب المتميزتسهم في بناء المجتمع ثمرة .

Transliteration: najAH AlTAlb Almtmyz vmrp tusohim fy bnA' AlmjtmE.

Gloss: An intelligent student's **success** is a **product** that contributes to society development.

Output: [NN-1/***I***, DTNN-2, DTJJ-3, NN-4/***P***, VBP-5, IN-6, NN-7, DTNN-8, PUNC-9]

In the example above:

- *NN-1* represents the inchoative ***I***

- *NN-4* represents the predicate ***P***.

4. EXPERIMENT

The evaluation techniques used to assess parsing performance varied across the studies. This study adopted a technique that compares the system's output against a gold standard dataset, such as the PATB gold standard. Because I and P detection is considered a task aligned with linguistic parsers, it is appropriate to apply parser evaluation metrics such as the PARSEVAL metric [49]. Accordingly, F-score remains a widely used general-purpose metric in most information system evaluations [50]. Standard evaluation measures, Precision, Recall, and F-measure, were used to evaluate the performance of the proposed tool. Precision was calculated using the following equation:

$$\textbf{\textit{Precision}} = \frac{number\ of\ correct\ constituents\ in\ P}{number\ constituents\ in\ P}$$

Recall is calculated by the following equation:

$$\text{Recall} = \frac{\text{Number of correct constituents in } P}{\text{Number constituents in } T}$$

The F-measure is calculated by taking the harmonic mean of both ***Precision*** and ***Recall*** as follows:

$$F\text{-}Measure = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Regarding the evaluation dataset used, a large set of sentences was cropped from the Prague Arabic Dependency Treebank (PADT) standard corpus [51]. The PADT corpus consists of morphologically and analytically annotated newswire texts in Modern Standard Arabic (MSA). These texts originate from the Arabic Gigaword corpus and plain data from PATB 1 and PATB 2. The PADT 1.0 distribution comprises over 113,500 tokens annotated analytically and is enriched with disambiguated morphological information. From the PADT categories, AFP, UMH, and XIA were used to construct the evaluation dataset. The resulting dataset includes 81,362 words across 2,500 sentences.

We relied on the Java environment as the primary programming language for testing the system, since most preprocessing tools are also written in Java.

From the initial experimentation, we noticed significant superiority of the system in correctly classifying the inchoative in almost every sentence in the dataset. Only a few sentences failed to be correctly processed. The primary issue was the identification of the X-head lexeme, which represents the inchoative within its extracted phrase. This difficulty typically occurs in phrases that begin with a title or surname, or sentences that start with pre-modifying expressions (e.g., Dr. Ali, company manager Ms. Nasreen, etc.).

Some of the failures observed during the experiment were not caused by the system itself, but by limitations in the preprocessing tools. Any errors originating from earlier layers inevitably propagated into the parsing stage.

Regarding for the shortcomings of our system, most occurred in long sentences, particularly those in which the predicate was delayed until the end of the sentence and interrupted by multiple intermediate complementary phrases.

Overall, the performance of the I and P tool was very encouraging, largely because of the elegance of the model (see Table 1). A comparative evaluation can be conducted against systems such as the Stanford parser, provided that the evaluation is restricted to first-level branching only. In other words, both systems are compared at the point where the parse tree initially splits into two branches: the left branch (I) and right branch (P) (see Fig. 3).

Table 1. Evaluation results of I and P tool vs. Stanford system

Model	Precision	Recall	F-measure
I & P tool	94.30%	91.75%	93.00%
Stanford	88.62%	82.48%	85.43%

The primary limitations of both systems lie in handling the unusual syntactic structures of I and P, and long-distance predicates relative to the inchoative.

5. CONCLUSION

This paper briefly discusses I and P annotation tool, outlining the procedures and methods assumed for both development tasks. Identification is a fundamental step that precedes any annotation process. Therefore, the proposed model leveraged the advantages of the shallow parsing approach.

The role of the tool is to label the two main components of the Arabic nominal sentence using two designated tags. This paper presents an easy-to-implement and expressive formalism to accomplish this task accurately.

This work paves the way for building complete syntactic parsers that can disambiguate all types of Arabic sentences, including the various syntactic structures involved. Furthermore, it supports conceptual analysis, which enhances semantic parsers' understanding and interpretation of sentence meaning. This is especially beneficial for parsing based on the core structure of nominal sentences and rhetorical analysis purposes.

An empirical evaluation was conducted using a dataset extracted from the PADT corpus, comparing our system with the Stanford parser (one of the most well-known Arabic parsers). The findings demonstrate that our approach yields substantial improvements in parsing nominal Arabic sentences. The evaluation results indicate that the system effectively achieved its intended objectives.

However, this tool exhibits some limitations, particularly when handling unusual I and P structures. Therefore, future studies can address these deficiencies to further enhance the robustness and accuracy of the system.

REFERENCES

[1] M. A. Al-Sheikh and A. A. K. I. Al-Doktor, *Jami' al-Durus al-Arabiyya*. Dar al-Kutub al-'Ilmiyya, 2016.

[2] G. Nelson and S. Greenbaum, *An Introduction to English Grammar*. Routledge, 2015.

[3] M. M. A. Shquier, M. S. Atoum, and O. M. A. J. N. T. I. Shqeer, "Arabic to english machine translation," in *Proceedings of [conference name not specified]*, 2017, p. 118.

[4] I. Zitouni, *Natural Language Processing of Semitic Languages*. Springer, 2014.

[5] N. Al-Horais, "Arabic verbless sentences: Is there a null vp?," 2006.

[6] M. A. Idrees, "The subject in causation and attribution: A study on function and meaning," 2017.

[7] J. Berant and P. Liang, "Semantic parsing via paraphrasing," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2014, pp. 1415–1425.

[8] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55–60.

[9] S. J. Abney, "Partial parsing via finite-state cascades," *Nat. Lang. Eng.*, vol. 2, no. 4, pp. 337–344, 1996.

[10] S. M. A. El-Morsy, M. Hussein, H. M. Mousa, and C. Engineering, "Arabic open information extraction system using dependency parsing," *Int. J. Eng. Comput. Sci.*, vol. 12, no. 1, pp. 541–551, 2022.

[11] C. Dyer, M. Ballesteros, W. Ling, A. Matthews, and N. A. Smith, "Transition-based dependency parsing with stack long short-term memory," in *arXiv preprint arXiv:1505.08075*, 2015.

[12] N. Ababou, A. Mazroui, R. Belehbib, and L. R. a. Evaluation, "From extended chunking to dependency parsing using traditional arabic grammar," *Lang. Resour. Eval.*, vol. 57, no. 3, pp. 1011–1043, 2023.

[13] A. Bouziane, D. Bouchiha, and N. Doumi, "Annotating arabic texts with linked data," in *Proceedings of the 2020 4th International Symposium on Informatics and its Applications (ISIA)*, IEEE, 2020, pp. 1–5.

[14] D. Zeman et al., "Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies," in *Proceedings of the CoNLL 2018 Shared Task*, 2018, pp. 1–21.

[15] N. G. Hammouda and K. Haddar, "Arabic nooj parser: Nominal sentence case," in *International Conference on Automatic Processing of Natural-Language Electronic Texts with NooJ*, Springer, 2017, pp. 69–80.

[16] N. Habash, R. Roth, O. Rambow, R. Eskander, and N. Tomeh, "Morphological analysis and disambiguation for dialectal arabic," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 426–432.

[17] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, D. Nouvel, and I. Sciences, "Arabic natural language processing: An overview," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 33, no. 5, pp. 497–507, 2021.

[18] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Trans. on Asian Lang. Inf. Process.*, vol. 8, no. 4, 2009.

[19] 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*, Association for Computational Linguistics, 2004, pp. 2–9.

[20] T. Buckwalter, "Buckwalter arabic morphological analyzer version 1.0," Linguistic Data Consortium, University of Pennsylvania, Tech. Rep., 2002.

[21] M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki, "The penn arabic treebank: Building a large-scale annotated arabic corpus," in *Proceedings of the NEMLAR Conference on Arabic Language Resources and Tools*, vol. 27, Cairo, 2004, pp. 466–467.

[22] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," *Comput. Linguist.*, vol. 31, no. 1, pp. 71–106, 2005.

[23] R. M. Duwairi and I. Qarqaz, "Arabic sentiment analysis using supervised classification," in *Proceedings of the 2014 International Conference on Future Internet of Things and Cloud (FiCloud)*, IEEE, 2014, pp. 579–583.

[24] R. M. Duwairi, "Sentiment analysis for dialectical arabic," in *Proceedings of the 2015 6th International Conference on Information and Communication Systems (ICICS)*, IEEE, 2015, pp. 166–170.

[25] K. Dukes and N. Habash, "Morphological annotation of quranic arabic," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, 2010.

[26] Y. Marton, N. Habash, and O. Rambow, "Improving arabic dependency parsing with lexical and inflectional morphological features," in *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, Association for Computational Linguistics, 2010, pp. 13–21.

[27] S. Green and C. D. Manning, "Better arabic parsing: Baselines, evaluations, and analysis," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, Beijing, China, 2010.

[28] A. M. Bsharat, "Methods of verbs negation in standard arabic and dialects: Statistical applications in analyzing morphological and syntactic changes," 2024.

[29] M. T. 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*, Association for Computational Linguistics, 2007, pp. 89–96.

[30] D. M. Bikel, "Intricacies of collins' parsing model," *Comput. Linguist.*, vol. 30, no. 4, pp. 479–511, 2004.

[31] A. Mouloudi, "Syntactic parsing of simple arabic nominal sentence using the nooj linguistic platform," in *Arabic Language Processing: From Theory to Practice, 6th International Conference, ICALP 2017, Fez, Morocco, October 11–12, 2017, Proceedings*, vol. 782, Springer, 2018, p. 244.

[32] Y. Zaki, H. Hajjar, M. Hajjar, and G. Bernard, "Towards the development of a statistical parser of arabic language," in *Proceedings of the 2017 Computing Conference*, IEEE, 2017, pp. 85–87.

[33] C. Gómez-Rodríguez, I. Alonso-Alonso, and D. Vilares, "How important is syntactic parsing accuracy? an empirical evaluation on rule-based sentiment analysis," *Artif. Intell. Rev.*, 2017.

[34] N. Ababou, A. Mazroui, R. Belehbib, and I. J. o. I. S. a. Applications, "Parsing arabic nominal sentences using context free grammar and fundamental rules of classical grammar," *Int. J. Inf. Sci. Appl.*, vol. 9, no. 8, p. 11, 2017.

[35] 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)*, IEEE, 2015, pp. 80–83.

[36] R. Collobert and B. Bai, "Method and apparatus for full natural language parsing," US Patent 8,874,434, 2014.

[37] Y. Marton, N. Habash, and O. Rambow, "Dependency parsing of modern standard arabic with lexical and inflectional features," *Comput. Linguist.*, vol. 39, no. 1, pp. 161–194, 2013.

[38] S. Al-Ghamdi, H. Al-Khalifa, and A. Al-Salman, "Fine-tuning bert-based pre-trained models for arabic dependency parsing," *Appl. Sci.*, vol. 13, no. 7, p. 4225, 2023.

[39] S. Petrov, L. Barrett, R. Thibaux, and D. Klein, "Learning accurate, compact, and interpretable tree annotation," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2006, pp. 433–440.

[40] S. Alqrainy, H. Muaidi, and M. S. Alkoffash, "Context-free grammar analysis for arabic sentences," *Int. J. Comput. Appl.*, vol. 53, no. 3, 2012.

[41] K. Dukes, "Statistical parsing by machine learning from a classical arabic treebank," *arXiv preprint arXiv:1501.00000*, 2015, Preprint.

[42] Y. Zaki, H. Hajjar, M. Hajjar, and G. Bernard, "A survey of syntactic parsers of arabic language," in *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies*, ACM, 2016, p. 31. DOI: 10.1145/3010089.3010105.

[43] S. Petrov and D. Klein, "Improved inference for unlexicalized parsing," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 2007, pp. 404–411.

[44] A. T. Al-Taani, M. M. Msallam, and S. A. Wedian, "A top-down chart parser for analyzing arabic sentences," *Int. Arab. J. Inf. Technol.*, vol. 9, no. 2, pp. 109–116, 2012.

[45] S. Majdi et al., "Morphologically-analyzed and syntactically-annotated quran dataset," *Data Brief*, vol. 58, p. 111 211, 2025. DOI: 10.1016/j.dib.2025.111211.

[46] A. Pasha et al., "Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic," in *Proceedings of LREC 2014*, vol. 14, 2014, pp. 1094–1101.

[47] N. Chomsky, *Aspects of the Theory of Syntax*. MIT Press, 2014.

[48] T. Stowell, "Subjects, specifiers, and x-bar theory," in *The Syntax of Parameterization*, Academic Press, 1989, pp. 232–262.

[49] E. Black et al., "A procedure for quantitatively comparing the syntactic coverage of english grammars," in *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19–22, 1991*, 1991.

[50] A. De Sitter, T. Calders, and W. Daelemans, "A formal framework for evaluation of information extraction," University of Antwerp, Tech. Rep., 2004, Working paper.

[51] J. Hajič, O. Smrž, P. Zemánek, J. Šnaidauf, and E. Beška, "Prague arabic dependency treebank: Development in data and tools," in *Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools*, 2004, pp. 110–117.