HPSG ANALYSIS OF TYPE-BASED ARABIC NOMINAL DECLENSSION

Mahmudul Hasan Masum, Muhammad Sadiqul Islam, M. Sohel Rahman, Reaz Ahmed
Department of Computer Science & Engineering,
Bangladesh University of Engineering and Technology, Bangladesh

Abstract: Semitic languages exhibit rich morphological operations, which can generate a myriad of inflected lexemes. Modeling the morphological effect of a rich declension system is a challenging problem and is essential for intelligent and automated processing of Arabic language. To capture these dimensions of rich morphology by natural linguistic processing, Head-driven Phrase Structure Grammar (HPSG) seems to be the best choice. In this paper, we propose a novel HPSG representation which includes the morphological, syntactical and semantic features for Arabic nominals. We analyze the declension system of Arabic nouns and design lexical type hierarchy by which the declension type of any noun lexeme will be determined by the lexical type. We also show construction rules to capture the morphological and syntactic effect of declension type dynamically. Finally, we have implemented the lexical type hierarchy, Attribute Value Matrix (AVM) and construction rules in TRALE platform to validate the proposed HPSG formalism. We believe our work in this paper shall remain as a milestone in the complete HPSG analysis of Arabic language.

Keyword: Arabic Declension, Head-driven Phrase Structure Grammar, Constraint-based Grammar, Arabic Verbal Declension

1 Introduction
Semitic languages like Arabic, Amharic and Hebrew exhibit rich morphological operations for construction of lexicons. We can have a large coverage of vocabulary in these languages by computational linguistic modeling of their morphology. In this paper we focus on Arabic for morphological analysis. The diversity and importance of Arabic nominals is broader than that of their counterparts in other languages. Modifiers, such as adjectives and adverbs, are treated as nominals in Arabic. Like others, Arabic nominals show two types of morpho-syntactic operations - derivation and inflection.

Inflection refers to the variation in the form of a word, typically by means of an affix that expresses a grammatical contrast which is obligatory for the stems in some given grammatical context. Formation of dual or plural from singular, formation of feminine gender from masculine and declension are some uses of inflection. Declension is the process of disambiguating the grammatical roles of words by slightly changing their end vowels. Arabic declension has some unique features. In this paper we limit our discussion of inflection on declension only.

For modeling Arabic morphology, we have chosen Head-driven Phrase Structure Grammar (HPSG) (11) which is an attractive tool for capturing complex linguistic constructs. HPSG is very suitable for Natural Language Processing (NLP) as it integrates all the essential linguistic layers (Phonology, Morphology, Syntax, Semantics, and Context etc.) of NLP. We have worked on Sign Based Construction Grammar (SBCG) (15) version of HPSG.

In this paper, we propose and analyze the HPSG constructs required for capturing the syntactic and semantic effects of rich morphology of Arabic nominals. Our contributions in this paper are as follows:

- We propose the HPSG type hierarchy of Arabic noun lexeme. It completes the type hierarchy proposed in (9).
- The proposed type hierarchy maps noun lexeme types to declension types. We devise an algorithm for this mapping. This mapping is used for typed based construction rule and replaces the usage of classical sixteen categories of noun for declension.
- We formulate the structure of attribute value matrix (AVM) for Arabic noun.
- We propose lexical construction rules for particular declension types of nouns to eliminate lexical entries.
- We verify the partial type hierarchy and construction rules in a lexical compiler named TRALE (12).
The rest of the paper is organized as follows. Section 2 presents preliminaries of Arabic morphology and related works. Section 3 presents our contributions. Section 4 briefly discusses the TRALE implementation of our work. We draw conclusion in Section 5.

2 Background
In this section we discuss the topics that serve as a background of the rest of the paper. Section 2.1 describes concatenative morphology with emphasis on declension. Section 2.2 discusses the state of research work on HPSG modeling of Arabic.

2.1 Arabic Concatenative Morphology
In this paper, we mainly focus on declension in concatenative morphology.

2.1.1 Arabic Declension
Grammatical declension is known in Arabic as I’rab (- irab). By definition, i’rab is the process of disambiguating the grammatical roles of words by slightly changing their end vowels. According to declension, we can classify Arabic lexemes into two types - declinable (- mu’rab) and indeclinable (- mabny). If a word experiences declension it is called declinable. If a word experiences declension it is called declinable, and if it does not experience declension, or experiences it but does not show it, it is called indeclinable.

In Arabic, there are three grammatical states of noun, i.e., three cases (14) as follows:
- The state of ( - ra’ / - raf) - nominative case
- The State of ( - nasab) - accusative case
- The state of ( - jarr) - genitive case

2.1.2 Role of declension in Arabic grammar
Declension plays more significant role in Arabic than most other languages. Because, in Arabic, subject, object, predicate everything is determined by the end vowel. As an example, in English, subject and object are determined by the sequence of words in a sentence. For example, consider the two sentences, “Zayeed beat him” and “He beat Zayeed”. In the former, ‘Zayeed’ is subject and ‘him’ is object whereas in the latter, ‘He’ is subject and ‘Zayeed’ is object. But in Arabic, زَيْدٌ (zaydun) is subject as it is ended with short form of (waw). Thus, end vowels imply grammatical cases for nominals in Arabic.

2.1.3 Declension types
Arabic declension can be classified according to three dimensions which are described below.

Visibility of declension: This dimension determines whether all possible forms of vowels are explicitly shown on the final letter of a noun for all three cases. Along this dimension, declension can be classified into three categories: Visible declension, Partially invisible declension and Completely invisible declension. There are nouns which cannot explicitly show one or more vowels on their final letters. In this case, it is assumed that if the noun were able to show vowels on its final letter, then it would do so. So, here assumed vowels are used to reflect the declension type. For example, حَسْنٰٓ (husna) cannot show (u) and (i) on its final letter because of the presence of (Ya maqṣura). So, حَسْنٰٓ (husna) follows invisible declension. On the other hand (hasanun) can show all short forms of vowels on its final letter. So, (hasanun) follows visible declension. There are some declension types where some forms are visible and rests are invisible. We call it partially invisible.

Vowel form used to decline: This dimension determines whether the vowel used for declension is in short or long form. There are some words which use (waw) to reflect the nominative case, (alf) to reflect accusative and ēp (ya) to reflect the genitive case.

Completeness of declension: This dimension determines whether all vowels are used in the declension forms. There are words which do not use all vowels to show all cases; rather same vowel is used to reflect accusative and genitive cases. For example, (masagidu) uses (a) to reflect both accusative and genitive cases. Here (i) is not used. For this reason this is an example of incomplete declension. On the other hand, (hasanun) is an example of complete declension as it can use all forms of short vowels to reflect the declensions. Notably, if a declension type is invisible, then the completeness dimension will not be applicable to that declension type.
<table>
<thead>
<tr>
<th>Partitions</th>
<th>Genitive</th>
<th>Accusative</th>
<th>Nominative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visibility</td>
<td>Visible declension</td>
<td>(hasanin)</td>
<td>(hasanan)</td>
</tr>
<tr>
<td>Invisable declension</td>
<td>(husna)</td>
<td>(husna)</td>
<td>(husna)</td>
</tr>
<tr>
<td>Vowel form</td>
<td>Declension with short vowel</td>
<td>(hasanin)</td>
<td>(hasanan)</td>
</tr>
<tr>
<td></td>
<td>Declension with long vowel</td>
<td>(jiy)</td>
<td>(ja)</td>
</tr>
<tr>
<td>Completeness</td>
<td>Complete declension</td>
<td>(hasanin)</td>
<td>(hasanan)</td>
</tr>
<tr>
<td></td>
<td>Incomplete declension</td>
<td>(masagida)</td>
<td>(masagida)</td>
</tr>
</tbody>
</table>

Table 1: Different dimensions of declension

Table 1 shows examples of the above three dimensions of declensions. Through inheriting different combination of these declension subtypes, nine ways can be found by which grammatical cases are represented. For example, visible complete declension with long vowel, partially invisible with long vowel, invisible declension with these short vowel etc. These nine declension types can be expressed as $T_{n1}$, $T_{n2}$, $T_{n3}$, ..., $T_{n9}$, which are shown in Table 2. In traditional Arabic, nouns are categorized into sixteen classes to reflect these nine declension types.

<table>
<thead>
<tr>
<th>Types</th>
<th>Noun class</th>
<th>Genitive</th>
<th>Accusative</th>
<th>Nominative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1</td>
<td>1. Triptote sound singular</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. Singular noun pseudo sound</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Triptote broken plural</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 2</td>
<td>4. Sound feminine plural</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 3</td>
<td>5. Diptote without prefixed by definite markness or not a mudaf in a Mudaf-Mudaf Ilayh sentence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 4</td>
<td>6. حم هن فم not towards first person singular number possessor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 5</td>
<td>7. Dual noun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8. and mudaf towards pronoun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9. and</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 6</td>
<td>10. Sound masculine plural</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11. Multiple of ten between twenty and ninety</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12. (plural of possessor)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 7</td>
<td>13. Noun mudaf towards first person personal pronoun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14. ending with ya maqsoora ( )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 8</td>
<td>15. Noun ending with</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 9</td>
<td>16. Sound masculine plural mudaf towards personal pronoun</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Noun classes according to 9 declension types

Some grammatical terms in Table 2 deserve some explanations. *Triptote* refers to words that take all three short vowel case endings, where each one differentiates a particular case. *Diptote* only
exhibits two case markers: (\textcircled{u}) for nominative and (\textcircled{a}) for both genitive and accusative. Sound nouns are those where end letters are consonants. For an unsound noun, end letter is a vowel (\textcircled{w}) or (\textcircled{y}). Pseudo sound ends with (\textcircled{w}) or (\textcircled{y}) and there is sakin (absence of short vowel on a letter) on the letter before the last letter. It is actually unsound, but it follows declension like a sound noun. Mudaf-Mudaf Ilayh is a possessive phrase of two terms, where the first term (Mudaf) can be interpreted as belonging to the second term (Mudaf Ilayh).

2.2 Related Works

HPSG analysis and modeling of Arabic grammar is a comparatively new area of research. Its precious morphology attracted several series of research projects (1; 4; 16). These research projects are mainly based on development of toolkit for Arabic morphological analysis.

In 2005, Melnik presented a comparison of the implementation platform of HPSG (10). She discussed the advantages and disadvantages of TRALE (12) and Linguistic Knowledge Building (LKB) (6) as implementation platforms of HPSG. This paper is very useful to choose the implementation platform of HPSG.

In 2006, an HPSG analysis of broken plurals and gerunds were presented (8). Main assumption in that work revolves around the Concrete Lexical Representations (CLR) located between an HPSG type lexicon and phonological realization. Here, HPSG sign was represented using CLR function and not by AVM. This function put more emphasis on phonology instead of morpho-syntactic operations.

HPSG modeling of Arabic triliteral strong verbs was proposed in 2008 (2; 3). The authors show regular morphology of Arabic verbs in these papers. They designed the SBCG AVM of Arabic verbs. They also designed several verb lexeme constructions and morphologically complex predicates (MCP). But they did not consider the morphological derivation of verbal nouns. Also, they did not give any distinct way to implement the construction proposed in their works.

An in-depth analysis of declensions in Arabic nouns has been presented in (7). But they did not show or discuss any mapping of lexical type to declension type which is necessary to implement the declension phenomenon in HPSG. Islam et al. also showed HPSG analysis of Arabic nominals from the perspective of verbal nouns (9). There, the authors showed the morphology of eight types of verbal nouns and their construction from verbs only. Notably, declension was beyond the scope of their work. Thus, to the best of our knowledge, the rich morphology of Arabic nominals has not yet been explored in the literature. This motivates us to do research in this particular area, which is the focus of this paper.

3 HPSG Formalism

In this section we discuss the formalism of our research. We propose complete type hierarchy in Section 3.1. Section 3.2 describes the mapping of lexical types to declension types. Related algorithm is presented in Section 3.3. We give noun AVM in Section 3.4. Lexical construction rules for declension are proposed in Section 3.5.

3.1 Type Hierarchy

We can classify Arabic nouns based on several dimensions e.g., number, derivation, gender, declension and ending type.

The type hierarchy of Arabic noun lexemes is shown in Figure 1 based on these dimensions. Derivation is explored by (9). In fact, classification along derivation is not subjected to declension. Among these dimensions, end letter type deserves some explanation. As discussed in Section 2.1.3, first level classification along ending type is Sound (end letter is consonant) and Unsound (end letter is vowel). Unsound can be further classified to Ism magsoor (alif ending) and Ending with ya/waw. Subtype Ending with ya/waw can be further classified to Pseudo sound and Pure unsound.
3.2 Mapping type hierarchy to declension type

As discussed in Section 2.1.3, there can be nine possible declension types of noun lexemes. Figure 2 shows the mapping of these declension types from lexical type hierarchy. Since derivation has no effect on declension, dimension along derivation is not shown in Figure 2. This figure shows subtypes which are formed by multiple inheritance indicated by dotted lines. For example, \textit{tripote-sound-sg-noun-lex} is type of a noun lexeme which is a subtype of triptote, sound and singular. Declension type for each sub type is shown inside parenthesis. Declension type of lexical type \textit{tripote-sound-sg-noun-lex} is $T_n1$ which indicates that lexical type \textit{tripote-sound-sg-noun-lex} follows declension type 1.
For simplicity, we have not mentioned lexical type for other subtypes though we have shown corresponding declension types. We can also observe that there are three lexical types which follow declension type 1. These are: triptote-sound-sg-noun-lex, triptote-pseudo-sg-noun-lex and triptote-broken-pl-noun-lex. Notably, Tₙ₄ and Tₙ₉ declension types are only found in phrase levels. That is why, these are not shown in this mapping.

3.3 Algorithm to find declension type of noun lexeme
In section 3.2, we have shown mapping of declension type from lexical type hierarchy. From this mapping, we devise an algorithm which determines the declension type of a noun lexeme. The flowchart of this algorithm is shown in Figure 3. This will be an offline method by which declension type of a lexeme will be identified.

From this flowchart, it is clear that each noun lexeme must have a declension type. This is because, at each decision maker, the noun lexemes are subjected to two new partitions. Thus, all noun lexemes are completely partitioned. In other words, every noun lexeme must have a declension type which can be determined from this flow chart.

3.4 AVM of Arabic Nouns
We have modified the AVM proposed in (9) to capture declension phenomenon. Our proposed AVM is illustrated in Figure 4. Here MORPH is to denote morphological features. It contains three features: ROOT, SKELETON and DEC. ROOT contains list of root letters and it bears same meaning as in (9). SKELETON contains not only stems but also inflected words. It is a sequence of morphological objects which are phonologically realized. It will include both lexical formatives and affixes. For example, for Arabic word (katiban), value of SKELETON will be <kaatib+an>. For (katibun), value of SKELETON will be <kaatib+un>. This feature is very much similar to the feature SKELETON used in (13).

DEC feature is a significant modification. This feature indicates declension type as discussed in Section 2.1.3. Declension type is a morphological feature. Hence DEC is placed under MORPH. It determines how the end vowel of a noun lexeme changes to reflect its case. The change of end vowel changes the form of a lexicon. As discussed in Section 2.1.3, there exists nine possible ways in which grammatical cases can be represented on an Arabic noun. So for declinable noun, value of the DEC feature will be one of Tₙ₁, Tₙ₂, Tₙ₃, …, Tₙ₉, corresponding to the nine declension types presented in Table 2. The value of this DEC feature can be determined from type hierarchy mentioned in Figure 2. For indeclinable nouns, the value of the DEC feature will be none.

3.5 Construction rules for declension
In Table 2 shows different types of declension follow different case markings. Section 3.2 shows different types of lexemes follow different declension types. Based on the mapping, for a particular type of lexeme, we develop construction rules to construct accusative and genitive forms from nominative lexemes. In this section, we present construction rule to construct accusative lexemes by Tₙ₁ declension.

A construction rule to construct the accusative form from a nominative lexeme following Tₙ₁
declension is shown in Figure 5. Here MTR (mother) is in accusative case and DTRS (daughter) contains only one lexeme which is in nominative case. MTR (accusative) is formed from DTRS (nominative) using $T_n1$ declension.

As in Table 2, case marking of the nominative case is $(un)$. But for accusative, the case marking is $(an)$. Our construction rule captures this change by changing the SKELETON feature under MORPH (morphology) from DTRS to MTR. SKELETON of MTR is ended with $-an$ but the same of DTRS with $-un$. Another change between MTR and DTRS lies in CASE under CAT (category) to denote the case of the lexemes. All other features are same for the two.

For example, we have given lexical entry of noun lexeme kaatibun which is in nominative case. But we have not given lexical entry for its inflected form kaatiban (accusative) or kaatibin (genitive). These will be identified by lexical construction rules.

5 Conclusion

This paper shows HPSG modeling of Arabic nominal declension. It is mainly focused on type hierarchy to declension type mapping and construction rules to eliminate lexical entries. We have proposed Arabic AVM for noun. We have devised algorithm to identify declension type of a noun which also proofs completeness of 9 types of declension and hence completeness of 16 nominal
categories. We have proposed construction rules to capture nominal declension for different cases based on different declension types. We have implemented all proposed construction rules in TRALE platform which not only helped us to formalize correct construction rules but also proves the correctness of the construction rules. We sincerely hope that our work in this paper shall remain as a milestone in the complete HPSG analysis of Arabic language.

References


