Chapter Five
HEAD-DRIVEN PHRASE STRUCTURE GRAMMAR AND ARABIC MACHINE TRANSLATION

5.0 Introduction

Natural Language Processing (NLP) mainly covers four processing levels: lexical, syntactic, semantic and Pragmatic treatment. Syntactic analysis is considered as the fundamental phase for other stages of language treatment. Natural language processing has many applications, where syntactic analysis is necessary such as machine translation, human-machine dialogue system and grammatical errors correction.

HPSG is highly lexicalized, non-derivational constraint-based, surface oriented grammatical architecture developed by Carl Pollard and Ivan Sag (Pollard & Sag, 1994). HPSG is a lexically based theory of phrase structure, so called because of the central role played by grammatical heads and their associated complements. Roughly speaking, heads are linguistic forms either words or phrases that make syntactic and semantic restrictions on the phrases. Depending on the type of phrase, the head is identified, for example, noun (N) is the head of the noun phrase (NP), verb (V) is the head of verb phrase (VP), and preposition (P) is the head of prepositional phrase (PP) and so forth.

Widely speaking, HPSG is highly appreciated in Natural Language Processing because it makes use of the best idea from its predecessors – Government and Binding theory (GB), Generalized Phrase Structure Grammar (GPSG), and Lexical Functional Grammar (LFG). And, in its detailed analysis it combines the different linguistic layers (Phonology, Morphology, Syntax, semantic, pragmatics, context etc.)

HPSG is one of the most successful grammars to process natural languages especially to process syntactic and semantic aspects but it has inadequate coverage on morphological construction especially for nonconcatenative morphology (Islam et al., 2010).

In fact, few works had been explored in Arabic syntactic analysis and have constructed grammars dealing with particular Arabic phenomena, (e.g., verbal
systems and nominal sentences). Arabic is a complex language with highly inflected morphology and complicated syntax, and there are several criteria to categorize words and thus it is a difficult task to represent their hierarchical type and grammars that cover all the Arabic specificities. Thus some modifications are necessary to HPSG to cover Arabic specificities. HPSG modeling of Arabic morphology is still a new area of research. There are few works dealing with Arabic morphology. Which have been extended to cover the Arabic verbal morphology such as derivational forms of Arabic verbs. A new feature MORPH is proposed and contains three sub-features: TYPE, ROOT and MEASURE. These variables are used for agreement purpose also.

![Figure 5.1 Standard SBCG Type Hierarchy](image)

### 5.1 HPSG Components

HPSG rests on two essential components:

a. A set of Attribute Value Matrix (AVM), to represent lexical entries features. Attributes are represented as linguistic objects. The attribute can have two values either atomic or complex i.e. function. Atomic values refer to the simple feature structures that represent the terminal linguistic attributes. On the other hand, Functions refer to the complex feature structures which are
described using an attribute value matrix (AVM) and can include other feature structures as their feature values.

b. And a set of Immediate Domination schemata (ID schemata), to describe syntactic phenomena (Haddar, e.t., 2010).

5.1.1 Features of HPSG

An AVM is composed of typed feature or components of structures which are paired with value. The features in HPSG are sorted where each node is labeled with a sort symbol. The sort symbol tells what type of object the structure is modeling; that is, there is one sort symbol for each basic type of construct (Pollard & Sag, 1994). Here, we adopted the system of features that is used in Sag, Wasow and Bender (2003). In general, these features refer to the linguistic information those talk about the words, for example the grammatical category of the word, its theta grid, and its level in the tree node.

Feature’s value can be any of the four possible types (Green, 1999):

- Atomic sort or single value
- A feature structure
- A set of feature structures
- List of feature structures

Each feature has a value and the attribute value pairs can also be represented by using Attribute value matrix (AVM). Feature structures have a variety of types. The first types are those indicate the word vs. phrase status of every constituent in a tree second is the features for a node. The attribute labels of each feature are determined by its sort depending on its ontological category. Highly speaking, the SORT WORD feature has the attribute labels which are classified into three main classes: the values of the features PHON, SYN, ARG-ST (argument structure) and SEM. The first feature PHON introduces phonetic information. SYN is the syntactic structure. ARG-ST (Argument structure) is a feature of sign which lists syntactico-semantic arguments. For example, for transitive verb donate has an ARG-ST list as follows: <NP, NP, PP>. Here first NP is verb’s subject, second NP is verb’s direct object and the last one is prepositional phrase (Islam et al., 2010). ARG-ST (argument structure) feature structures are theta grid features representatives and the SEM feature has
different values represented in the semantic feature properties of the node. SEM feature has two values INDEX and FRAMES. INDEX indicates the referent of an expression, in the case of noun phrase. The INDEX indicates the subject and in the case of verb phrase it indicates the verb. FRAMES indicate predications and its value includes a list of frames. A frame is an elementary scene in which certain semantic roles are specified and specific participants are assigned to them (Islam et al., 2010). For example, the participants of the cleaning frame are an actor (who does the cleaning) and the place (which is cleaned). SIT (situation) indicates verb index in a frame. The figure 5.2 below illustrates the SEM values.

![Figure 5.2 Example of Semantic Object](image)

Each feature is associated with a determined value as illustrate in the following figure:

![Figure 5.3: The General Structure of an AVM](image)
As illustrated in the figure 5.3 above each feature in the AVM has a determined type of information. The first feature is PHON which represents the sign’s sound content which serves as the basic block of the phonetic and phonological interpretation.

The second feature SYNSEM is a collection of both the syntactic and semantic information. SYN feature structure contains the CAT (category), VAL (valence) and MRKG (marking).

CAT is a complex feature which has a complex value (set of features) like CASE that indicates the case of nouns, MOOD which indicates the mood of verbs, VFORM which represents the morpho-syntactic category of a verb, AUX (auxiliary) which tells whether a verb is an auxiliary and VOICE for verbs. VAL represents the degree of saturation. MRKG has two values. It can be unmk (unmarked) or any one of these; that, whether, than, det etc. for example, that lesson is written. Here, the MKNG value of this sentence is that and the MRKG value of the lesson is written is unmk. The figure 5.4 bellow shows the SYN feature value.

![Figure 5.4: Example of Syntactic Object](image)

The SNSEM feature is divided into two other features: LOC and NON-LOC feature. The LOC feature has other features such as TETE and VAL. Again, the feature TETE represents the characteristics that describe the represented entry. On the other hand,
the feature VAL represented the compounds that are categorized by the represented entry. For the feature NON-LOC, it represents the relation between both the represented compounds and other compounds.

As illustrated in the figure 5.5 the SYNSEM feature again is subdivided into two features or attributes labels: LOCAL (LOC) and NON-LOCAL (NON-LOC). LOC information in turn is divided into CATEGORY, CONTENT, and CONTEXT features. The NON-LOC feature represents the relation between both the represented compound and other compounds. With reference to the linguistic connotation of the feature structure description, the value of category represents both the syntactic category of the word and the grammatical argument it requires. The CONTENT value contains context-independent linguistic information. It represents the semantic interpretation aspects of any phrase contains it. And the CONTEXT value in turn contains context-dependent linguistic information usually discussed under such rubrics as indexicality, presupposition, and/or conventional implicature (Pollard & Sag, 1994).

The CATEGORY value (CAT) has two attributes HEAD and SUBCAT. The HEAD value is the part of speech of the mentioned sign and it contains two sorts which
represent the different parts of speech substantive (subst) and functional (funct). Substantive sort in turn has four sub Sorts; nouns, verbs, adjective (adj), and preposition (prep). On the other hand, Functional sort has two sub Sorts, determiner (det) and Marker.

The CONTENT value and the CONTEXT value constitute the semantic contribution of the sing and the matters of reference. The CONTENT value of nominals both the nouns and their phrasal projections is called nominal-object (nom-obj), this feature has the attribute INDEX (IND). This feature (IND) refers to the reference marker and indices that the thematic roles are assigned. According to their mode of referring, INDICES are divided into three sub-Sorts referential (ref), there, and it.

Referential (ref) indices are used for semantically content nouns as well as for non-predicative prepositions (PPs). Whereas the indices there and it are used for the expletive (dummy or pleonastic) pronouns. Indices have three agreement features PERSON (PER), NUMBER (NUM), and GENDER (GEND). Nominal-objects again are classified into two sub Sorts nonpronoun (npro) and pronoun (pron). Pronoun (pron) is divided into two sub Sorts personal-pronoun (ppro) and anaphor (ana). Anaphora has two sub Sorts reflexive (refl) and reciprocal (recp).

With reference to the figure above, the left subscripts refer to the sort assignments, it can be a word, phrase or sentence. Here the sort of the assignment is word (she).

The boxed numerals (tags) indicate the structure sharing. For example the boxed numeral indicates that both INDX and INST share the same values.

Descriptions of lists are usually abbreviated by the use angle-bracket notation. Here `<>` describes the empty list.

Curly braces `{}` describes the sets and `{}` refers to the empty set.

As mentioned above each attribute label must be appropriate to the sort feature which represents in the structure, the value of that label must also be appropriate to both the SORT WORD and the ATTRIBUTE VALUE. For example, as depicted in the figure above, the value of CATEGORY attribute label of sort LOC is of sort CATEGROY (CAT).
5.1.2 ID Schemata of HPSG

Head driven phrase structure grammar is a compositional grammar that means the meaning of the complete sentence is determined by looking at the constituent structure of that sentence. HPSG has six basic and different schemata. These schemata introduce syntactic rules which are applied to compose both phrases and sentences. The feature in HPSG represents the syntactic relations between the different components. The most important principles of HPSG are Head complement principle, Head Feature Principle, Valence principle, Head Modifier principle, Head SPEC principle and Marker ones.

5.1.2.1 Head Complement Principle

This principle focuses on the head word and marks it as H. according to this principle if the head is in sequence with other arguments which are in correspondence with its complements requirements; it licenses a phrase with an AVM like those of the HEAD. The complement features which are deleted are not considered in this license. Here the head features of the HEAD are the head features of the whole phrase. If we apply this rule to a transitive verb and an NP, we can see the outputs of this rule as shown in the following figure 5.6

![Figure 5.6: Head Complement Principle](image)
5.1.2.2 Head Feature Principle

Head Feature Principle determines the HEAD value of the headed phrase with its HEAD-DTRS in other words the HEAD value of any headed phrase is similar to the HEAD value of the head daughter. This can be illustrated in the figure 5.7, one more feature to be noticed is that the HFP must be taken into account in the construction of all phrases.

![Diagram of Head Feature Principle]

**Figure 5.7: Head Feature Principle**

As shown in the figure 5.7 the HEAD of HEAD-DTRS, indexed 1, must be similar the head of the mentioned phrase.

5.1.2.3 Valence Principle (VALP).

The Third Important principle is The Valence Principle (VALP). According to this principle the mother’s SPR and COMPS values are identical to those of the head daughter if there is no exception according to the applied rules.

5.1.2.4 Head Specifier Principle

The Forth Principle is the Head Specifier principle. This principle allows sharing of the marker daughter’s SPEC value with the head daughter’s SYNSEM value. In other words, according to this principle the phrase which has a non-empty SPR value will be combined with another item which satisfies the SPR value and after that a phrase without an SPR value will be generated.
As it is illustrated in the figure 5.8 the MARER-DTRS’s SPEC value which is indexed 4 represents the HEAD-DTRS’s SYSTEM value.

5.1.2.5 Marking Principle

The Fifth Principle is the Marking principle. This principle proposes that the HEAD-MARKER phrase takes its MARKING value with the marker daughter if there is one, otherwise with the head daughter.

The HPSG schemata can be applied in Arabic syntactic phenomenon. But the HPSG features are not sufficient to cover all the Arabic lexical entries.

Some Arabic features are totally different from those of the HPSG features and others features are not represented and need to be added.
5.1.3 AVM of Arabic

5.1.3.1 AVM of Arabic Nouns

Here, the Sign-Based Construction Grammar (SBCG) version of HPSG for English is modified and adopted for Arabic language. With reference to the SBCG, we have kept some features and have added some others according to the proposed type’s hierarchy. For example, in Arabic language a linguistic sign (word or phrase) is characterized by its declination. This leads to the addition of the feature DEC (declination) that determines whether a word is سُرْعَة (m?urab) means the end of the word can be changed; or مبني (mabni) means the end of the word cannot changed. In the case of Arabic noun, declinable nouns are like common nouns or proper nouns. Indeclinable nouns are like personal pronouns, conjunctive nouns, relative pronouns, and demonstrative nouns. The table 6.1 represents the major features of Arabic nouns based on them we proposed our modified AVM for Arabic noun.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>NFORM</td>
<td>Declined / Indeclinable</td>
</tr>
<tr>
<td>DEFINITE</td>
<td>Definite/ Indefinite</td>
</tr>
<tr>
<td>NAT</td>
<td>Demonstrative/ conjunctive</td>
</tr>
<tr>
<td>ADJ</td>
<td>Adjective/ not</td>
</tr>
</tbody>
</table>

Figures 5.10 and 5.11 represent the Attributive value matrix for noun in both Arabic and English. The focus in these figures is on three main features; MORPH, SYN and SEM.

Figure 5.10: AVM for English Nouns
The feature PHON represents the phonological information without any indication to the morphological information. The MORPH feature indicates the morphological information of signs and in this figure it replaces the FORM feature which has a value of different morphological information.

Figure 5.11: AVM for Arabic Noun

The figure 5.11 above represents the AVM for Arabic noun which is extracted from AVM of English noun with some extension to cover the additional various linguistic features of Arabic. The first feature is PHON, as mentioned above this feature
represents the phonological information of the sings. MORHP is a function feature and it includes three main features – ROOT, SKELETON and DEC.

**ROOT** feature represents the root letters for the following cases:

a) The root is characterized as a part of a lexeme, and is common to a set of derived or inflected forms  
b) The root cannot be further analyzed into meaningful units when all affixes are removed  
c) The root carries the principal portion of the meaning of the lexeme

In other cases, the value of this feature is empty (Islam et al., 2010).

**SELEKTON** feature includes both the word stem and the inflected word. In other words this feature includes both lexical formatives and affixes. It includes all the letters of the given word. For example, in the case of the word *arsalna* means *(we sent)*, the FORM value is <arsala + na > both the stem of the inflected word and the suffix attached to that stem.

**DEC** feature refers to the declension type which is an additional Arabic feature to the MORPH feature in the AVM of Arabic. Declination is the morphological feature that determines the end of the vowel of a noun based on its case. A declinable noun in Arabic language has nine grammatical cases. The value of the DEC feature is one of those nine cases; Tn1, Tn2, Tn3 …Tn9 as illustrated in the table below. In the case of the indeclinable nouns, the value of the DEC feature is *none*.

**ARG-ST** feature refers to the syntactico-semantic arguments. This feature is based on the notion of governor. And as the noun lexeme is detached, the noun lexeme has no arguments except the accusative lexeme it needs governor requirements that can be captured by the AVM of the governor itself. For example the ARG-ST value of the accusative lexeme رسالات (means letter) is empty. In the case of the verb ارسلَ arsala (means sent), the ARG-ST of this verb ارسلَ arsala (means sent) contains رسالات (means letter) because the verb here is the governor of the noun lexeme رسالات.

**SYN** feature includes three features; CAT, VAL, and MRKG. The feature CAT indicates the different word categories. Arabic has three main categories; nouns, verbs
and particles. Based on a particular part of speech, the CAT feature has different values. The CAT feature of Arabic noun includes CASE, DEF, SELECT, XARG and LID features. The values of the first feature CASE are nominative, accusative and genitive. DEF feature indicates the value of definiteness of an Arabic noun and the syntactic agreement in the phrasal construction. The value of the DEF feature can be either yes or no. If the noun lexeme is defined its DEF value is yes, otherwise it is no. (Jamil, 2003-2011) states eight ways for noun lexeme to be definite:

1. A word made definite by means of the definite article: ﻣُؤْمَنَ (al )
2. A sentence can be made definite by means of a relative pronoun: “the car that was driven”
3. Demonstrative pronouns: “This”, “That”.
4. Proper nouns are also definite.
5. Personal pronouns such as “he”, “I” and “you” are inherently definite.
6. Objects of vocation: “O car!”
7. A noun which is possessive to any of the above: “Zahid’s car”
8. Special category: ﻣُؤْمَنَ (al-1-ahu ) is another instance of definite lexeme.

In the case of the importance of the DEF feature for the syntactic agreement in the phrasal construction. The DEF feature of the noun must agree with its modifier.

In the case of Arabic language the SEM feature includes two features – INDEX and FRAMES. In Arabic language the INDEX feature contains PERSON, NUMBER, GENDER and HUM (humanness) under SEM. And, PERSON, NUMBER and GENDER are utilized in English for semantic agreement. INDEX is used for index based semantic agreement as opposite to putting the agreements under the AGR feature. The reason behind the selection of the INDEX for semantic agreement is more used and more customary in HPSG. FRAMES refer to the list of the frames that includes semantic information in Minimal Recursion Semantic (MRS).

The feature PERSON has three values – 1st, 2nd and 3rd. Arabic has three values of NUMBER – sg (singular), dual (dual) and pl (plural). Arabic has only two values of the feature GENDER – masc (masculine) and fem (feminin). It is important to note that Arabic has no neutral gender as it is there in English.
In Arabic HUM feature refers to the humanness. This feature has a significant grammatical role for the agreement with other constituents of a certain phrase or clause within a sentence and for the prediction of certain plural formation. The HUM feature has two values either *yes* or *no*. if the noun lexeme refers to a human being, the value of HUM is *yes* otherwise it is *no*.

**Table 5.2: Noun Classes Based on the 9 Declension Types Adopted from (Islam et al., 2010)**

<table>
<thead>
<tr>
<th>Types</th>
<th>Declension</th>
<th>Noun class</th>
<th>Genitive</th>
<th>Accusative</th>
<th>Nominative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1</td>
<td>$T_1$</td>
<td>1. Triptote sound singular</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Singular noun pseudo sound</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. Triptote broken plural</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 2</td>
<td>$T_2$</td>
<td>4. Sound feminine plural</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 3</td>
<td>$T_3$</td>
<td>5. Diptote without prefixed by definite markness or not a possessed in a possessive phrase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\text{أَبِّ أَخِ حَمْ حِمْ هُنِم} \quad \text{مُذِوٌّ}$ possessed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>not towards first person singular number possessor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 5</td>
<td>$T_5$</td>
<td>7. Dual noun</td>
<td></td>
<td>$\text{اٍ}$</td>
<td>$\text{ي}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8. $\text{كِلَّا}$ and $\text{kَلَّا}$ possessed towards pronoun</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\text{إِنْتَاْنَ} \quad \text{إِنْتَاْنَ}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 6</td>
<td>$T_6$</td>
<td>10. Sound masculine plural</td>
<td></td>
<td>$\text{ي}$</td>
<td>$\text{ي}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11. Multiple of ten between twenty and ninety</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>12. $\text{أُولَو} \quad \text{(plural of possessor)}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 7</td>
<td>$T_7$</td>
<td>13. Noun possessed towards first person personal pronoun</td>
<td></td>
<td></td>
<td>$\text{ي}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14. ending with ya maqoora (ى)</td>
<td></td>
<td></td>
<td>$\text{ي}$</td>
</tr>
<tr>
<td>Type 8</td>
<td>$T_8$</td>
<td>15. Noun ending with $\text{ي}$</td>
<td></td>
<td>$\text{ي}$</td>
<td>$\text{ي}$</td>
</tr>
<tr>
<td>Type 9</td>
<td>$T_9$</td>
<td>16. Sound masculine plural possessed towards personal pronoun</td>
<td></td>
<td>$\text{ي}$</td>
<td>$\text{ي}$</td>
</tr>
</tbody>
</table>
**ARG-ST** feature includes syntactico-semantic arguments.

**SYN** feature includes three features; CAT, VAL and MRKG features. In the case of Arabic language there are three parts of speech; noun, verb and particle. In Arabic language the feature CAT contains CASE, DEF, SELECT, XARG and LID features. The value of the CASE can be one of these cases; *nominative, accusative* or *genitive*. DEF feature is the feature that is used in phrasal construction for the syntactic agreement. This feature indicates the value of definiteness of Arabic noun. If the noun is definite, the value of the feature DEF is *yes*, otherwise it is *no*.

### 5.1.3.1.1 An example of an Arabic Lexical discipline

If we consider a partial lexical entry of the Arabic noun (*duruusa*) means lessons in the following Figure.

![Figure 5.12: A basic Lexical Entry](image)

As illustrated in the figure 5.12 above the basic lexical entry consists of an ordered list of the form of the word and the large AVM. The ordered lists of the form of the word are represented with angled brackets (<>). The top-left corner specified the type word, it tell us this is a word and not a phrase or a sentence. Here the Attribute-Value
Matrix (AVM) describes objects of type word *duruusa* means *lessons*. Below this we have the SYN feature which has three values: HEAD, SPR, and COMPS features. The HEAD feature determines the category of the node and its inflectional properties. Here the HEAD feature tells us the kind of the lexical item. It is a noun, it triggers masculine in gender, plural in number and third person agreement. It is accusative in case. The SPR feature introduces an ordered list of the items such as determiners that can fill the specifier position. The SPR in this lexical entry includes only one item; that is the plural countable noun determiner. The last value of the SYN feature is COMPS feature that tells us PP complement may be optional.

For example:

a. Katabu al duruusa fi al nahuu.

They wrote the Lessons *in grammar*. Or

S V O PP. comp

b. Katabu al duruusa.

They wrote the lessons.

S V DET O

The other feature is the ARG-ST feature. The value of this feature is an ordered list of all the arguments associated with the word and represents the theta grid of the word. As illustrated in the figure above the boxed number < 1 > is called a (coreference). The tag < 1 > In both the SPR feature and the ARG-STR feature indicates that both of them share the same AVM. This means that the noun is preceded by a DET (definite article).

The boxed numbers indicate token-identity of the values specified. Instead of specifying such linking in each lexical entry, it can be derived from general linking principles (Koenig and Davis, 2003).

Finally we have the SEM (semantic) features which indicate to the information about the interpretation of both the words and sentences. As mentioned in the figure above,
the MODE tells us that the semantic type of the node (proposition, interrogative, directive and referential item). The INDEX features tell us the various participants or situations described in the sentence. Last, we have the RESTR (restriction) feature that indicates at the properties that must hold true for the sentence to be true.

5.1.3.2 AVM for Arabic Verb

In this thesis the developed AVM of Arabic Verb is extracted from Bhuyan et al. model with significant modifications, particularly the INDEX feature. If we compare the AVM of Arabic noun in figure 5.11 with the AVM of Arabic verb in figure 5.10 we can see that both the verb AVM and the noun AVM have the same MORPH feature value except the VDEC (verb declination). Based on the VDEC feature verbs can be declined \textit{m?urab} or in declined \textit{mabni}. The table 5.3 below represents the major features of Arabic verb.

<table>
<thead>
<tr>
<th>Features</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>RADICAL</td>
<td>trilateral / quadrilateral</td>
</tr>
<tr>
<td>VFORM</td>
<td>sound / defective</td>
</tr>
<tr>
<td>TYPE</td>
<td>intact / doubled</td>
</tr>
<tr>
<td>VOICE</td>
<td>Passive / Active</td>
</tr>
<tr>
<td>ASPECT</td>
<td>accomplished / unaccomplished / Imperative</td>
</tr>
<tr>
<td>ROOT</td>
<td>verb’s root</td>
</tr>
</tbody>
</table>

In the verb AVM, the MORPH feature takes the declension type of verbs. The DEC feature determines how the mood of the verb can change the shape of the end vowel of the verb. Verbs have four types of declensions and thus the VDEC values are Tv1, Tv2 . . . Tv4 denoting the four declension types of verbs as illustrated in the table 6.4 below.

<table>
<thead>
<tr>
<th>Types</th>
<th>Verb class</th>
<th>Jussive</th>
<th>Subjunctive</th>
<th>Indicative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1</td>
<td>1. Sound singular</td>
<td>jawazim</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 2</td>
<td>2. Unsound by</td>
<td>Extinction</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Unsound by not ending with</td>
<td>Extinction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 3</td>
<td>4. Unsound by</td>
<td>Extinction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 4</td>
<td>5. Imperfect ending with</td>
<td>Extinction</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The AVM of the Arabic verb includes VFORM (VERBFORM), VOICE and MOOD. The VFORM feature has three values—*perf* (perfect), *imperf* (imperfect) and *imperative*. *Perfect* denotes that the event has been completed, *imperfect* denotes that the event has not yet been completed, and *imperative* denotes that the event is a command. The VOICE feature also has two values—*passive* and *active*. The MOOD feature has three values—*subjunctive*, *indicative*, and *jussive*.

**SYN** and **SEM features** of Arabic are like that of SBCG English except in the case of INDEX. In SBCG there is no any distinction between the INDEX and SIT-INDEX features and deals with them as one feature. Arabic INDEX feature has a SIT-INDEX (situation index) which is utilized for index based semantic agreement and it has one atomic attribute that is SITUATION. The SIT-INDEX feature is utilized in event-fr feature.

**FRAMES** in AVM of Arabic verb are like those of the AVM of noun. The frames of the verb include indices of both SIT-INDEX and INDEX. The figure 5.13 represents an example of AVM model of Arabic verb.

![AVM model of Arabic verb](image-url)

**Figure 5.13: AVM of the Verb katab-uu**
The verb katab-uu means (they wrote) consists of the implicit subject they and the verb wrote. As shown in figure 5.13, the verb has a complete representation.

5.1.3.3 AVM of Arabic Particles

Arabic particles are considered as the third category of Arabic word. The PFORM feature is categorized into two main types – operative and non-operative particles. And the NATP features Table 6.5 represents the major features of Arabic particles.

Table 5.5: The Major Features of Arabic Particles

<table>
<thead>
<tr>
<th>Feature</th>
<th>values</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFORM</td>
<td>Operative / non-operative</td>
</tr>
<tr>
<td>NATP</td>
<td>Elision particle / subjunctive particle</td>
</tr>
</tbody>
</table>

As shown in figure 5.14, we defined different features and values characterizing this particle, at the level of TETE feature. Moreover, the feature SPEC specifies an object indexed 1. According to PFORM and NATP features, the object of this particle must be an elided verb.

![Figure 5.14: The AVM model of the Arabic elision particle (لم, lam) as a type of Arabic negation.](image)

5.3 Example based Machine Translation and HPS Grammar

It is a challengeable task to combine different ideas and approaches into a comprehensive whole. In this thesis we propose an approach to Machine Translation that combines the notions and techniques of Example-based machine translation model and Lexicalist theoretical frameworks that is Head driven Phrase Structure
Grammar (HPSG). This framework is chosen due to its rich and efficient computational characteristics. This approach has been implemented in our proposed Example-based Machine Translation system. The EBMT model and the HPSG have been integrated into one system that encompasses all of them. Despite the differences between their paradigms and methods, this combination calls the advantages of each approach. It makes use of the detailed linguistic information represented by HPSG which helps much in solving the linguistic ambiguities in an efficient way. On the other hand, EBMT approach allows the removal of the bilingual database’s redundancy resulted from the examples overlapping.

This theory is highly lexicalized and represents a detailed linguistic characteristics and information and this helps much the proposed system to overcome and solve most of the different linguistic ambiguities (lexical, morphological, syntactic, semantic and pragmatic) arising in the different steps of translation processes – matching, transfer and recombination. Here a unified a MT architecture can be proposed that encompasses the methodologies and advantages of both approaches, by, by utilizing bilingual information to drive the translation process, while preserving the modularity of the EBMT system.

A common characteristic of all EBMT approaches is that the translation process is driven by the content of the bilingual knowledge base and the bilingual corpus based on the AVM features represents all the information needed in EBMT. The bilingual knowledge base stating equivalences between the minimal translation units, e.i. the alignment here is a mapping between bags of linguistic features and values as they represented in the AVM models of both the source and the target texts. “A bilingual knowledge base as described above is not only a source of bilingual information, but it also encodes a considerable amount of monolingual linguistic information, on either side. A multi-word bilingual entry gives syntactic and semantic information about the analysis of phrasal expressions, collocations and idioms. Even single-word entries give clues about the analysis of lexically ambiguous items” (Turcato, 1999).

The following example represents how translation process works with English and Arabic in the verb phrase *she walks*: 

96
Figure 5.15: AVM (linguistic features of walks)

Figure 5.16: AVM (Linguistic Features of tamshi, Walks)
As we see in the above figures the words are annotated at different linguistic levels (morpho-syntactic and semantic levels). Besides the monolingual annotation and parsing of the verb in each language, this framework also has the advantage of biasing annotating and parsing towards analyses that are supported by the bilingual lexicon based on the AVM MODEL. In the case of the lexical ambiguity the semantic and indications and values will determine the right equivalent of the source word.

The first module is the pre-processing step in which the input is analysed based on the syntactic framework; HPSG.