

M.Sc. Engineering Thesis

Nonconcatenative Morphology in HPSG

by
Md. Shariful Islam Bhuyan

Submitted to

Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
Master of Science in Computer Science and Engineering

Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology (BUET)
Dhaka 1000

November, 2008



#105987#

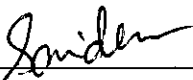
The thesis titled "Nonconcatenative Morphology in HPSG", submitted by Md. Shariful Islam Bhuyan, Roll No. 100605036P, Session October 2006, to the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, has been accepted as satisfactory in partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Engineering and approved as to its style and contents. Examination held on November, 2008.

Board of Examiners

1.  _____

Dr. Reaz Ahmed
Assistant Professor
Department of CSE
BUET, Dhaka 1000

Chairman
(Supervisor)

2.  _____


Dr. Md. Saidur Rahman
Professor & Head
Department of CSE
BUET, Dhaka 1000

Member
(Ex-officio)

3.  _____


Dr. Muhammad Masroor Ali
Professor
Department of CSE
BUET, Dhaka 1000

Member

4.  _____

Dr. Masud Hasan
Assistant Professor
Department of CSE
BUET, Dhaka 1000

Member

5.  _____

Dr. Chowdhury Mofizur Rahman
Professor
Department of CSE
United International University, Dhaka 1209

Member
(External)

Candidate's Declaration

This is to certify that the work entitled "Nonconcatenative Morphology in HPSG" is the outcome of the investigation carried out by me under the supervision of Dr. Reaz Ahmed in the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka-1000. It is also declared that this thesis or any part of it has not been submitted elsewhere for the award of any degree or diploma.



Md. Shariful Islam Bhuyan
Candidate

Abstract

Broad-coverage precision grammar and constraint-based lexicon development for deep linguistic processing is a research-intensive area with several potential applications. Amidst the vast literature on formal linguistic theory, Head-driven Phrase Structure Grammar (HPSG) has a unique position, since it combines the best features of the contemporary approaches as well as establishes an integrated framework for cross-layer representation of linguistic objects comprising phonology, morphology, syntax, semantics, pragmatics and discourse. In spite of being a successful syntactic theory in many respects, HPSG has inadequate coverage for morphological constructions, especially for nonconcatenative morphology, which is prominent in the Semitic languages such as Arabic, Hebrew, etc. Moreover, there is very few HPSG analysis of Arabic morphological, syntactic and semantic features. Arabic is the best instance of nonconcatenative morphology among the living languages. Arabic verb system shows a rich morphology, capable of lexically expressing diverse syntactic and semantic phenomena. Formalisms of existing morphological analyzers for Arabic cannot capture this higher layer diversity due to a lack of mathematical rigor and expressiveness. In this thesis, we extend the HPSG framework to support rich nonconcatenative morphology of Arabic verbs. We present HPSG analysis of the agency of Arabic passives and reflexives as well as morphologically complex predication of causatives in accordance with the existing analysis in other languages.

Acknowledgments

First of all, I would like to thank my supervisor Dr. Reaz Ahmed for giving the opportunity to explore the beautiful and fascinating field of mathematical modeling of natural languages, and for teaching me how to carry on a research work. I really appreciate his extraordinary patience in reading my numerous inferior drafts for making them publishable. I again express my heart-felt and most sincere gratitude to him for his constant supervision, valuable advice and continual encouragement, without which this thesis would have not been possible.

I would like to express my special thanks to Maolana Ahmad Ali and Maolana Ruhul Amin Qasemi, two great scholar of Arabic for the valuable discussions frequently made with them. They have provided many suggestions regarding the contents.

Finally, I owe my loving thanks to my family. Without their encouragement it would have been impossible for me to finish this work. Above all, I am grateful to Almighty ALLAH Subhanahu wa-ta'ala (glorious and exalted is He), who gave me the strength to finish this work.

Contents

<i>Board of Examiners</i>	1
<i>Candidate's Declaration</i>	2
Abstract	3
Acknowledgments	4
1 Introduction	9
2 Background	12
2.1 Linguistic Background	12
2.1.1 Morphology	13
2.1.2 Syntax	16
2.1.3 Semantics	18
2.2 An HPSG Primer	20
2.3 Morphology in Arabic	27
2.4 Summary	32
3 An HPSG analysis of Arabic	33
3.1 Modeling of Arabic Verb	33
3.2 Perfect Form	36
3.3 Imperfect Form	39
3.4 Imperative Form	41
3.5 Summary	43

4 Agency in Arabic	44
4.1 Arabic Passive	44
4.1.1 Passive Perfect	44
4.1.2 Passive Imperfect	46
4.1.3 Passive Imperative	48
4.2 Arabic Causative	50
4.2.1 Active Causative	51
4.2.2 Passive Causative	53
4.3 Arabic Reflexive	56
4.3.1 Action Reflexive	56
4.3.2 Reflexive for <i>cause</i> -predicate	59
4.3.3 Reflexive for <i>think</i> -predicate	62
4.4 Summary	64
5 Conclusion	65
Appendix	68

List of Figures

2.1	CFG for a small fragment of English	21
2.2	A parsetree of <i>I eat rice</i>	22
2.3	A parsetree of <i>I *eats rice</i>	22
2.4	HPSG for a small fragment of English	23
2.5	An HPSG Sign and construct	23
2.6	A Standard SBCG type hierarchy	25
2.7	SBCG Sign for the English verb “write”	27
2.8	Derivational Paradigm of root ktb	29
3.1	HPSG Sign for an Arabic verb	34
3.2	A partially modified HPSG type hierarchy for Arabic	35
3.3	HPSG Sign for <i>kataba</i>	37
3.4	HPSG Sign for <i>yaktubu</i>	40
3.5	HPSG Sign for <i>uktub</i>	42
4.1	HPSG Sign for <i>kutiba</i>	45
4.2	HPSG Sign for <i>yuktabu</i>	47
4.3	HPSG Sign for <i>lituktab</i>	49
4.4	HPSG Sign for <i>kattaba</i>	52
4.5	HPSG Sign for <i>kuttiba</i>	54
4.6	HPSG Sign for <i>igtasala</i>	57
4.7	HPSG Sign for <i>taTahhara</i>	60
4.8	HPSG Sign for <i>takabbara</i>	62

List of Tables

2.1	Thematic role	20
2.2	Derivational Paradigm of root ktb	29
2.3	Inflectional Paradigm of Form I-Active-Perfect	30
2.4	Inflectional Paradigm of Form I-Passive-Perfect	30
2.5	Attributes Governing Morphological Paradigm	31
5.1	Transliteration Table of Arabic Alphabet	68



Chapter 1

Introduction

Natural languages processing is an inter-disciplinary field of research where several well established branches of knowledge such as artificial intelligence, linguistics, mathematics and philosophy just to name a few, conjoined to answer some of the most difficult questions ever posed to mankind. Here a complete success will result in the ability to emulate cognitive capabilities at such a level that we can have artifacts which will understand human conversations, communicate linguistically and also pass the famous Turing-test. Even a partial success will help us in several practical applications, namely machine translation, natural language interfaces to computer systems, speech recognition, text to speech generation, automatic summarization, e-mail filtering, intelligent search engines and many more.

Broad-coverage precision grammar and computational lexicon development [1-7] for deep linguistic processing [8] is a research-intensive area with several potential applications [8-10]. Amidst the vast literature on formal linguistic theory [11], Head-driven Phrase Structure Grammar (HPSG) [12] has a unique position since it combines the best features of the contemporary approaches as well as establishes an integrated framework for cross-layer representation comprising phonology, morphology, syntax, semantics, pragmatics and discourse. In spite of being a successful syntactic theory in many respects, HPSG has inadequate coverage for morphological constructions [13], especially for nonconcatenative morphology [14, 15], which is prominent in the Semitic languages such as Arabic, Hebrew, etc. Arabic verb system shows a rich nonconcatenative templatic morphology, capable of lexically expressing diverse syntactic and semantic phenomena. Formalisms of existing morphological analyzers [16-18] for Arabic are

not powerful enough to capture this higher layer diversity.

This thesis focus on the following two objectives. First, we extend the HPSG framework to support rich nonconcatenative templatic morphology. Second, we define the first comprehensive HPSG-construction for the morphology of derivational and inflectional paradigm of Arabic verbs based on sound root class, to the best of our knowledge using our extension. This includes capturing several morpho-syntactic and semantic features such as agreement, agency, subcategorization, event structure, complex predication and modality. Arabic is the best instance of nonconcatenative morphology among living languages as well as 6th ranked language with approximately 422 million native speakers. It is also the intellectual and liturgical language of the Islamic World.

The main results of this thesis are as follows:

1. We extend the HPSG framework to support rich nonconcatenative templatic morphology. To support generic nonconcatenative morphology, the feature MORPH in the attribute value matrix (AVM) of the HPSG Sign is modified. Three new features, e.g., TYPE, ROOT and MEASURE have been added.
2. We define the first comprehensive HPSG-construction for the morphology of Arabic verbal system using our extension, including the agency of Arabic passives and reflexives as well as morphologically complex predication of causatives. Co-indexing of semantic arguments is used to capture agency in Arabic passives and reflexives. However, predicate constituent is used to capture the phenomena such as event structure and modality. Finally, morphologically complex predicates e.g., causative constructions, are captured with a blend of predicate constituent and co-indexing.
3. We construct an initial type hierarchy for an Arabic constraint-based lexicon. The TYPE feature is used to denote the class of roots, which share a common derivational paradigm. The attributes governing derivational and inflectional paradigm of an Arabic word are identified.

Some of these results are also presented in [32–34]. The thesis is organized as follows.

Chapter 2 gives a background by explaining the basic ingredients and necessary tools. It discusses about several linguistic topics ranging from morphology, syntax to semantics. Next, it gives a brief introduction about HPSG, the mathematical theory of

languages used in our thesis. Then it provides a sketch of Arabic grammar, mainly the morphology associated to its rich word construction. Chapter 3 describes our contribution of the development of a generic structure of the attribute value matrix of an Arabic verb. Next, it discuss about the introduced features and their implementation for various Arabic verb form. Chapter 4 explains about the very rich agency in Arabic. It mainly analyzes three types of agency construction; passive, causative and reflexive. Finally, Chapter 5 gives the conclusion.

Chapter 2

Background

We discuss several topics in this chapter, which serve as a background of the rest of the thesis. In the Section 2.1 we explain some linguistic background. Section 2.1.1 gives an introduction of morphology, Section 2.1.2 gives an introduction of syntax and Section 2.1.3 gives an introduction of semantics. Next, the Section 2.2 gives an overview of Head-driven Phrase Structure Grammar. Finally, Section 2.3 gives a short introduction of Arabic verbal morphology.

2.1 Linguistic Background

Language understanding is quite a complicated task. Many of the developed algorithms are computationally intractable. Necessary knowledge for processing is enormous and most stages of the process involves ambiguity. The whole process requires layered information that can be summarized as follows:

- **Phonology:** Study of speech sound
- **Morphology:** Study of word formation
- **Syntax:** Study of sentence construction
- **Semantics:** Study of meaning
- **Pragmatics:** Study of situational context
- **Discourse:** Study of connected speech

In this thesis, we mainly focus on morphology, especially nonconcatenative morphology, with its implication on syntax and semantics. For this reason we need to discuss some of the concepts related to the morphology, syntax and semantic layer. We took the linguistic definitions from [31].

2.1.1 Morphology

Morphology deals with the study of the patterns of word formation in a particular language, description of such patterns and the behavior and combination of morphemes. It is difficult to define any linguistic object precisely, since they vary from language to language. However, we can identify some properties of the concept **word**, which is a grammatical unit and used as a minimal possible unit in a reply. Word boundaries impose restrictions over the phonological stress. A word is the largest unit, which denies the insertion of new constituents within its boundaries. It is also the smallest constituent that can move within a sentence without making the sentence grammatically incorrect.

A **morpheme** is the smallest meaningful unit in the grammar of a language. The word *dogs* consists of two morphemes: *dog*, and *-s*, a plural marker on nouns. A **bound morpheme** is a grammatical unit that never occurs by itself, but always attached to some other morpheme. Hence, the plural morpheme *-s* in *dogs* is a bound morpheme. A **free morpheme** is a grammatical unit that can occur alone. However, other morphemes such as affixes can attach to it. Here the word *dog* is a free morpheme.

We can form new words from existing ones by **morph-syntactic** operations. There are two kinds of morpho-syntactic operation, **inflection** and **derivation**. Inflectional operations create forms that can be readily embedded in the sentence with discourse compliance, whereas derivational operations create forms that cannot be necessarily embedded in the sentence and which may still require inflectional operations before they can be integrated into discourse. In the example,

(2.1) He speaks for people.

the word *speak* inflects to *speaks*. This represents the temporal aspect of the action to the time of utterance as well as third-person, singular-number actor attributes. However, in the example,

(2.2) There are many speaker.

the word *speaker* cannot be readily integrated. Although, this word is derived from *speak*, it needs to be inflected to *speakers*. Inflection does not change the lexical category of the word and contribute syntactically constrained information, such as number, gender, or aspect. However, derivation often changes the lexical category of the word (e.g., *speak* is verb and *speaker* is noun) and contribute different lexical meaning.

In the previous examples, the bound morphemes *-er* and *-s*, which are joined after the word *speak*, are called **affix**. A **root** is characterized, as the part of a word, which is common to a set of derived or inflected forms, cannot be further analyzed into meaningful units when all affixes are removed, and carries the principle portion of meaning of the words. For example, *speak* is the root of the words *speaks*, *spoke*, *spoken*, *speaking*, *speaker*, *speakers*, *spokesman* etc. There are two kinds of affixes, **inflectional affix** and **derivational affix**. A derivational affix is an affix by means of which one word is derived from another. For example, *-er* of the word *speaker* is a derivational affix. An inflectional affix is an affix by means of which one word is inflected from another. For example, *-s* of the word *speakers*. **Stem** is a root with any derivational operation, to which inflectional affixes are added. In our case, the word *speaker* is a stem.

Morphology deals with two kinds of information. First, what information is encoded by the morpheme. For example, we can take an Arabic word *kataba* - he wrote. In this thesis, we use a romanized transliteration of Arabic alphabet from the Table 5.1. A variety of information is encoded in this word and its other inflected or derived form. Some are listed below:

- **Agreement:** كَتَبَ - *kataba* – he wrote. Person – 3rd, Number – Singular, Gender – Masculine., Mood – Indicative.
- **Event structure:** كَتَبَ - *kataba* – he wrote. Tense – Past, Aspect – Perfect.
- **Agency:** كُتِبَ - *kutiba* – it was written. Voice – Passive.
- **Illocutionary force:** اُكْتُبْ - *uktub* – Write. Mode – Command.

- **Part-of-Speech:** كِتَابٌ - *kitaabun* – a book. *kataba* – verb, *kitaabu* – noun.
- **Definiteness:** الْكِتَابُ - *al-kitaabu* – the book Determiner – Definite.
- **Complex Predicate:** كَتَّبَ - *kattaba* – he made to write. Semantic relation – Causation.

There are many more syntactic and semantic phenomena those can be expressed using morphology. Second issue, with which morphology deals with, is how information is encoded in the morpheme. Morpho-syntactic operations performed over the morphemes come with two flavors: concatenative and nonconcatenative.

- **Concatenative** operations are those where morphemes are linearly concatenated. For example:
 - **Prefixation:** Morphemes concatenated at the front, e.g., clear – *unclear*
 - **Suffixation:** Morphemes concatenated at the back, e.g., walk – *walked*
 - **Circumfixation:** Morphemes concatenated both at the front and back, e.g., mind – *unmindful*
- **Nonconcatenative** operations are those where morphemes are nonlinearly embedded. For example:
 - **Infixation:** Root letter morphemes embedded at the middle, e.g., *kataba* — *kartaba*
 - **Simulfixation:** Front morpheme shifted to the back, e.g., *eat* — *ate*
 - **Modification:** Middle vowel changed, e.g., *man* — *men*
 - **Suppletion:** Whole stem changed, e.g., *go* — *went*

There are many other morpho-syntactic operations also. In this thesis, we mainly focus on nonconcatenative operation as well as concatenative operation and give a mathematical formalism to capture their rich diversity.

2.1.2 Syntax

Syntax is the study of the rules of construction of phrases (including sentences). We start with the notion of **construction**, which is an ordered arrangement of grammatical units forming a larger unit. For examples in English,

- *subject + verb + object* — forms a clause
- *preposition + noun* — forms a prepositional phrase

There are many kinds of constructions such as, sentence, clause, direct/indirect speech, elliptical, idiom, phrase and many more. We also consider stem and word as a kind of lexical construction. A **sentence** is a grammatical unit that is composed of one or more clauses. A **clause** is a grammatical unit that includes, at minimum, a predicate and an explicit or implied subject, and expresses a proposition. A **phrase** is a syntactic structure that consists of more than one word but lacks the subject-predicate organization of a clause. A **head** is a constituent of a headed construction that, if standing alone, could perform the syntactic function of the whole construction. It may govern the agreement of grammatical categories, such as person and number, or occurrence of other constituents. The characteristics of a head is determined by its syntactic category. A **syntactic category** is a set of words and/or phrases in a language which share a significant number of common characteristics. The classification is based on similar structure and sameness of distribution (the structural relationships between these elements and other items in a larger grammatical structure), and not on meaning. Among the major syntactic categories there are phrasal syntactic categories like **NP** (noun phrase), **VP** (verb phrase), **PP** (prepositional phrase) and lexical categories that serve as heads of phrasal syntactic categories like **noun**, **verb** and others. For example a prepositional phrase (PP) is a phrase that has a preposition as its head. The definition is similar for noun phrase (NP) and A verb phrase (VP).

A **constituent** is one of two or more grammatical units that enter syntactically or morphologically into a construction at any level. For example, the sentence, *He gave a book to me yesterday.* – contains the following constituents:

1. **Immediate constituents:** He, gave a book to me yesterday
2. **Ultimate constituents:** He, gave, a, book, to, me, yesterday

There are several related, cross-cutting and sometimes confusing concepts related to constituents. We explain the phenomena at syntactic level. **Syntactic constituents** can be classified under **syntactic category** means the constituent head will be a noun, verb, adjective, adverb, preposition or something like that. Constituents can perform **syntactic functions** in the construction. A **syntactic function** is the grammatical relationship of one constituent to another within a syntactic construction. There are various kinds of syntactic functions such as subject, predicate, object, complement, adjunct, modifier and others.

A constituent performing the syntactic function of a **complement** has a phrasal or clausal syntactic category and is *subcategorized* (selected) by the head of a phrase. A selected, or subcategorized, phrase is obligatory, as contrasted with **adjuncts**, which are, broadly defined, an optional constituent of a construction. For instance, the **direct object** of a transitive verb is obligatory and therefore a complement, whereas **adverbial modifiers** are generally optional, and therefore non-complements. However, the distinction is not always clear, particularly for oblique objects. An **oblique object** is a grammatical relation proposed for a noun phrase clause constituent whose nature and behavior are more readily describable in semantic terms than syntactic. Also, while the subject of a clause is often considered a core argument of the verb, it is not normally considered to be a complement. This is because in most languages, the subject appears to be a clause-level constituent, rather than a constituent of the verb phrase. However, in Arabic, subjects appear to be a complement of verbs along with the concept of hidden pronoun. Different heads can have complements from different syntactic category, as explained in following sentences:

- He eats *rice*. (direct object NP complement of the verb)
- He gave *me* the book. (indirect object NP complement of the verb)
- He put it *on the desk*. (obligatory locative complement PP of the verb)
- This problem seems *very easy*. (adjective phrase or AP complement of the verb)
- They doubted *whether it was possible*. (sentential complement of the verb)
- ... under *the table* (NP complement of a preposition)
- ... hard *to understand* (VP complement of an adjective)

Depending on the count of required objects we classify verbs into different categories. For example, if a verb does not require any object then it is called **intransitive verb**. If a verb requires a direct object then it is called **transitive verb**. If a verb requires both direct and indirect object then it is called **ditransitive verb**. Depending on the argument optionality we can further classify verbs. For example, if an intransitive verb can never take an argument, it is called strictly intransitive verb. In this way we can also define transitive and ditransitive verbs with optional argument. Another type of restriction that can be found in Arabic is dependent optionality. This phenomenon can be seen in a ditransitive verb which can take two objects or leave both of them. It is not possible to select any one of them.

Constituents performing the role of a *modifier* in a headed construction restrict or qualify some other constituents relating to the head of the construction. In the headed construction *the very hot soup*, the constituents *the* and *very hot* are modifiers of *soup*, the head of the construction. There is little distinction between modifier and adjunct, since most of the modifiers are optional.

A grammatical category is a set of syntactic features that express meanings from the same conceptual domain, occur in contrast to each other, and are typically expressed in the same fashion. The term **grammatical category** has been used to cover a wide variety of things, including what traditional grammars call **parts of speech**. Some dimensions of grammatical category are person, number, gender, definiteness, class, case, tense, aspect, mood, voice, polarity, form, declination, transitivity and many more.

2.1.3 Semantics

Construction of an appropriate semantic representation for natural languages remains one of the most difficult problems in the area of knowledge representation. There are several proposals but none of them are accepted by everyone. This is indeed a very hard problem due to its abstract nature. We look at two of the most important properties required to define the semantics of a verb; first, its **valence** feature (i.e., the number of basic arguments that it requires) and its **thematic roles** (i.e., the semantic roles played by the basic arguments). Combining them we can find verb feature which is called the argument structure of verb. To understand them we can take the following example,

(2.3) He broke the window with a hammer.

In this example, the verb *break* requires two arguments: the subject *He* and the object *the window*. Both the arguments are required because, if any one of them are missing, the sentence would be ungrammatical. But the following sentence is correct.

(2.4) He broke the window.

For the verb *break*, the semantic role of the subject is *actor*, and indicates the entity responsible for the event. The semantic role of the object is *undergoer*, and indicates the entity which experiences the state or change of state described by the verb. In other words, the argument structure of the English verb *break* requires two arguments: the first argument (i.e., the subject) must be a semantic agent, and the second argument (i.e., the object) must be a semantic undergoer. Arguments required by a verb are called *core arguments*.

The phrase *with a hammer* is what is called an oblique argument since it is not essential for the sentence to be grammatical. It simply provides additional peripheral information about what happened. In this sentence, it indicates the *instrument* of the event.

Thematic roles are values such as actor, undergoer, soa, goal, etc, assigned to the arguments of verbs and other predicates. They are used to give a semantic classification of arguments and to express generalizations with regard to the syntactic realization of arguments. We adopt the influential proposal of [25, 26] for our analysis. The proposal is given within the framework of Head-driven Phrase Structure Grammar (HPSG) which we use as our mathematical formalism. To be able to express generalizations on thematic roles, instead of specific thematic roles like eater, writer, player and others, [25] uses proto-roles like actor, undergoer etc. Specific roles seem unnecessary for linking with the syntactic arguments. In the Table 2.1, we give description of some proto-roles used in our analysis.

Meaning of verbs are expressed using predicates which is a relation with semantic roles as its argument. For example, in our example of the verb *break*, it introduces an event predicate name *break(breaker, broken)*. This is an example of an actor-undergoer relation.

Table 2.1: Thematic role

Proto-role attribute	Entailments
ACTOR	<p>Causally affects or influences other participant(s) or event(s). Volitionally involved at the event. Has a perception of other participant(s) in event or state. Exerts forceful contact on other participant(s) in event. Includes another participant in state or event. Is superior compared to another participant. Possesses another participant in state or event.</p>
UNDGR (Undergoer)	<p>Causally affected or influenced by another participant in event. Undergoes change of state in event. Is an incremental theme in event. Moves with respect to another participant in event.</p>
SOA (Statement of affairs)	<p>Is conceived of or perceived by another participant in event or state. Is a resulting event or state caused in event. Is an event or state that necessarily accompanies another event.</p>
GRND (Ground)	<p>Path traversed by another participant in event.</p>

2.2 An HPSG Primer

Natural languages generally consist of two components. First, the utterances that can be used by human. Second, the linguistic rules that *license* those utterances. For example, in English, *He writes books*, *writes books*, *writes* – all are valid utterances. However, *Writes he books*, *writes he*, *rwite* are not valid, since the rules do not *license* them. HPSG is a mathematical theory for natural languages that formally captures these two core linguistic components. Utterances are modeled using a mathematical object **Sign**, which is a formal representations of words, phrases as well as sentences. Rules are captured using another mathematical object **Construct**, which is a formal representations of grammar rules or schema that are used to license signs. Both sign and construct are described using feature structure - a collection of features of corresponding linguistic objects along with their values.

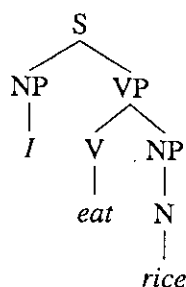
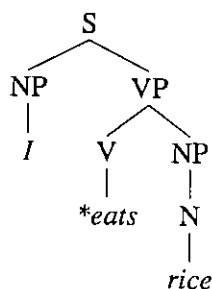
1. $\Sigma = \{\text{eat, eats, } \dots, \text{rice, Jo, } \dots, \text{I, you, he, } \dots\};$
2. $V = \{S, VP, NP, V, N, P\};$
3. $S = \text{Start symbol};$
4. $P = \{$
 - $S \rightarrow NP VP$
 - $VP \rightarrow V NP$
 - $VP \rightarrow V$
 - $NP \rightarrow N$
 - $NP \rightarrow P$
 - $V \rightarrow \text{eat, eats, } \dots$
 - $N \rightarrow \text{rice, Jo, } \dots$
 - $P \rightarrow \text{I, you, he, } \dots$

Figure 2.1: CFG for a small fragment of English

To understand the motivation of HPSG we need to start from its predecessor Context Free Grammar (CFG). A Context Free Grammar G is a 4-tuple $G = (\Sigma, V, S, P)$ where,

1. Σ is a finite, non-empty set of terminals, the alphabet;
2. V is a finite, non-empty set of non-terminals;
3. $S \in V$ is the start symbol;
4. P is a finite set of production rules, each of the form $A \rightarrow \alpha$, where $A \in V$ and $\alpha \in (V \cup \Sigma)^*$

For example, let us have CFG in Figure 2.1 for very small fragment of English. Using this CFG, the sentence *I eat rice* can be analyzed by the derivation in Figure 2.2.

Figure 2.2: A parsetree of *I eat rice*Figure 2.3: A parsetree of *I *eats rice*

However, there are problems with CFG. The above definition also generates the sentence *I eats rice* as in Figure 2.3. Second derivation is grammatically wrong. We do not capture the accurate agreement information with the grammar G . Basic problem with CFG is that its terminals are non-informative. This leads to Head-driven Phrase Structure Grammar (HPSG), a constraint-based lexicalist formalism of natural language. In HPSG our grammar fragment will look like the grammar in Figure 2.4.

Now, we put the agreement information inside the terminals. We do not handle the transitivity here. HPSG does not license the second derivation, since it violates the constraint in the first phrase structure rule - agreement of the noun phrase and agreement of the subject of the verb phrase must match. Here, a technique, called **structure-sharing** is used. There are two boxed-one in the phrase structure rule of HPSG fragment in Figure 2.4. It means that the value of these two agreements share the same value. We call the information-bearing terminals as **lexical sign** and non-terminals as **phrasal sign**, Σ as lexicon, phrase structure rules as **constructs** and the matrix associated with each

1. $\Sigma = \left\{ \left[\begin{array}{l} \text{eats} \\ \text{HEAD } \textit{verb} \\ \text{SUBJ } \left[\begin{array}{l} \text{AGR } \left[\begin{array}{l} \text{NUM } \textit{sg} \\ \text{PERS } \textit{3rd} \end{array} \right] \end{array} \right] \\ \text{VAL } \langle \text{[NP]} \rangle \end{array} \right], \left[\begin{array}{l} \text{I} \\ \text{HEAD } \textit{noun} \\ \text{AGR } \left[\begin{array}{l} \text{NUM } \textit{sg} \\ \text{PERS } \textit{1st} \end{array} \right] \end{array} \right], \dots \right\};$
 2. $V = \{S, VP, NP, V, N, P\};$
 3. $S =$ Start symbol;
 4. $P = \{$
 - $S \rightarrow NP \left[\begin{array}{l} \text{AGR } \square \end{array} \right] VP \left[\begin{array}{l} \text{AGR } \square \end{array} \right]$
 - $VP \rightarrow \left[\begin{array}{l} \text{HEAD } \textit{verb} \end{array} \right]$
 - $NP \rightarrow \left[\begin{array}{l} \text{HEAD } \textit{noun} \end{array} \right]$
 - \vdots
- }

Figure 2.4: HPSG for a small fragment of English

<i>sign</i>		<i>construct</i>	
PHON	<i>φ-phr</i>	MTR	<i>sign</i>
MORPH	<i>morph-obj</i>	DTRS	<i>list(sign)</i>
SYN	<i>syn-obj</i>		
SEM	<i>sem-obj</i>		
SYN	<i>syn-obj</i>		
⋮	⋮		

Figure 2.5: An HPSG Sign and construct

sign as attribute value matrix, according to HPSG terminology.

Grammatical objects of all kinds (including signs, case values, parts of speech, and

constructions) are modeled as feature structures. I make the further assumption that feature structures are either atoms (like *pl(ural)*, *acc(usative)*, *+*, etc.) or else functions from features to feature structures. This is a simple, but powerful way of modeling linguistic objects,

It is important to notice that although feature structures themselves are complete, feature structure descriptions may be partial. Lexical entries will be formulated as partial feature structure descriptions (typically being true of many feature structures), as will grammatical constructions of all kinds. Yet underlying all our concerns will be the set of feature structures that are specified by the theory we present. An HPSG must neither overgenerate (by delimiting a set of feature structures that includes some that do not model expressions of the target language), nor undergenerate (by failing to provide descriptions of some feature structures that do model expressions of the target language). Our feature structures have one more property that is not part of the basic theory of functions, as standardly presented we assume that feature structures are organized in terms of a theory of linguistic types. A type is associated with a set of feature structures that have certain stated properties in common. One benefit derived from assigning feature structures to types is that we can thereby better organize the properties that classes of grammatical objects have and simplify their description in the process. Intuitively, certain grammatical feature specifications are appropriate only for certain kinds of grammatical objects. This intuition is given formal expression in terms of the types that particular feature structures instantiate. Each feature structure instantiates a particular maximal (most specific) type. This type assignment, together with the general structure of the space of types, determines that the feature structure in question specifies values for a particular set of features and that each features value is a particular kind of feature structure (possibly, a function of a particular type; possibly an atom, e.g. *nominative* or *+*).

Here, we face the problem of selecting appropriate attributes for a particular language. Attributes are selected by linguistic motivation as well as language independent requirements. A linguistic object can be captured at multiple layers. For example, a sign have attributes that can be phonological, morphological, syntactic, semantic, pragmatic and so on. Well-established representation of signs captures these different aspects of a linguistic object using an attribute value matrix. To capture these features, the description of a typical HPSG sign looks like Figure 2.5. To capture grammatical rules, the feature structure of a construct has a mother (MTR) feature and a daughters (DTRS)

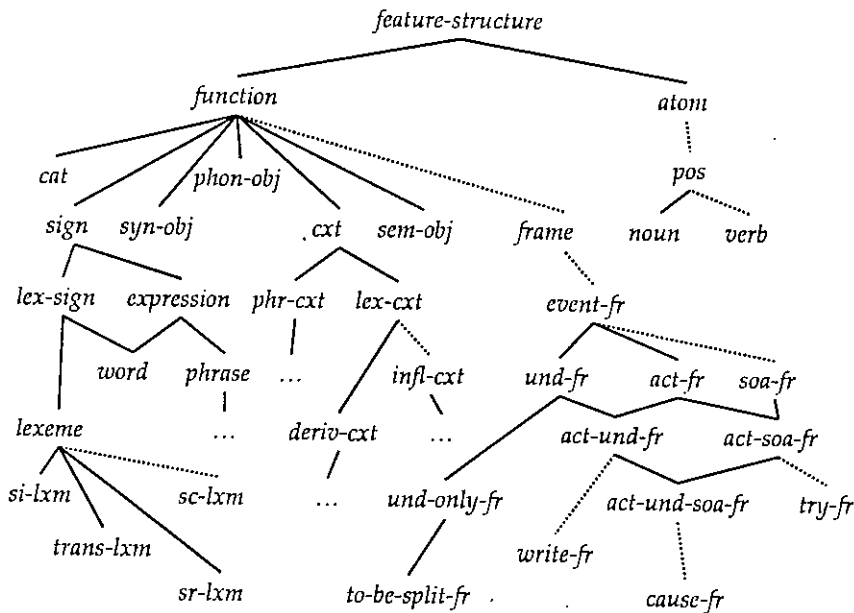


Figure 2.6: A Standard SBCG type hierarchy

feature. The value of the MTR is a sign and the value of the DTRS is a nonempty list of signs. the description of a typical HPSG construct looks like Figure 2.5. The features and their values are modeled using feature structure. They are organized into a type hierarchy which forms the grammar core.

As stated before, HPSG modeling of any language starts from building a very detailed type hierarchy which is both linguistically motivated as well as captures the language independent constraints. From this type hierarchy we can construct the corresponding attribute value matrix for linguistic signs. In this thesis, we use the ideas of Sign-Based Construction Grammar (SBCG) which is “an attempt to adapt ideas developed over a twenty year period of research in Head-Driven Phrase Structure Grammar (HPSG) to the analysis of constructional phenomena of the sort studied in the tradition of Berkeley Construction Grammar (BCG)” [27]. A Standard SBCG type hier-

archy is given in Figure 2.6. From the type hierarchy we know that every linguistic object can be modeled using **feature-structure**. There are two types of Feature structure. **Atoms** are simple feature structures, which indicate the terminal value of various linguistic attributes. **Functions** are complex feature structure, which are expressed using attribute value matrix and can contain other feature structures as their feature values. **Sign** and **cxt(construct)** both are feature-structure. The attribute of signs are also feature-structure; **phon-obj**, **syn-obj**, **sem-obj**, etc. **Frames** are semantic representation of events which can be classified according to [27]. There are two types of constructions; **phr-cxt**(phrasal) and **lex-cxt**(lexical). There are also two types of sign; **lex-sign** and **expression**. For the detail description of type hierarchy of HPSG, see [27]. We demonstrate a typical attribute value matrix for the English verb *write* in the Figure 2.7, according to [27].

If we look closely, we see that the paper [27] does not provide an in-depth discussion of the first two sign-level features, e.g., **PHON** and **FORM**. The value of **FORM** feature should capture the morphological constituents, which are phonologically realized in the **PHON** feature. **FORM** contains lexical formatives and affixes. Next, the **ARG-ST** is a list feature that encodes the combinatoric potential of a lexical sign. This maintains a rank-based representation of list elements which indicates their grammatical functions, constrains the placement of anaphora and has other purpose. The **SYN** feature express the syntactic constraints of sign. Among its sub-features **CAT** describes the complex grammatical category associated with the sign. For verbs, it captures the morphosyntactic category of verb form through **VF** and describes whether the verb is an **AUX**(iliary).

XARG is the argument of an argument-taking expression outside the phrase it projects. For example, in English the external argument of a clause is its subject. In Arabic subjects are phrase level constituent. So, we do not use this feature. **MRKG** and **SELECT** expression select what it can modify or combine with as a marker and they are not subjects of our concern.

We need an analysis of **SEM** feature for present purpose. **INDEX** individuate the referent of an expression essentially a variable assigned to an individual (**NP**) or a situation (**VP**). **FRAMES** specify the predications that together determine the meaning of a sign. This picture an elementary scene in which certain roles are specified and particular participants are assigned to them. For example, in an *eating* frame the participants are an actor, who does the eating, and the food, which gets eaten. In the representation

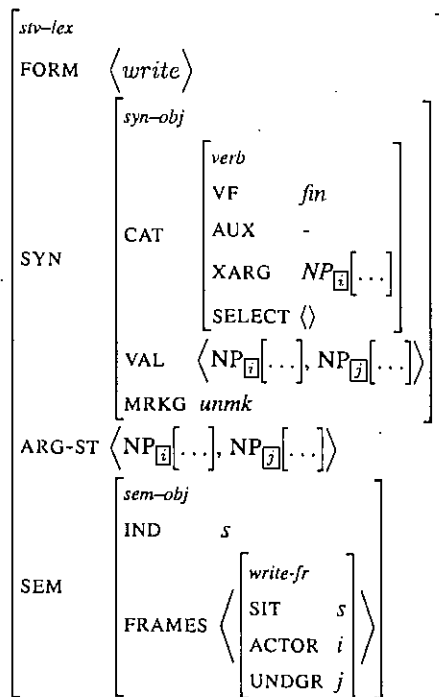


Figure 2.7: SBCG Sign for the English verb “write”

frames in the form of feature structures, each role is denoted by a feature and the corresponding participant by an index. In the case of most (if not all) verbs, an event index will be encoded as a SIT(uation) feature, whose value is a situational index, and there will often be an ACTOR feature, whose value is an individual index. It is important to understand that SBCG is in fact compatible with most approaches to semantic analysis.

2.3 Morphology in Arabic

Classical Arabic exhibits an extremely rich morphology [13–18]. Both concatenative and nonconcatenative operations take place in the formation of an Arabic word. Inflection is made by concatenative operations whereas derivation is made by nonconcatenative operations.

Arabic word formation is an excellent example of root-pattern morphology. A combination of root letters are plugged in a variety of morphological patterns with priory fixed letters and particular vowel melody that gives rise to corresponding syntactic and semantic phenomena. However, verb formation in this manner is not strictly productive. Of the major ten templates, normally only two are three are extant for a given root, and often their meanings are idiosyncratic. Our discussion does not intend to be complete,

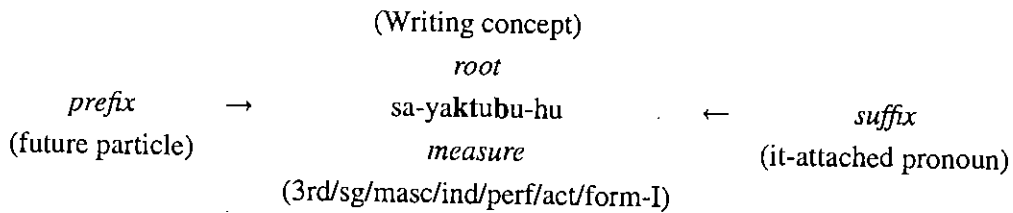
but rather to give an overview of how the system works. To feel the richness of Arabic morphological patterns, which we call "measure" in this thesis, following example is given. Here, the root letters *k, t, b* bearing a concept of writing, is plugged in various measures to get a myriad of syntactic and semantic phenomena. The measures with a particular semantic paradigm are called FORM. Arabic has many forms. Among them, ten forms are used regularly. The root letters *ك، ت، ب* (*k, t, b*) can be plugged in among nine of them.

1. **Form I** (Transitive): *كَتَبَ* (kataba) - He wrote.
2. **Form II** (Causative): *كَتَّبَ* (kattaba) - He caused to write.
3. **Form III** (Ditransitive): *كَاتَبَ* (kaataba) - He corresponded.
4. **Form IV** (Factitive): *أَكْتَبَ* (aktaba) - He dictated.
5. **Form V** (Reflexive): *تَكَتَّبَ* (takataba) - It was written on its own.
6. **Form VI** (Reciprocity): *تَكَاتَبَ* (takaataba) - They wrote to each other.
7. **Form VII** (Submissive): *اِنْكَبَ* (inkataba) - He was subscribed.
8. **Form VIII** (Reciprocity): *اِكْتَبَ* (iktataba) - They wrote to each other.
9. **Form X** (Control): *اِسْتَكَبَ* (istakataba) - He asked to write.

The above example illustrates the derivational paradigm of Arabic word. However, there is also an inflectional paradigm, which is governed by the agreement information. Every entry of the Table 2.2, can take twelve inflectional form according to there number, gender and person. For imperfect form, there are three such inflectional paradigms. Table 2.3 and 2.4 show the inflectional paradigm for active perfect and passive perfect entry of form I.

Table 2.2: Derivational Paradigm of root **ktb**

	FORM I	FORM II	FORM III	FORM ...
Active perfect	كَتَبَ	كَتَّبَ	كَاتَبَ	...
Passive perfect	كُنِبَ	كُنَّبَ	كُوْتِبَ	...
Active imperfect	يَكْتُبُ	يُكْتَبُ	يُكَاتِبُ	...
Passive imperfect	يُكْتَبُ	يُكْتَبُ	يُكَاتِبُ	...
Active imperative	اُكْتُبْ	كُتِّبْ	كَاتِبْ	...
Passive imperative	اِنْكْتُبْ	اِنْكُتِّبْ	اِنْكَاتِبْ	...
Verbal noun	كِتَابَةٌ	تَكْتِيبٌ	مُكَاتَبَةٌ	...
Active participle	كَاتِبٌ	مُكْتَبٌ	مُكَاتِبٌ	...
Passive participle	مَكْتُوبٌ	مُكْتَبٌ	مُكَاتِبٌ	...
...

Figure 2.8: Derivational Paradigm of root **ktb**

An Arabic word can encode a complete sentence. For example, *سَيَكْتُبُهُ* (sayak-tubuhu) - *He will write it.* We can break the word in the following component.

From the Figure 2.8, we can conclude that an Arabic word has four components.

1. **Prefix:** *sa* – the particle indicating future
2. **Suffix:** *hu* – the object pronoun attached as a clitic
3. **Root:** *k, t, b* – the root letters bearing the concept of writing

Table 2.3: Inflectional Paradigm of Form I-Active-Perfect

Ind/Sub/Juss	Singular	Dual	Plural
3 rd /Masc.	كَتَبَ	كَتَبَا	كَتَبُوا
3 rd /Fem.	كَتَبَتْ	كَتَبَتَا	كَتَبْنَ
2 nd /Masc.	كَتَبْتِ	كَتَبْتُمَا	كَتَبْتُمْ
2 nd /Fem.	كَتَبْتِ	كَتَبْتُمَا	كَتَبْتُنَّ
1 st	كَتَبْتُ	-	كَتَبْنَا

Table 2.4: Inflectional Paradigm of Form I-Passive-Perfect

Ind/Sub/Juss	Singular	Dual	Plural
3 rd /Masc.	كُتِبَ	كُتِبَا	كُتِبُوا
3 rd /Fem.	كُتِبَتْ	كُتِبَتَا	كُتِبْنَ
2 nd /Masc.	كُتِبْتِ	كُتِبْتُمَا	كُتِبْتُمْ
2 nd /Fem.	كُتِبْتِ	كُتِبْتُمَا	كُتِبْتُنَّ
1 st	كُتِبْتُ	-	كُتِبْنَا

4. **Measure:** *ya__u__u* – bearing the syntactic and semantic information of the event

It may be possible to concatenate multiple prefixes and suffixes. However, there must be a single measure and single set of root letters, where the measure packages syntactic and semantic features and root supplies the core concept. If we plug in another set of root letters, for example, *ر، ص، ن* (*n, S, r*) which bears the concept of helping, we get *سَيَنْصُرُهُ* (*sayanSuruhu*) – *He will help him*.

Depending on this analysis, we can give the following model of an Arabic word [36].

A Root-Derived Arabic word = Prefix + Measure(Root) + Suffix

There are syntactic and semantic features, which governs the derivational and inflectional paradigms for Arabic roots. With a linguistic investigation, we list some features that we use in this thesis. Attributes in the Table 2.5, govern the derivational and inflectional paradigm for an Arabic root respectively.

Table 2.5: Attributes Governing Morphological Paradigm

	Attribute	Values
Attributes governing derivational paradigm	ROOT-TYPE POS FORM VOICE VFORM	sound, weak, hamzated, geminate, ... noun, verb, particle I, II, III, IV, ... active, passive perfect, imperfect, imperative
Attributes governing inflectional paradigm	MOOD PERSON NUMBER GENDER CASE DEFINITENESS POLARITY	indicative, subjunctive, jussive 1st, 2nd, 3rd singular, dual, plural masculine, feminine nominative, accusative, genitive definite, indefinite affirmative, negative

The attribute ROOT-TYPE indicates the characteristics of the constituent root letters of an Arabic word. For example, *sound* root contains only consonants excluding *hamza*. ROOT-TYPE drives the stem measures. The attribute POS gives the parts-of-speech tag of an Arabic word. Arabic has three types of POS; *noun*, *verb* and *particle*. There are different stem measures for nouns and verbs. Arabic exhibits the characteristics of having several derived forms of a single verb root with syntactic or semantic increases. They are represented by the attribute FORM and have their corresponding derivational paradigm with exclusive stem measures. There are around fifteen such forms among which ten of them are in common use. Finally within a single form, derivational paradigms exist according to the dimension of VOICE and VFORM. There are two types of VOICE in arabic; *active* and *passive*. There are three types of verb form in Arabic; *perfect* – indicates that the event has been completed, *imperfect* – indicates that the event has not yet been completed, and *imperative* – indicates that the event is a

command.

There are also attributes which governs the inflectional paradigm of an Arabic stem, settled by its corresponding attributes governing derivational paradigm. The attribute MOOD governs the diacritic of a verbs last character. There are three types of MOOD in Arabic; *indicative*, *subjunctive* and *jussive*. For nouns, this attribute is called as CASE. There are three types of CASE in Arabic; *nominative*, *accusative* and *genitive*. There are three semantic attributes which also govern a very regular inflectional paradigm of Arabic verbs; PERSON, NUMBER and GENDER. Unlike many other languages, these information is morphologically embedded in the Arabic verbs.

2.4 Summary

In this chapter, we have highlighted several topics including basic linguistic comprising morphology, syntax and semantics. We have explained the related linguistic terms. Moreover, we have also given an overview of HPSG and Arabic verbal morphology. These will help us to understand the rest of the thesis.

Chapter 3

An HPSG analysis of Arabic

Arabic verbs are grammatically complex object containing a myriad of linguistic information. HPSG modeling of such verbs is challenging and touches a broad range of topic in Arabic grammar. We could not use the de-facto tools for building resource grammars in HPSG such as LKB (Linguistic Knowledge Builder) due to their limitations in addressing the complex morphological operation such as infixation, stem alternation, etc, found in the nonconcatenative templatic morphology [2]. However, for our purpose, we present our analysis using the theoretically sound framework of HPSG. In the Section 3.1 we give our HPSG model for Arabic verbs. Section 3.2 gives an introduction of verbs in *perfect* form, Section 3.3 gives an introduction of verbs in *imperfect* form and Section 3.4 gives an introduction of verbs in *imperative* form.

3.1 Modeling of Arabic Verb

We propose the sign of Figure 3.1, to capture the attribute value matrix of a typical Arabic verb. We do not discuss PHON and CNTXT as they are out of our scope. Also the SEM feature has not changed from [27]. We have modified the SYN feature. We retain the verb form VF as VFORM but with a different value set where the possible values are perfect, imperfect and imperative. We eliminate the AUX feature since there are only two verbs that can act as auxiliary in Arabic. They should be taken care of in the construct. A newly introduced feature is VOICE with two possible values in its value set; active and passive. Arabic exhibit lexical passives. So, there are separate verb-stem for active and passive. Another newly introduced feature is MOOD that captures the

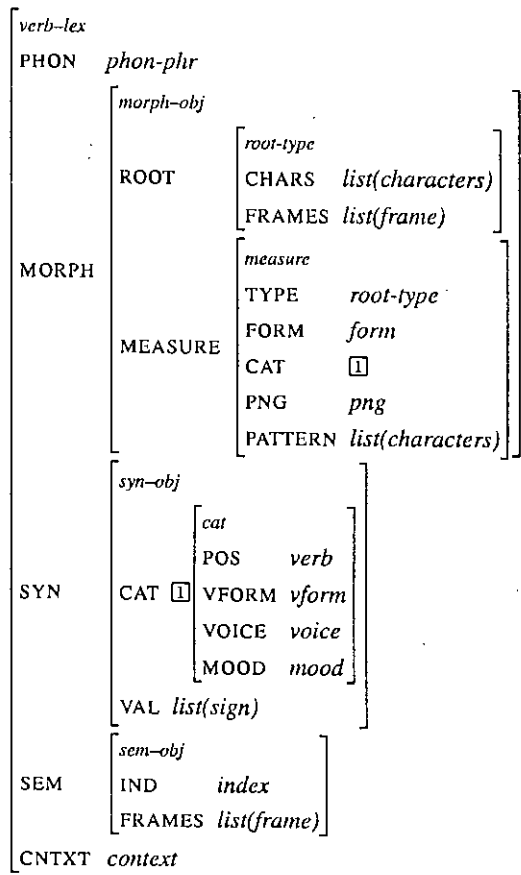


Figure 3.1: HPSG Sign for an Arabic verb

case-marking for verbs. Case-marking is expressed by the diacritic of the last letter in Arabic. There are three possible values for MOOD; indicative, subjunctive and jussive

We define the feature for morph-obj as MORPH, which captures the nonconcatenative templatic morphology. This introduces several new features. First, the ROOT feature which consists of two feature; CHARS and FRAMES. CHARS is the list of root letters. FRAMES is the list of predicates contributed to the semantic content. Next, the feature MEASURE which consists of five features: TYPE, FORM, PATTERN, CAT and PNG. Arabic roots are classified among several root classes according to their deriva-

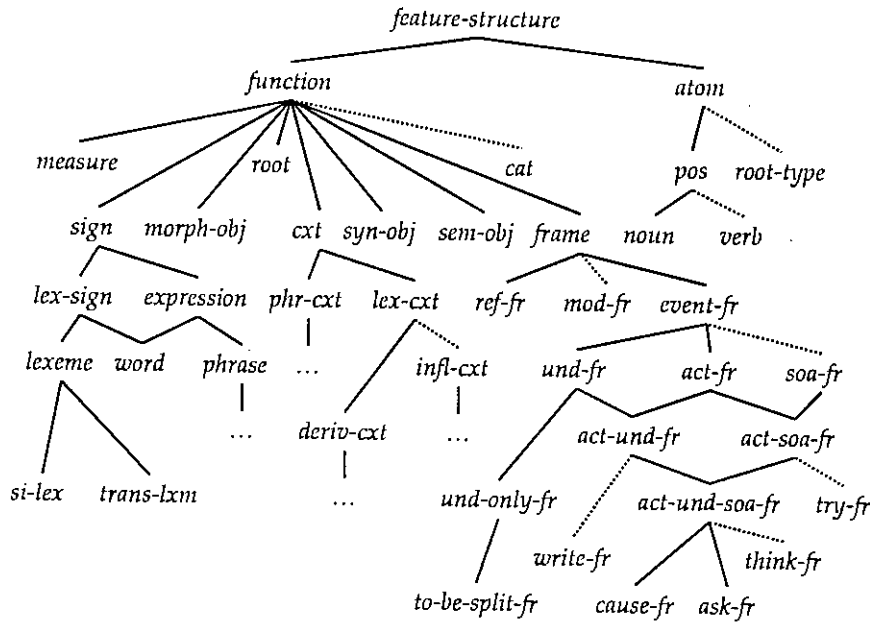


Figure 3.2: A partially modified HPSG type hierarchy for Arabic

tional and inflectional paradigm. The feature TYPE captures the corresponding root class. In this thesis, we only deal with the *sound* root class, where all of the the root letters are strong consonants. The feature FORM indicates one of the derived form of a base verb with corresponding semantic increase. The feature PATTERN contains a list of character capturing the morphological template where the root letters are embedded and co-indexed with the characters of CHARS feature. The feature CAT indicates the syntactic VFORM, VOICE and MOOD of the morphological measure. Its value is structure-shared with the syntactic CAT feature. Finally the PNG feature captures the PERSON, NUMBER and GENDER of the morphological measure which must agree with the PNG of the reference frame in FRAMES. Depending on the above discussion we now propose the partial Type Hierarchy for Arabic. In the Figure 3.2, we give the core HPSG hierarchy of feature structure.

Now, we give an illustrative example of lexical entries for a sample Arabic verb according to our formalism. We take the Arabic tri-lateral root ب، ت، ك (k, t, b) for this purpose.

3.2 Perfect Form

In *perfect* verb form the root ب، ت، ك (k, t, b) gives the Arabic verb كَتَبَ (*kataba*) – *He wrote*. We give the corresponding attribute value matrix in Figure 3.3.

The AVM of the form-I base entry for Arabic root ب، ت، ك (k, t, b) is illustrated in Figure 3.3. Among the morphologically associated features of MORPH, first comes the feature ROOT. This feature contains the list CHARS of root letters *k, t, b* as well as the CONTENT feature that gives the semantic contribution made by CHARS. In this example, the value of CONTENT is structure-shared with the predicate *write-fr* in the FRAMES feature. This indicates, the core meaning that the root letters contribute is somehow associated with the concept of writing. Next comes the feature MEASURE, which contains the morphological, syntactic and semantic information contributed by the templatic measures. First, the feature TYPE, which denotes the associated root class. This feature is determined by the root letters in CHARS and affects their corresponding available MEASUREs. Not every MEASURE is available to every root class. In our case, its value is *sound* denoting that all root letters are strong consonants. Next, the feature FORM, which denotes the specific derivational paradigm of corresponding root. Here, *kataba* - is a FORM-I derivative. Next, the feature PATTERN, which captures the stem of the lexeme along with the root letters of CHARS using structure sharing. Here, its value is *_a_a_a*. The blanks inside the measure indicate the placeholder for corresponding root characters. Then, the feature CAT, which contains the syntactic category for this measure. Its value is structure-shared with the syntactic feature CAT. Finally, the feature PNG, which captures the PERSON, NUMBER and GENDER information

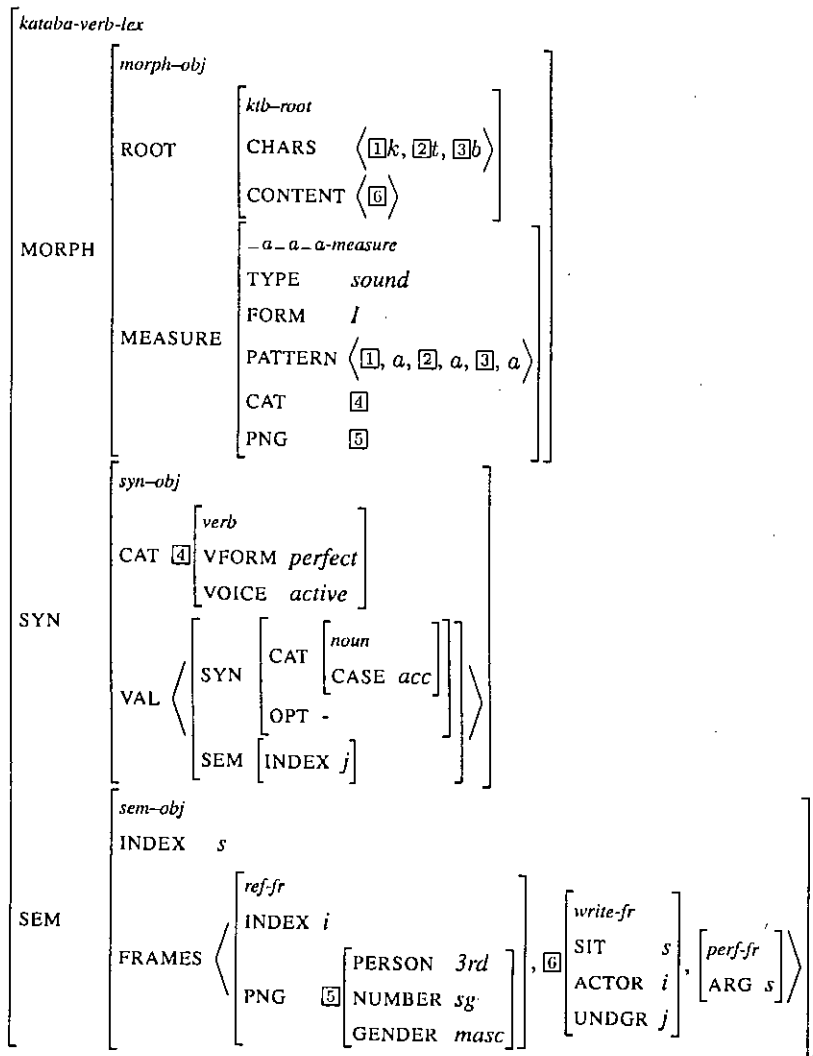


Figure 3.3: HPSG Sign for *kataba*

of our semantic agent in the case where it is not syntactically realized. In this case, its value is *3rd/Singular/Masculine*.

We present the syntactic and semantic information using the SYN and SEM feature, for the word - *kataba*. First, the CAT features identifies the syntactic category of -

kataba. It contains the VFORM and VOICE feature of Arabic, which governs the derivational paradigm of verb lexeme. In this case their values are *perfect* and *active* respectively. Next, the VAL feature, which captures the subcategorization of verbs. VAL is a list of signs, which are required by the syntactic head. In this case, the verb - *kataba*, requires an object. The verb - *write* is a transitive verb that takes an object. We should note that the hidden pronoun. *he* is encoded by the inflectional morphology, when no explicit subject is used. The semantic actor is not realized syntactically. So, the verb only subcategorizes for syntactic object. We can also see the constraints imposed over the object. In this version, its syntactic head should be a noun phrase with the value of its CASE feature set to *accusative*. The negative value of the OPT feature indicates that this object is not optional, rather required to be syntactically correct.

Next, we need to consider some semantic features. Here, we use a type feature version of predicate logic to capture semantics of natural language. First, we consider the INDEX feature, which is a reference to a discourse entity. Then, the PNG feature, which capture the semantics of PERSON, NUMBER and GENDER. Next, the FRAMES feature, which serves as a bag for elementary predicates to describe the situation at hand. For example, in the case of *kataba*, the event of writing is expressed. The event is completed in the past and there is a discourse referent to the actor. To capture the core event, *write-fr* is introduced. To capture the temporal constraint, we use the *perf-fr*. Finally, to express the actor of the event, the hidden pronoun, we introduce a discourse referent with corresponding PNG feature. Predicates have their respective arguments. *write-fr* has a situation hook, expressed by the feature SIT. There are two semantic role associated with this predicate. First, we consider the role of writer, who plays a doer role, expressed by the feature ACTOR. Second, we consider the role of written, who plays an undergoer role, expressed by the feature UNDGR. The *perf-fr* takes a situation hook as an argument, which is expressed as the feature ARG. We use the technique of co-indexing for sharing semantic objects. The discourse referent predicate is actually the actor of the *write-fr*. To denote this constraint, the INDEX value of hidden pronoun and the ACTOR value of the write-predicate are co-indexed, both are given the value *i*. This is an example of reference co-indexing. We also use event co-indexing. The event hook SIT of *write-fr*, situation hook of the entire scenario and argument ARG of the *perf-fr*, all are co-indexed and expressed using the value *s*. Another important issue of HPSG representation is the syntax-semantics interface. In this example, this is done by co-indexing the INDEX value of the syntactic object and the UNDGR value of the

write-fr with a value j . This indicates that the syntactic object is our semantic undergoer whereas from our previous discussion we can note that the semantic actor is not syntactically realized.

3.3 Imperfect Form

In *imperfect* verb form the root k, t, b gives the Arabic verb **يَكْتُبُ** (*yaktubu*) – *He writes or will write*. We give the corresponding attribute value matrix in Figure 3.4.

We have all the three features associated with morphology in the Figure 3.4. First, the feature TYPE, which denotes the associated root class, takes its value *sound*. Next, the feature ROOT, which is the list of root letters k, t, b as well as the CONTENT feature giving the semantic contribution made by root letters. Here, its value is structure-shared with the *write-fr* in the FRAMES feature. This indicates that the core meaning which the root letters contribute are somehow associated with the concept of writing. Next, the feature MEASURE, which contains the morphological, syntactic and semantic information contributed by measure. First, the feature FORM, which denotes the semantic paradigm of *yaktubu* as a FORM-I derivative. Next, the feature PATTERN captures the stem measure *ya__u__u*. Then, the feature CAT, which contains the syntactic category for this measure, structure-shared with the syntactic feature CAT. Finally, the feature PNG, which captures the PERSON, NUMBER and GENDER information of our semantic actor whose value is *3rd/Singular/Masculine*.

The SYN feature captures syntactic information. First, the CAT features identifies the syntactic category of - *yaktubu*. Its VFORM and VOICE features which govern the derivational paradigm of verb lexeme, take the values *imperfect* and *active* respectively. A new feature MOOD is also introduced which is available in *imperfect* form. The value of MOOD is set to *indicative*. Next, the VAL feature, which captures the subcategorization of verbs. As a transitive verb it requires an object. We should also note that the hidden pronoun, *he* is encoded by the inflectional morphology, when no explicit subject is used. The semantic actor is not realized syntactically. So, the verb only subcategorizes for syntactic object. The object's syntactic head should be a noun phrase with the value of its CASE feature set to *accusative*. The negative value of the OPT feature indicates that this object is not optional, rather required to be syntactically correct.

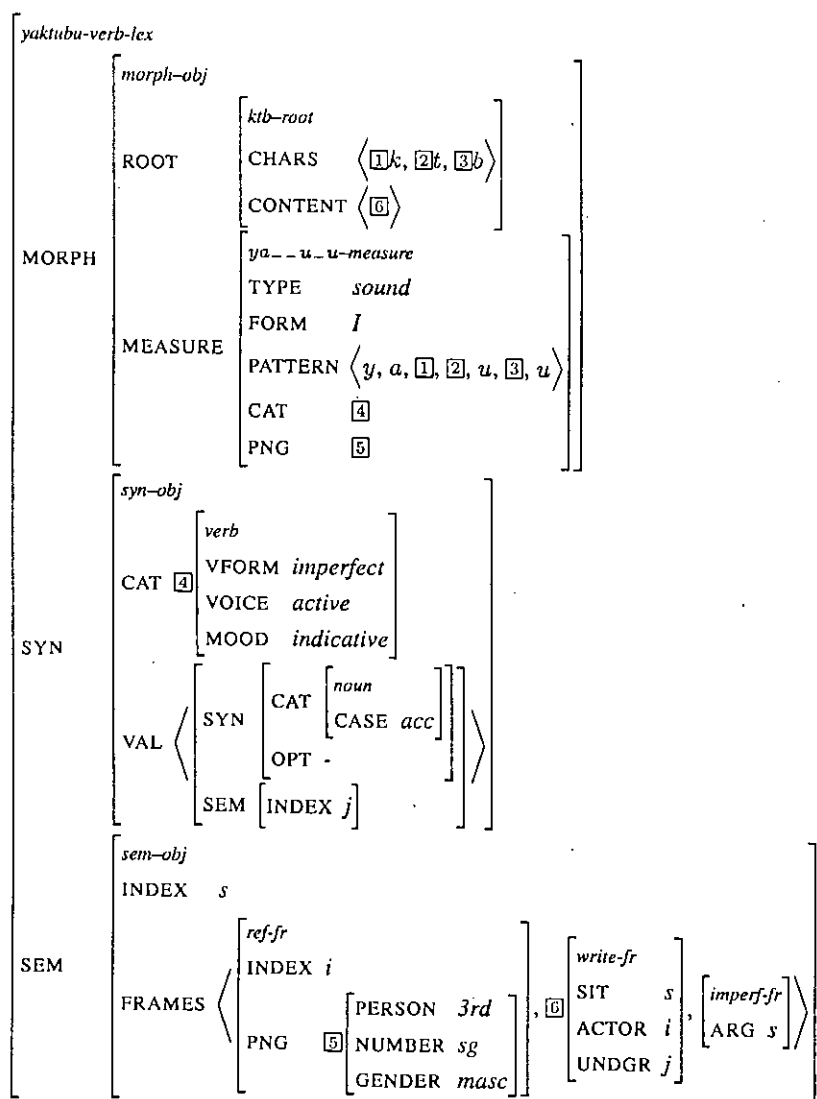


Figure 3.4: HPSG Sign for *yaktubu*

The SEM feature captures semantic information. In the case of *yaktubu*, the event of writing is expressed. The event has not yet been *completed* and there is a discourse referent to the *actor*. To capture the core event, *write-fr* is introduced. To capture the

temporal constraint, we use the *imperf-fr*. Finally, to express the *actor* of the event, the hidden pronoun, we introduce a discourse referent with corresponding PNG feature. The *imperf-fr* takes the situation hook *s* of the *write-fr* as an argument, which is expressed as the feature ARG. The INDEX value of discourse referent and the ACTOR value of the *write-fr* are co-indexed, both are given the value *i*. We express the syntax-semantic interface by co-indexing the INDEX value of the syntactic object and the UNDGR value of the *write-fr* with a value *j*. This indicates that the syntactic object is our semantic undergoer whereas from our previous discussion we can note that the semantic actor is not syntactically realized.

3.4 Imperative Form

In *imperative* verb form the root *k, t, b* gives the Arabic verb **اكتب** (*uktub*) – *Write!*. We give the corresponding attribute value matrix in Figure 3.5.

We have all the three features associated with morphology in the Figure 3.5. First, the feature TYPE, which denotes the associated root class, takes its value *sound*. Next, the feature ROOT, which is the list of root letters *k, t, b* as well as the CONTENT feature giving the semantic contribution made by root letters. Here, its value is structure-shared with the *write-fr* in the FRAMES feature. This indicates that the core meaning which the root letters contribute are somehow associated with the concept of writing. Next, the feature MEASURE, which contains the morphological, syntactic and semantic information contributed by measure. First, the feature FORM, which denotes the semantic paradigm of *uktub* as a FORM-I derivative. Next, the feature PATTERN captures the stem measure *u__u_*. Then, the feature CAT, which contains the syntactic category for this measure, structure-shared with the syntactic feature CAT. Finally, the feature PNG, which captures the PERSON, NUMBER and GENDER information of our semantic actor whose value is *2nd/Singular/Masculine*.

The SYN feature captures syntactic information. First, the CAT features identifies the syntactic category of - *uktub*. Its VFORM and VOICE features which govern the derivational paradigm of verb lexeme, take the values *imperative* and *active* respectively. The feature MOOD is also introduced which is also available in *imperfect* form. The value of MOOD is set to *jussive*. Next, the VAL feature, which captures the subcategorization of verbs. As a transitive verb it requires an object. We should also note that the

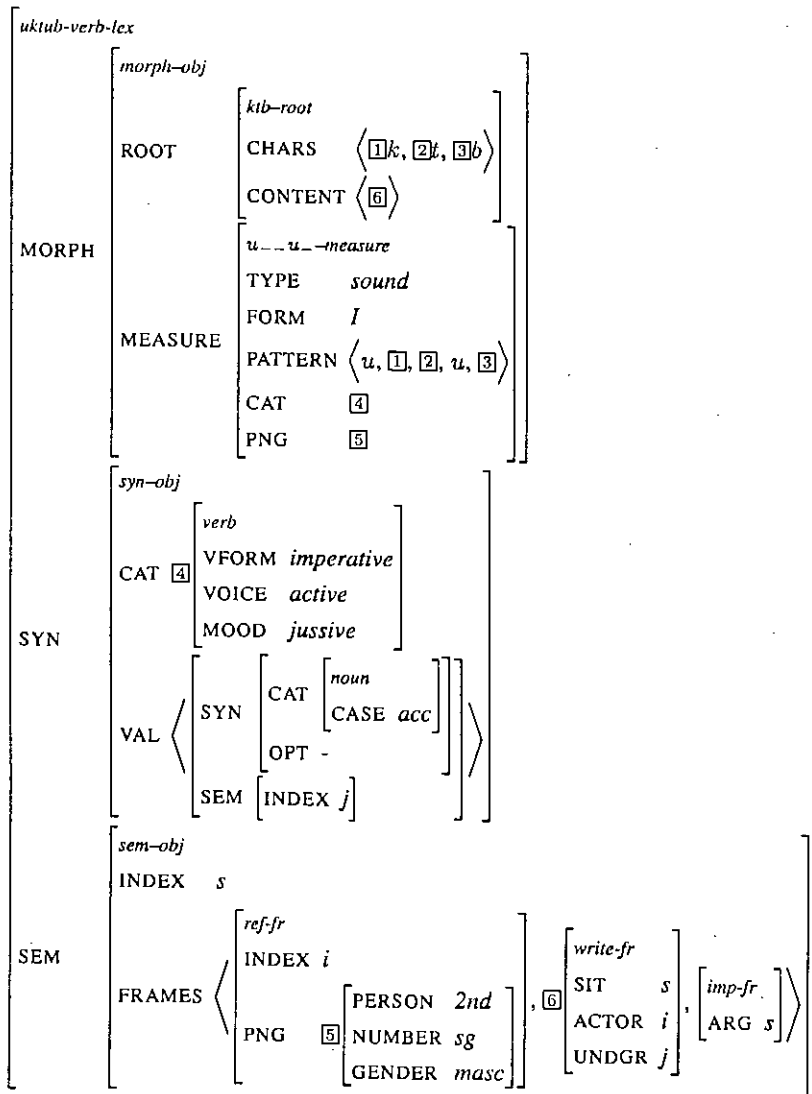


Figure 3.5: HPSG Sign for *uktub*

hidden pronoun, *he* is encoded by the inflectional morphology, when no explicit subject is used. The semantic actor is not realized syntactically. So, the verb only subcategorizes for syntactic object. The object's syntactic head should be a noun phrase with

the value of its CASE feature set to *accusative*. The negative value of the OPT feature indicates that this object is not optional, rather required to be syntactically correct.

The SEM feature captures semantic information. In the case of *uktub*, the event of writing is expressed. The event is a *command* and there is a discourse referent to the *actor*. To capture the core event, *write-fr* is introduced. To capture the modal constraint, we use the *imp-fr*. Finally, to express the *actor* of the event, the hidden pronoun, we introduce a discourse referent with corresponding PNG feature. The *imp-fr* takes the situation hook *s* of the *write-fr* as an argument, which is expressed as the feature ARG. The INDEX value of discourse referent and the ACTOR value of the *write-fr* are co-indexed, both are given the value *i*. We express the syntax-semantic interface by co-indexing the INDEX value of the syntactic object and the UNDGR value of the *write-fr* with a value *j*. This indicates that the syntactic object is our semantic undergoer whereas from our previous discussion we can note that the semantic actor is not syntactically realized.

3.5 Summary

In this chapter, we have given an HPSG analysis for Arabic verb as well as presented a generic model for them. We have also explained our model in case of three Arabic verb form including *perfect*, *imperfect* and *imperative* in form I. However, they can also be extended to other forms.

Chapter 4

Agency in Arabic

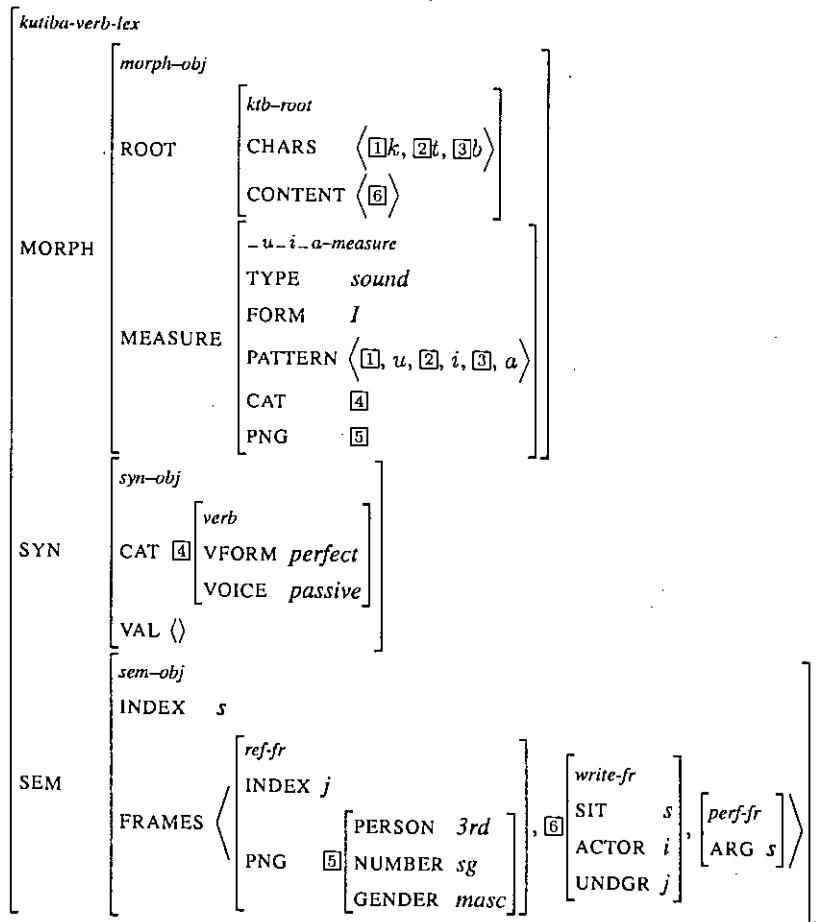
The analysis of agency has always enjoyed deep investigations from linguistics. Agency express the characters associated with the situation and their orientation. In this chapter, we discuss three type of Arabic agency. In the Section 4.1 we explain HPSG construction of Arabic passive. In the Section 4.2 we explain HPSG construction of Arabic causative. In the Section 4.3 we explain HPSG construction of Arabic reflexive.

4.1 Arabic Passive

We give here, the attribute value matrix (AVM) for the passive form *كُتِبَ* (*kutiba*) – *It was written in perfect form* and *يُكْتَبُ* (*yuktabu*) – *It is written or will be written* for the Arabic verb *كَتَبَ* (*kataba*) – *He wrote* and *يَكْتُبُ* (*yaktubu*) – *He writes or will write*.

4.1.1 Passive Perfect

We have all the three features associated with morphology in the Figure 4.1. First, the feature TYPE, which denotes the associated root class, takes its value *sound*. Next, the feature ROOT, which is the list of root letters *k*, *t*, *b* as well as the CONTENT feature giving the semantic contribution made by root letters. Here, its value is structure-shared with the *write-fr* in the FRAMES feature. This indicates that the core meaning which the root letters contribute are somehow associated with the concept of writing. Next, the feature MEASURE, which contains the morphological, syntactic and semantic information contributed by measure. First, the feature FORM, which denotes the semantic

Figure 4.1: HPSG Sign for *kutiba*

paradigm of *kutiba* as a FORM-I derivative. Next, the feature PATTERN captures the stem measure *-u-i-a*. Then, the feature CAT, which contains the syntactic category for this measure, structure-shared with the syntactic feature CAT. Finally, the feature PNG, which captures the PERSON, NUMBER and GENDER information of our semantic *undergoer* whose value is *3rd/Singular/Masculine*.

The SYN feature captures syntactic information. First, the CAT features identifies the syntactic category of *-kutiba*. Its VFORM and VOICE features which govern the

derivational paradigm of verb lexeme, take the values *perfect* and *passive* respectively. Next, the VAL feature, which captures the subcategorization of verbs. Unlike English, which can have a prepositional complement in passives, Arabic passives do not subcategorize for a subject or any other argument. For this reason, the VAL list is empty. We should also note that the hidden pronoun, *it* is encoded by the inflectional morphology, when no explicit subject is used. The semantic actor is not realized syntactically.

The SEM feature captures semantic information. In the case of *kutiba*, the event of writing is expressed. The event has been *completed* and there is a discourse referent to the *undergoer*. To capture the core event, *write-fr* is introduced. To capture the temporal constraint, we use the *perf-fr*. Finally, to express the *undergoer* of the event, the hidden pronoun, we introduce a discourse referent with corresponding PNG feature. The *perf-fr* takes the situation hook *s* of the *write-fr* as an argument, which is expressed as the feature ARG. discourse referent in the feature FRAMES is now co-indexed with the UNDGR feature of the *write-fr*, expressed by the value *j*. Semantic actor *i* is now completely unknown by not having any syntactic or semantic reference, which is a distinctive property of Arabic passive.

4.1.2 Passive Imperfect

We have all the three features associated with morphology in the Figure 4.2. First, the feature TYPE, which denotes the associated root class, takes its value *sound*. Next, the feature ROOT, which is the list of root letters *k, t, b* as well as the CONTENT feature giving the semantic contribution made by root letters. Here, its value is structure-shared with the *write-fr* in the FRAMES feature. This indicates that the core meaning which the root letters contribute are somehow associated with the concept of writing. Next, the feature MEASURE, which contains the morphological, syntactic and semantic information contributed by measure. First, the feature FORM, which denotes the semantic paradigm of *yuktabu* as a FORM-I derivative. Next, the feature PATTERN captures the stem measure *yu...a...u*. Then, the feature CAT, which contains the syntactic category for this measure, structure-shared with the syntactic feature CAT. Finally, the feature PNG, which captures the PERSON, NUMBER and GENDER information of our semantic *undergoer* whose value is *3rd/Singular/Masculine*.

The SYN feature captures syntactic information. First, the CAT features identifies the syntactic category of - *yuktabu*. Its VFORM and VOICE features which govern

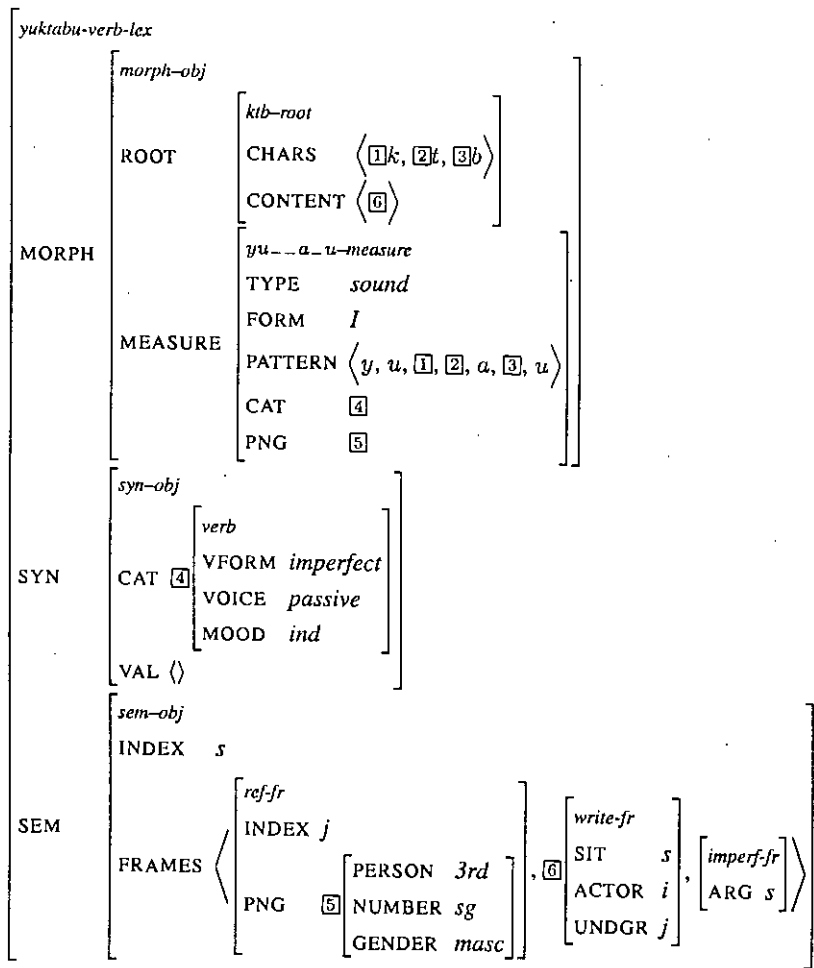


Figure 4.2: HPSG Sign for *yuktabu*

the derivational paradigm of verb lexeme, take the values *imperfect* and *passive* respectively. The MOOD feature is also introduced which is available in *imperfect* form. The value of MOOD is set to *indicative*. Next, the VAL feature, which captures the subcategorization of verbs. Unlike English, which can have a prepositional complement in passives, Arabic passives do not subcategorize for a subject or any other argument. For this reason, the VAL list is empty. We should also note that the hidden pronoun, *it* is

encoded by the inflectional morphology, when no explicit subject is used. The semantic actor is not realized syntactically.

The SEM feature captures semantic information. In the case of *yuktabu*, the event of writing is expressed. The event has not yet been *completed* and there is a discourse referent to the *undergoer*. To capture the core event, *write-fr* is introduced. To capture the temporal constraint, we use the *imperf-fr*. Finally, to express the *undergoer* of the event, the hidden pronoun, we introduce a discourse referent with corresponding PNG feature. The *imperf-fr* takes the situation hook *s* of the *write-fr* as an argument, which is expressed as the feature ARG. discourse referent in the feature FRAMES is now co-indexed with the UNDGR feature of the *write-fr*, expressed by the value *j*. Semantic actor *i* is now completely unknown by not having any syntactic or semantic reference.

4.1.3 Passive Imperative

We have all the three features associated with morphology in the Figure 4.3. First, the feature TYPE, which denotes the associated root class, takes its value *sound*. Next, the feature ROOT, which is the list of root letters *k, t, b* as well as the CONTENT feature giving the semantic contribution made by root letters. Here, its value is structure-shared with the *write-fr* in the FRAMES feature. This indicates that the core meaning which the root letters contribute are somehow associated with the concept of writing. Next, the feature MEASURE, which contains the morphological, syntactic and semantic information contributed by measure. First, the feature FORM, which denotes the semantic paradigm of *lituktab* as a FORM-I derivative. Next, the feature PATTERN captures the stem measure *litu__a__*. Then, the feature CAT, which contains the syntactic category for this measure, structure-shared with the syntactic feature CAT. Finally, the feature PNG, which captures the PERSON, NUMBER and GENDER information of our semantic *undergoer* whose value is *2nd/Singular/Masculine*.

The SYN feature captures syntactic information. First, the CAT features identifies the syntactic category of *-lituktab*. Its VFORM and VOICE features which govern the derivational paradigm of verb lexeme, take the values *imperative* and *passive* respectively. The MOOD feature is also introduced which is available in *imperative* form. The value of MOOD is set to *jussive*. Next, the VAL feature, which captures the subcategorization of verbs. Unlike English, which can have a prepositional complement in passives, Arabic passives do not subcategorize for a subject or any other argument. For

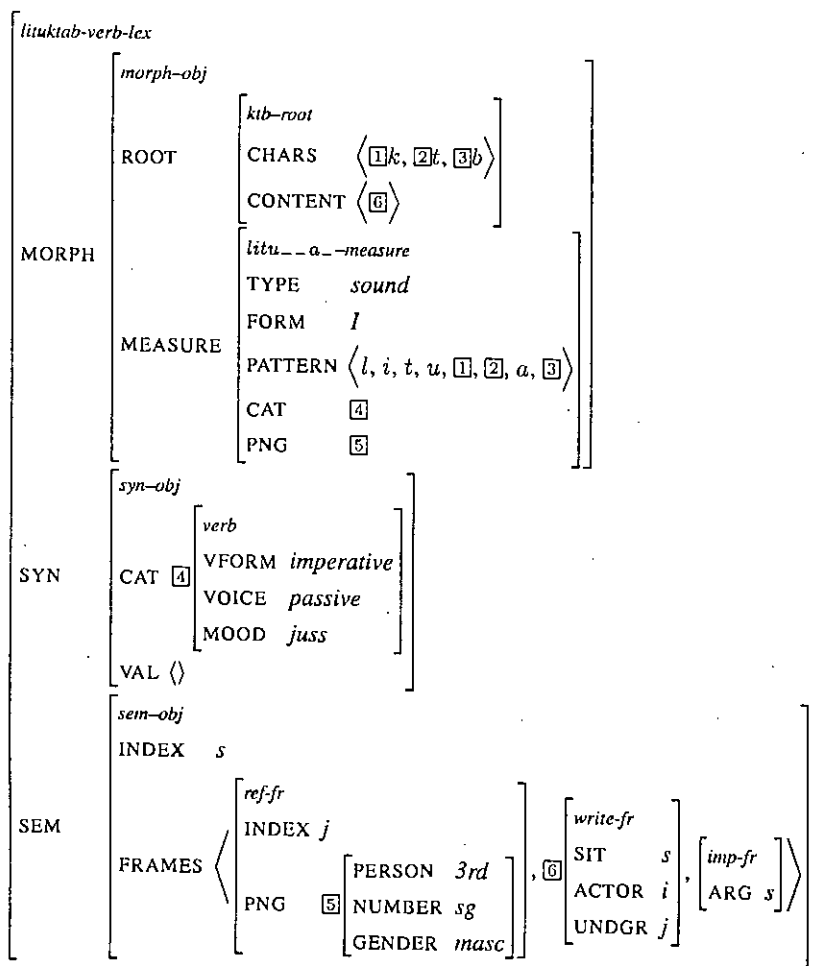


Figure 4.3: HPSG Sign for *lituktab*

this reason, the VAL list is empty. We should also note that the hidden pronoun, *it* is encoded by the inflectional morphology, when no explicit subject is used. The semantic actor is not realized syntactically.

The SEM feature captures semantic information. In the case of *lituktab*, the event of writing is expressed. The event expresses a *command* and there is a discourse referent to the *undergoer*. To capture the core event, *write-fr* is introduced. To capture the modal

constraint, we use the *imp-fr*. Finally, to express the *undergoer* of the event, the hidden pronoun, we introduce a discourse referent with corresponding PNG feature. The *imp-fr* takes the situation hook *s* of the *write-fr* as an argument, which is expressed as the feature ARG. discourse referent in the feature FRAMES is now co-indexed with the UNDGR feature of the *write-fr*, expressed by the value *j*. Semantic actor *i* is now completely unknown by not having any syntactic or semantic reference.

4.2 Arabic Causative

Causative construction is one of the highly investigated research area within the domain of theoretical and computational linguistics. HPSG is used in successful analysis of causative for several languages [22, 23]. However, there is no HPSG analysis for Arabic causative. Arabic verb system exhibit several lexical causative. In this section, we extend the HPSG framework to support Arabic causative along with its nonconcatenative morphology. Arabic causatives come from multiple verb forms. Among them, we discuss form II.

A range of verbs take causative meaning in form II. This is the most used form of causative. We give a few examples in the following.

(4.1) كَتَبَ (*kataba*) – to write ... كَتَّبَ (*kattaba*) – to make (someone) write

(4.2) فَزَّقَ (*faraqa*) – to split ... فَزَّقَ (*farraqa*) – to make (something) split

(4.3) عَلَّمَ (*3alima*) – to learn ... عَلَّمَ (*3allama*) – to make (someone) learn

(4.4) بَلَغَ (*balaga*) – to reach ... بَلَغَ (*ballaga*) – to make (something) reach

(4.5) فَضَّلَ (*faDala*) – to be superior ... فَضَّلَ (*faDDala*) – to make (someone) be superior

(4.6) *ظَهَرَ* (*Tahara*) – to be pure . . . *ظَهَّرَ* (*Tahhara*) – to make (something) be pure

(4.7) *كَثُرَ* (*kaCura*) – to be numerous . . . *كَثَّرَ* (*kaCCara*) – to make (something) numerous

Causatives can also have active and passive form, since causatives are always transitive. In the following two sections we give the HPSG construction for both of them.

4.2.1 Active Causative

Here we give the attribute value matrix (AVM) for the causative form *كَتَبَ* (*kattaba*) – *He caused (someone) to write* for the Arabic verb *كَتَبَ* *kataba* – *He wrote*.

We have all the three features associated with morphology in the Figure 4.4. First, the feature TYPE, which denotes the associated root class, takes its value *sound*. Next, the feature ROOT, which is the list of root letters *k*, *t*, *b* as well as the CONTENT feature giving the semantic contribution made by root letters. Here, its value is structure-shared with the *write-fr* in the FRAMES feature. This indicates that the core meaning which the root letters contribute are somehow associated with the concept of writing. Next, the feature MEASURE, which contains the morphological, syntactic and semantic information contributed by measure. First, the feature FORM, which denotes the semantic paradigm of *kattaba* as a FORM-II derivative. Next, the feature PATTERN captures the stem measure *_a__a_a_*. Then, the feature CAT, which contains the syntactic category for this measure, structure-shared with the syntactic feature CAT. Finally, the feature PNG, which captures the PERSON, NUMBER and GENDER information of our semantic actor whose value is *3rd/Singular/Masculine*.

The SYN feature captures syntactic information. First, the CAT features identifies the syntactic category of *-kattaba*. Its VFORM and VOICE features which govern the derivational paradigm of verb lexeme, take the values *perfect* and *active* respectively. Next, the VAL feature, which captures the subcategorization of verbs. Generally, causative verbs subcategorizes for at least one argument, *the causee*, with another optional complement which is the result of the cause. In this version we capture this by specifying one compulsory and one optional argument in the VAL list. Unlike English, Arabic causatives are realized lexically and they increase their requirement of syntactic arguments. Here, the original transitive verb subcategorizes for single object and in

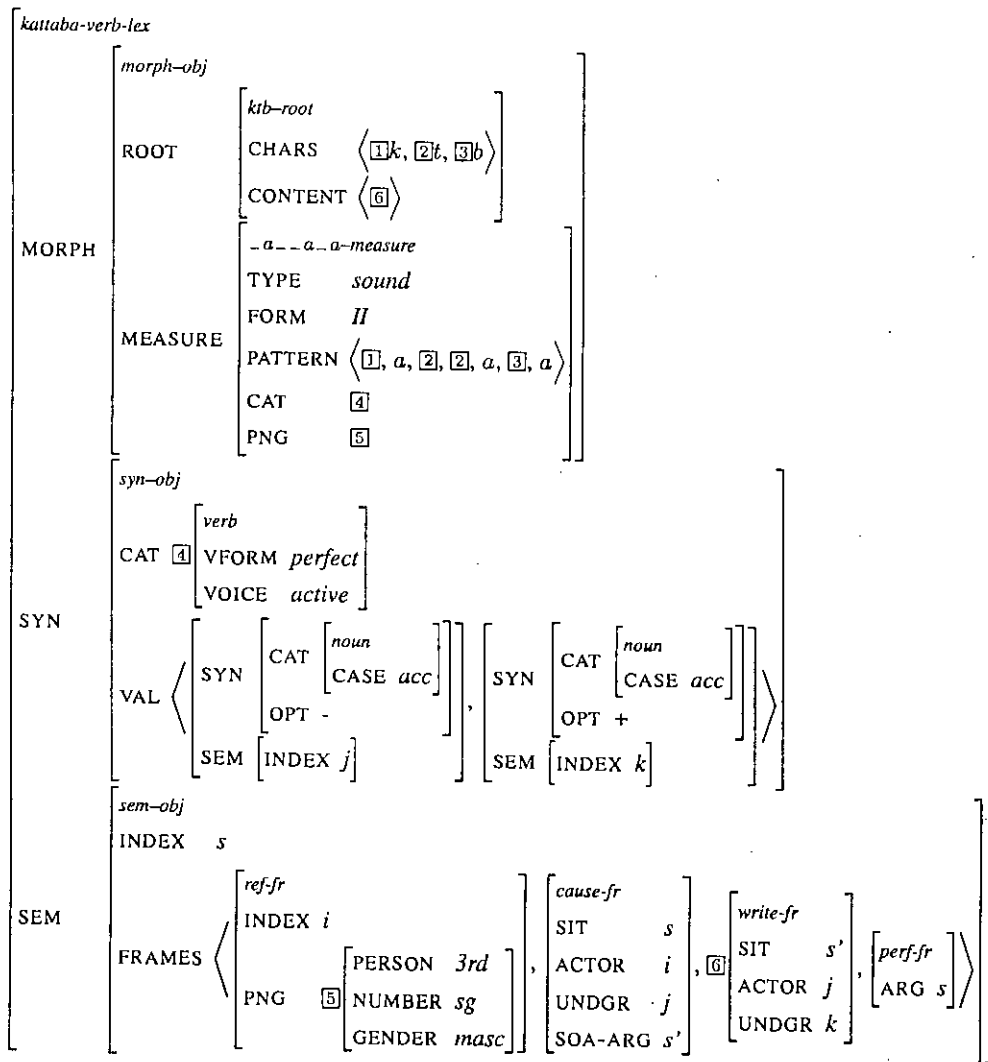


Figure 4.4: HPSG Sign for *kattaba*

causative version it becomes doubly transitive. We should also note that the hidden pronoun, *he* is encoded by the inflectional morphology, when no explicit subject is used. The semantic actor is not realized syntactically. So, the verb only subcategorizes for syntactic objects. The syntactic head of the objects should be a noun phrase with the

value of its CASE feature set to *accusative*.

The SEM feature captures semantic information. In the case of *kattaba*, the event of writing as well as the event of causation is expressed. The events have been *completed* and there is a discourse referent to the *causer*. In our analysis, the semantics of causative construction introduces at least two predicates; first, the base predicate and the complement predicate. In this case the base predicate is *cause-fr* and the complement or core event is *write-fr*. This phenomenon is called *morphologically complex predicate* [35]. To capture the temporal constraint, we use the *perf-fr*. This *cause-fr* is a semantic increase of the FRAMES feature. The discourse referent in the *ref-fr* of the feature FRAMES is co-indexed with the value of ACTOR feature of the *cause-fr*, both are given the value *i*. This indicates that the *causer* is the actor of this situation, expressed by the value *i*. The UNDGR *j* of the *cause-fr* is co-indexed with the ACTOR of the *write-fr* which is the complement predicate. SOA-ARG feature in the *cause-fr* and the event hook SIT of *write-fr* are co-indexed using the value *s'*. This indicates that the event of writing is the consequence or result of the event of causing. The *perf-fr* takes the situation hook *s* of the *cause-fr* as an argument, which is expressed as the feature ARG. We express the syntax-semantic interface by co-indexing the INDEX of the first syntactic object and the UNDGR value of the *cause-fr* with a value *j*. This indicates that the first syntactic object is *causee* whereas we can note that the semantic actor is not syntactically realized. We express another syntax-semantic interface by co-indexing the INDEX of the second syntactic object and the UNDGR value of the *write-fr* with a value *k*. This indicates that the second syntactic object is *undergoer* of the *write-fr*.

4.2.2 Passive Causative

Here we give the attribute value matrix (AVM) for the passive causative form كُتِّبَ (*kuttiba*) – *He was caused to write (by someone)* for the Arabic verb كَتَبَ *kataba* – *He wrote*.

We have all the three features associated with morphology in the Figure 4.5. First, the feature TYPE, which denotes the associated root class, takes its value *sound*. Next, the feature ROOT, which is the list of root letters *k, t, b* as well as the CONTENT feature giving the semantic contribution made by root letters. Here, its value is structure-shared with the *write-fr* in the FRAMES feature. This indicates that the core meaning which the root letters contribute are somehow associated with the concept of writing. Next,

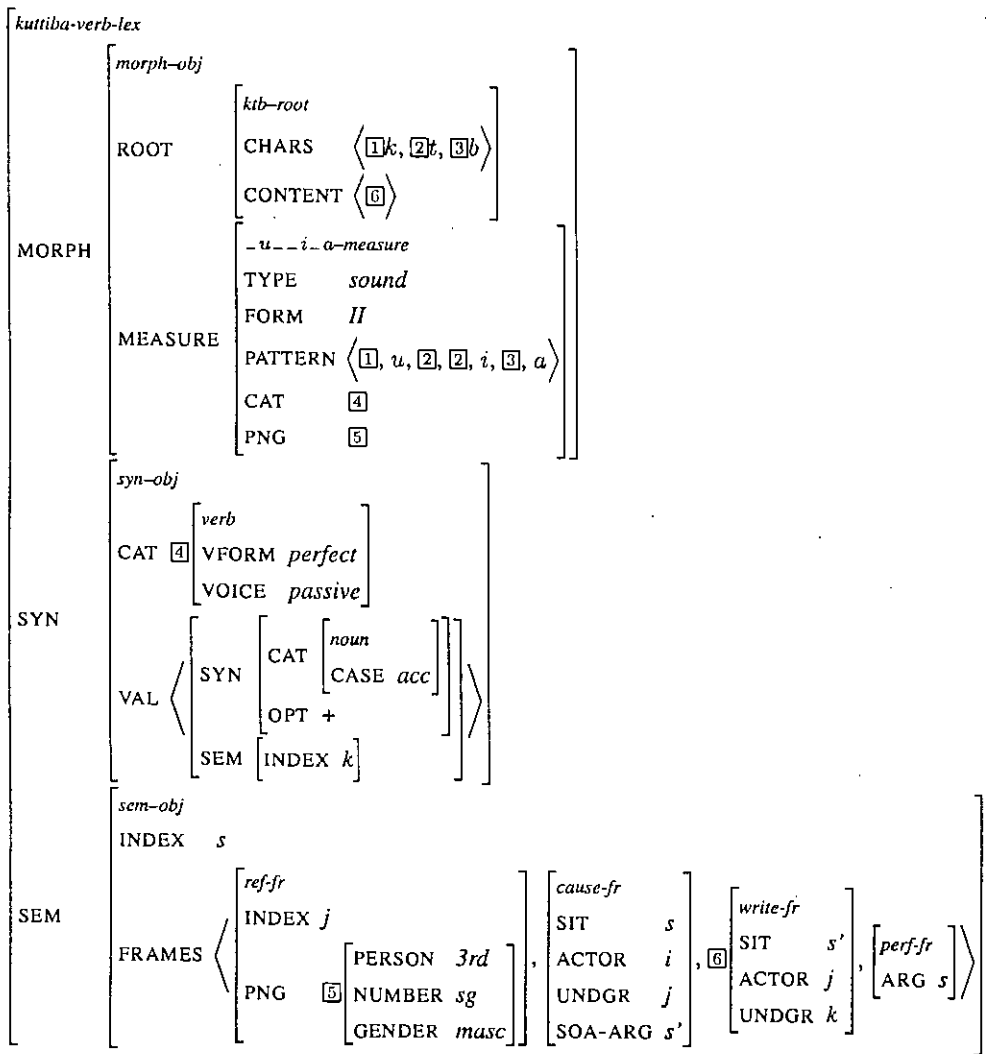


Figure 4.5: HPSG Sign for *kuttiba*

the feature MEASURE, which contains the morphological, syntactic and semantic information contributed by measure. First, the feature FORM, which denotes the semantic paradigm of *kuttiba* as a FORM-II derivative. Next, the feature PATTERN captures the stem measure *-u--i-a*. Then, the feature CAT, which contains the syntactic category

for this measure, structure-shared with the syntactic feature CAT. Finally, the feature PNG, which captures the PERSON, NUMBER and GENDER information of our semantic *undergoer* whose value is *3rd/Singular/Masculine*.

The SYN feature captures syntactic information. First, the CAT features identifies the syntactic category of *-kuttiba*. Its VFORM and VOICE features which govern the derivational paradigm of verb lexeme, take the values *perfect* and *passive* respectively. Next, the VAL feature, which captures the subcategorization of verbs. Unlike English, which can have a prepositional complement in passives, Arabic passives do not subcategorize for the previous subject. So, the passive of causative verbs does not subcategorize for the syntactic argument expressing *the causer*. However, the previous optional complement which is the result of the cause, still remains. Here, the passive of original transitive verb subcategorizes for no object and in causative version it becomes singly transitive. We should also note that the hidden pronoun, *he* is encoded by the inflectional morphology, when no explicit subject is used. The semantic undergoer is not realized syntactically. So, the verb only subcategorizes for syntactic object. The syntactic head of the objects should be a noun phrase with the value of its CASE feature set to *accusative*. In this version we capture this by specifying one optional argument in the VAL list.

The SEM feature captures semantic information. In the case of *kuttiba*, the event of writing as well as the event of causation is expressed. The events have been *completed* and there is a discourse referent to the *causee*. Here, the base predicate is *cause-fr* and the complement or core event is *write-fr*. To capture the temporal constraint, we use the *perf-fr*. This *cause-fr* is a semantic increase of the FRAMES feature. The discourse referent in the *ref-fr* of the feature FRAMES is co-indexed with the value of UNDGR feature of the *cause-fr*, both are given the value *j*. This indicates that the *causee* is the undergoer of this situation, expressed by the value *j*. The UNDGR *j* of the *cause-fr* is also co-indexed with the ACTOR of the *write-fr* which is the complement predicate. SOA-ARG feature in the *cause-fr* and the event hook SIT of *write-fr* are co-indexed using the value *s'*. This indicates that the event of writing is the consequence or result of the event of causing. The *perf-fr* takes the situation hook *s* of the *cause-fr* as an argument, which is expressed as the feature ARG. We express the syntax-semantic interface by co-indexing the INDEX of the syntactic object and the UNDGR value of the *write-fr* with a value *k*. This indicates that the syntactic object is *writee*. Semantic causer *i* is now completely unknown by not having any syntactic or semantic reference,

which is a distinctive property of Arabic passive.

4.3 Arabic Reflexive

Like the previous fields, reflexive construction also draws attention of many researchers in the field of theoretical linguistic. HPSG is used in successful analysis of reflexive for several languages [28–30]. However, there is no HPSG analysis for Arabic passives. Arabic verb system exhibit lexical passive. Arabic has a rich system for expressing reflexive. When the discourse reference of ACTOR and UNDGR refer to the same entity in the real world that is called reflexive construction. Arabic reflexives can also come from multiple verb forms namely form V, VIII and others. They can also encompass various semantic structure. In this section, we discuss three types of reflexives. First, we discuss reflexive for action verbs whose undergoer refers to the actor. Next, we discuss reflexive of causatives, where the causer and causee refer to the same entity. Finally, we discuss the reflexive of other semantically complex predicates where the situation semantics consists of multiple elementary predicates.

4.3.1 Action Reflexive

In this section, we discuss how an action verb can take a reflexive form. In the following, we present a few example of action verbs with their reflexive counterpart. All of them come from the form VIII.

(4.8) غَسَلَ (*gasala*) – to wash ... اِغْتَسَلَ (*igtasala*) – to wash oneself

(4.9) سَتَرَ (*satara*) – to cover ... اِسْتَتَرَ (*istatara*) – to cover oneself

(4.10) عَزَلَ (*3aJala*) – to remove ... اِعْتَزَلَ (*i3taJala*) – to remove oneself

(4.11) عَصَمَ (*3aSama*) – to preserve ... اِعْتَصَمَ (*i3taSama*) – to preserve oneself

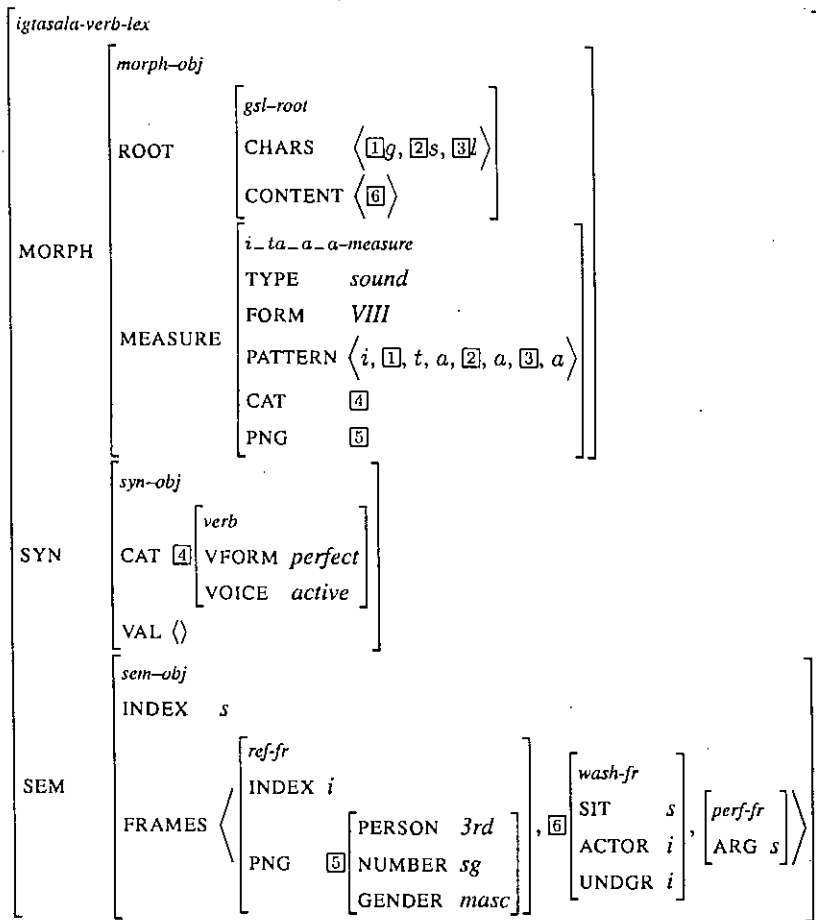


Figure 4.6: HPSG Sign for igtasala

(4.12) فَرَّقَ (*faraqa*) – to split ... اِفْتَرَقَ (*iftaraqa*) – to split oneself

Here we give the attribute value matrix (AVM) for the reflexive form اِغْتَسَلَ (*igtasala*) – He washed himself for the Arabic verb غَسَلَ (*gasala*) – He washed.

We have all the three features associated with morphology in the Figure 4.6, where TYPE takes its value *sound*. ROOT is the list of root letters *g, s, l* as well as the CONTENT gives the semantic contribution with the *wash-fr* in the FRAMES feature. This

indicates that the core meaning which the root letters contribute are somehow associated with the concept of splitting. Next, the feature MEASURE, which contains the morphological, syntactic and semantic information contributed by measure. First, the feature FORM, which denotes the semantic paradigm of *igtasala* as a FORM-VIII derivative. Next, the feature PATTERN captures the stem measure *i_ta_a_a*. Then, the feature CAT, which contains the syntactic category for this measure, structure-shared with the syntactic feature CAT. Finally, the feature PNG, which captures the PERSON, NUMBER and GENDER information of our semantic actor whose value is *3rd/Singular/Masculine*.

The SYN feature captures syntactic information. First, the CAT features identifies the syntactic category of *-igtasala*. Its VFORM and VOICE features which govern the derivational paradigm of verb lexeme, take the values *perfect* and *active* respectively. Next, the VAL feature, which captures the subcategorization of verbs. As a transitive verb it requires an object. However the meaning of its object is absorbed in its verb sense. The object express the entity which is washed. The reflexive sense place this meaning inside its verb sense. We should also note that the hidden pronoun, *he* is encoded by the inflectional morphology, when no explicit subject is used. The semantic actor is not realized syntactically. So, the verb does not subcategorize for neither syntactic subject nor object and the VAL is empty.

The SEM feature captures semantic information. In the case of *igtasala*, the event of washing is expressed. The event has been *completed* and there is a discourse referent to the *actor*. To capture the core event, *wash-fr* is introduced. To capture the temporal constraint, we use the *perf-fr*. Finally, to express the *actor* of the event, the hidden pronoun, we introduce a discourse referent with corresponding PNG feature. The *perf-fr* takes the situation hook *s* of the *wash-fr* as an argument, which is expressed as the feature ARG. The INDEX value of discourse referent and the ACTOR value of the *wash-fr* are co-indexed, both are given the value *i*. The key change is that we express the reflexive sense by co-indexing the ACTOR value and the UNDR value of the *wash-fr* with the value *i*. This indicates that the semantic ACTOR is our semantic UNDR whereas from our previous discussion we can note that the semantic actor is not syntactically realized.

4.3.2 Reflexive for *cause-predicate*

In this section, we discuss how a causative verb can take a reflexive form. In the following, we present a few example of causative verbs with their reflexive counterpart. All of them come from the form V.

- (4.13) ظَهَّرَ (*Tahhara*) – to make (someone) pure ... تَطَهَّرَ (*taTahhara*) – to make oneself pure
- (4.14) عَلَّمَ (*3allama*) – to make (someone) learn ... تَعَلَّمَ (*ta3allama*) – to make oneself learn
- (4.15) فَرَّقَ (*farraga*) – to make (something) split ... تَفَرَّقَ (*tafarraga*) – to make oneself split
- (4.16) فَضَّلَ (*faDDala*) – to make (someone) superior ... تَفَضَّلَ (*tafaDDala*) – to make oneself superior

Here we give the attribute value matrix (AVM) for the reflexive form تَطَهَّرَ (*taTahhara*) – *He make himself pure* for the Arabic verb ظَهَّرَ (*Tahhara*) – *He make (someone) pure*.

We have all the three features associated with morphology in the Figure 4.7. First, the feature TYPE, which denotes the associated root class, takes its value *sound*. Next, the feature ROOT, which is the list of root letters *T, h, r* as well as the CONTENT feature giving the semantic contribution made by root letters. Here, its value is structure-shared with the *to-be-pure-fr* in the FRAMES feature. Here, the point to note is that our predicate exhibits a patient oriented meaning. This indicates that the core meaning which the root letters contribute are somehow associated with the concept of being pure. Next, the feature MEASURE, which contains the morphological, syntactic and semantic information contributed by measure. First, the feature FORM, which denotes the semantic paradigm of *taTahhara* as a FORM-V derivative. Next, the feature PATTERN captures the stem measure *ta_a_a_a*. Then, the feature CAT, which contains the syntactic category for this measure, structure-shared with the syntactic feature CAT. Finally, the feature PNG, which captures the PERSON, NUMBER and GENDER information of our semantic actor whose value is *3rd/Singular/Masculine*.

The SYN feature captures syntactic information. First, the CAT features identifies the syntactic category of *-taTahhara*. Its VFORM and VOICE features which govern

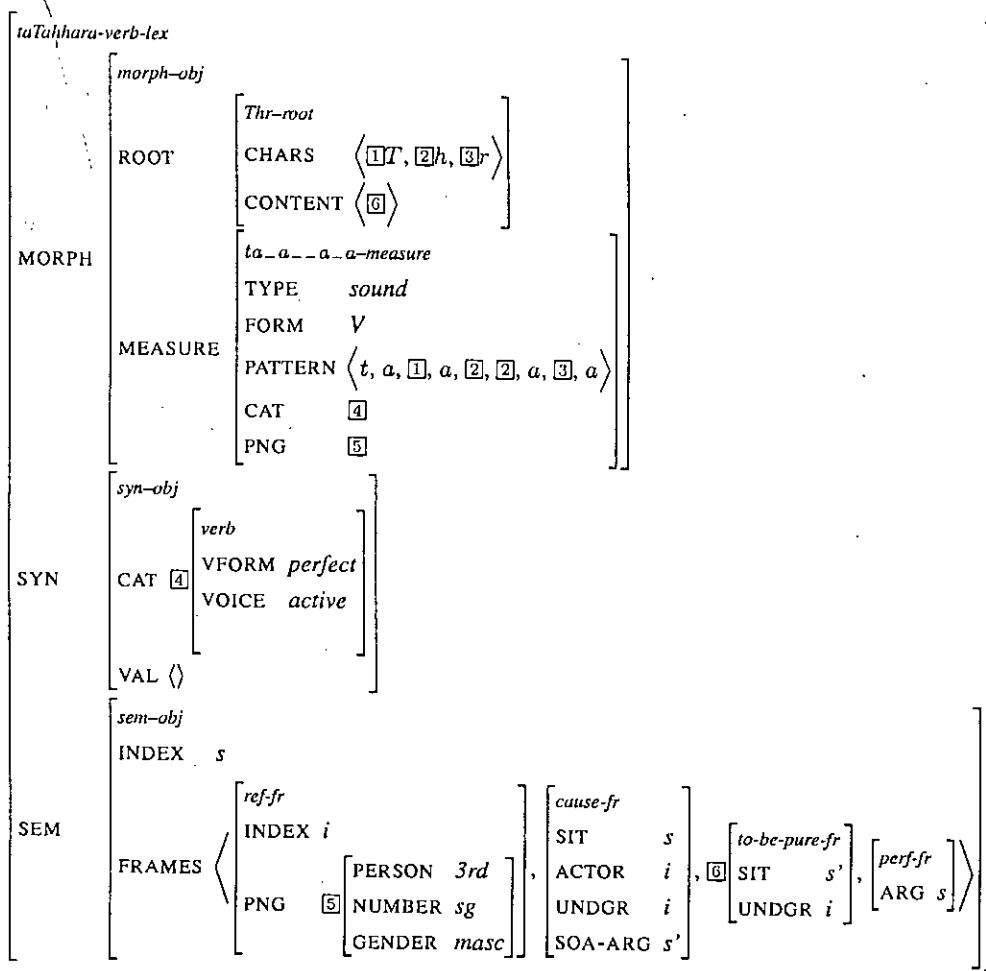


Figure 4.7: HPSG Sign for *taTahhara*

the derivational paradigm of verb lexeme, take the values *perfect* and *active* respectively. Next, the VAL feature, which captures the subcategorization of verbs. Generally, causative verbs subcategorizes for at least one argument, *the causee*. However, in reflexive sense the *causee* is implicit in the verb sense which indicates same reference to the *causer* and *causee*. There may be another optional complement which is the result of the cause. The predicate *to-be-pure-fr* does not introduce any required object that

needs to be syntactically realized. Here, the original intransitive verb subcategorizes for no object and in the reflexive of its causative version it remains intransitive. In this version we capture this by specifying an empty VAL list. We should also note that the hidden pronoun, *he* is encoded by the inflectional morphology, when no explicit subject is used. The semantic causer is not realized syntactically.

The SEM feature captures semantic information. In the case of *taTahhara*, the event of being pure as well as the event of causation is expressed. The events have been *completed* and there is a discourse referent to the *causer*. In this case the base predicate is *cause-fr* and the complement or core event is *to-be-pure-fr*. To capture the temporal constraint, we use the *perf-fr*. This *cause-fr* is a semantic increase of the FRAMES feature. The discourse referent in the *ref-fr* of the feature FRAMES is co-indexed with the value of ACTOR feature of the *cause-fr*, both are given the value *i*. This indicates that the *causer* is the actor of this situation, expressed by the value *i*. The key change is that we express the reflexive sense by co-indexing the ACTOR and the UNDGR of the *cause-fr* with the value *i*. This indicates that the semantic ACTOR is our semantic UNDGR. The UNDGR of the *cause-fr* is also co-indexed with the UNDGR of the *to-be-pure-fr* which is the complement predicate using the value *i*. SOA-ARG feature in the *cause-fr* and the event hook SIT of *to-be-pure-fr* are co-indexed using the value *s'*. This indicates that the event of being pure is the consequence or result of the event of causing. The *perf-fr* takes the situation hook *s* of the *cause-fr* as an argument, which is expressed as the feature ARG.

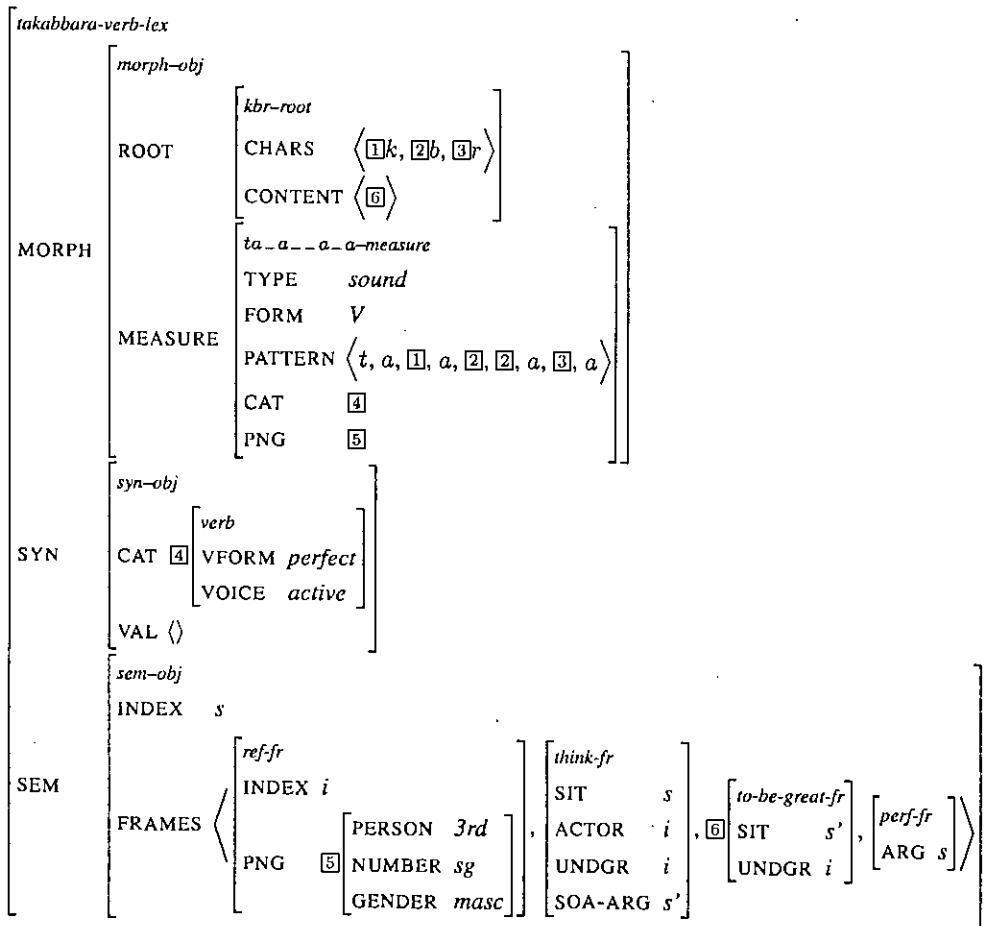


Figure 4.8: HPSG Sign for *takabbara*

4.3.3 Reflexive for *think*-predicate

In this section, we discuss the reflexive of semantically complex predicates where the situation semantics consists of multiple elementary predicates. We take an example of *تَكَبَّرَ* (*takabbara*) – *to think oneself great*. Here, the form I verb is *كَبَّرَ* (*kabbara*) – *to be great*. Its form II verb is *كَبَّرَ* *kabbara* – *to think (someone) great*. Here, the *think*-predicate is base predicate and *to-be-great*-predicate is complement predicate.

We have all the three features associated with morphology in the Figure 4.8. First, the feature TYPE, which denotes the associated root class, takes its value *sound*. Next, the feature ROOT, which is the list of root letters *k, b, r* as well as the CONTENT feature giving the semantic contribution made by root letters. Here, its value is structure-shared with the *to-be-great-fr* in the FRAMES feature. Here, the point to note is that our predicate exhibits a patient oriented meaning. This indicates that the core meaning which the root letters contribute are somehow associated with the concept of being great. Next, the feature MEASURE, which contains the morphological, syntactic and semantic information contributed by measure. First, the feature FORM, which denotes the semantic paradigm of *takabbara* as a FORM-V derivative. Next, the feature PATTERN captures the stem measure *ta_a__a_a*. Then, the feature CAT, which contains the syntactic category for this measure, structure-shared with the syntactic feature CAT. Finally, the feature PNG, which captures the PERSON, NUMBER and GENDER information of our semantic actor whose value is *3rd/Singular/Masculine*.

The SYN feature captures syntactic information. First, the CAT features identifies the syntactic category of - *takabbara*. Its VFORM and VOICE features which govern the derivational paradigm of verb lexeme, take the values *perfect* and *active* respectively. Next, the VAL feature, which captures the subcategorization of verbs. Generally, semantically complex verbs with *think*-predicate subcategorize for at least one argument, *the thinker*. However, in reflexive sense the *thinker* is implicit in the verb sense which indicates same reference to the *thinker* and *thinker*. There may be another optional complement which is the topic of thinking. The predicate *to-be-great-fr* does not introduce any required object that needs to be syntactically realized. Here, the original intransitive verb subcategorizes for no object and in the reflexive of its causative version, it remains intransitive. In this version we capture this by specifying an empty VAL list. We should also note that the hidden pronoun, *he* is encoded by the inflectional morphology, when no explicit subject is used. The semantic causer is not realized syntactically.

The SEM feature captures semantic information. In the case of *takabbara*, the event of being great as well as the event of thinking is expressed. The events have been *completed* and there is a discourse referent to the *thinker*. In this case the base predicate is *think-fr* and the complement or core event is *to-be-great-fr*. To capture the temporal constraint, we use the *perf-fr*. This *think-fr* is a semantic increase of the FRAMES feature. The discourse referent in the *ref-fr* of the feature FRAMES is co-indexed with the value of ACTOR feature of the *think-fr*, both are given the value *i*. This indicates

105987

that the *thinker* is the actor of this situation, expressed by the value *i*. The key change is that we express the reflexive sense by co-indexing the ACTOR and the UNDGR of the *think-fr* with the value *i*. This indicates that the semantic ACTOR is our semantic UNDGR. The UNDGR of the *think-fr* is also co-indexed with the UNDGR of the *to-be-great-fr* which is the complement predicate using the value *i*. SOA-ARG feature in the *think-fr* and the event hook SIT of *to-be-great-fr* are co-indexed using the value *s'*. This indicates that the event of being great is the topic of the event of thinking. The *perf-fr* takes the situation hook *s* of the *think-fr* as an argument, which is expressed as the feature ARG.

4.4 Summary

In this chapter, we have given an HPSG analysis for various derivational forms of Arabic verb including *passive*, *causative* and *reflexive*. They exhibit a myriad of agency features. They are also extendible to other morphologically complex predicates.

Chapter 5

Conclusion

In this thesis, we give a pioneering proposal about how to capture nonconcatenative templatic morphology, especially Arabic verb morphology within the framework of HPSG. Our detailed contributions are given in the following list.

- We have identified the linguistically motivated *types* for Arabic as well as customized the cross-linguistic features according to the requirement. They are arranged under an HPSG type hierarchy.
- Using the identified features, we have given a *generic attribute value matrix* for an Arabic verb. This matrix gives the representation of a typical Arabic verb. The contrast is explained with the AVM of a typical English verb presented in [27].
- We have constructed the MORPH feature so that it can capture the traits of nonconcatenative templatic morphology. MORPH is capable to model the regular concatenative morphology as well as nonconcatenative templatic morphology found in Semitic languages. We have introduced the features ROOT and MEASURE, which denote the root letter contributions as well as templatic effects respectively.
- We have modified several syntactic and semantic features such as VFORM, VOICE, MOOD, VAL etc. We have explained the variations among the verbs in *perfect*, *imperfect* and *imperative* forms with concrete examples. *Imperfect* and *imperative* form introduced new feature MOOD and in semantics, we distinguished their effects using modifier frame.

- We have covered three agentive aspects of Arabic morphology. Arabic can express *passive*, *causative* and *reflexive* constructions using templatic morphology.
- Passive construction deviates from their active counterpart in several ways. Among them, most important is that the discourse referent now *co-refer* the event undergoer instead of actor. Moreover, the event actor is completely unknown by not having a syntactic argument. This is a special characteristic of Arabic passive.
- Causatives are expressed using the concept of *MCP* (morphologically complex predicate). Two predicate have been used to capture the event semantics where the *cause-fr* is used as the base predicate. Undergoer of *cause-fr* co-refer the corresponding role in complement predicate. This may also increase the requirement of syntactic argument.
- We capture reflexives by the *co-reference* of actor and undergoer. We have also covered the reflexive construction of morphologically complex predicates.

There is a massive amount of works to do in the future. To construct matrices from the Table 2.2, which shows the derivational paradigm of Arabic roots, we need to cope with a wide range of diversity that an Arabic verb can take.

- In this thesis, we have dealt with the verbs those generate from sound root class. There are many different kinds of root class, which need to be addressed including weak and four-letter roots. Ongoing works are in place to classify the root classes and their corresponding attribute value matrices.
- A complete treatment of Arabic syntactic valancy is necessary to construct the appropriate type hierarchy. Arabic shows the unusual phenomenon of dependent valancy. Future work should give its syntactic analysis in detail.
- Arabic exhibits various type of MCP. We have treated causative and estimative in some detail. Semantic composition, regularities of predications among various roots, distribution of complex predicates among multiple forms, lexical rule treatment of semantic templates, are some of the issues need to be deeply analyzed.
- We need to develop a detailed treatment of lexical rules to cope up with the numerous derived and inflected forms that can be generated from a single tri-lateral root. Avoidance of unnecessary generation is a big issue in this respect.

- We need to extend the formalism to capture the root-derived Arabic nominal system for a complete development of Arabic computational lexicon. Arabic nouns also exhibit moderately regular morphological system. However, presence of anomalies are higher in nominal system requiring special treatment.
- A completely different direction, which is related to our works is the development of Arabic phrasal constructs. Which is well worth due to their uniqueness inherited from the Semitic origin as well as for the development of a full-fledged computational resource grammar for Arabic.

However, these are not the only directions. New direction can be spawned from other dimensions also. Results will be immensely helpful for the construction of resource grammar for other languages with rich nonconcatenative morphology.

Appendix

In the Table 5.1, we give the romanized transliteration of Arabic alphabet.

Table 5.1: Transliteration Table of Arabic Alphabet

Arabic Letter	Transliteration	Arabic Letter	Transliteration
ا	a	ظ	Z
ب	b	ع	3
ت	t	غ	g
ث	C	ف	f
ج	j	ق	q
ح	H	ك	k
خ	kh	ل	l
د	d	م	m
ذ	z	ن	n
ر	r	و	w
ز	J	أ	h
س	s	ه	y
ش	sh	ي	D
ص	S	ط	T
ض	a	أ	a
ط	i	أ	u
ق	u	ا	i

Bibliography

- [1] Copestake, A. and Flickinger, D., *An open-source grammar development environment and broad-coverage English grammar using HPSG*, Proceedings of the 2nd conference on Language Resources and Evaluation, 2000.
- [2] Tseng, J., *La Grenouille: Grammar Report*, Delph-In Summit, 2007.
- [3] Marimon, M., Bel, N., Espeja, S. and Seghezzi, N., *The Spanish resource grammar: pre-processing strategy and lexical acquisition*, Proceedings of the Workshop on Deep Linguistic Processing, Association for Computational Linguistics, 2007.
- [4] Comrie, B., Fabri, R., Hume, B., Mifsud, M., Stolz, T. and Vanhove, M., (Eds), *Towards an HPSG analysis of Maltese*, Proceedings of the 1st International Conference on Maltese Linguistics, 2007.
- [5] Siegel, M. and Bender, E. M., *Efficient deep processing of Japanese*, Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization, Coling, 2002.
- [6] Hellan, L. and Haugereid, P., *NorSource - an exercise in the Matrix Grammar building design*, Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Engineering, ESSLLI 2003.
- [7] Bender E. M., Flickinger, D. and Oepen, S., *The grammar Matrix. An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammar*, Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics, 2002.

- [8] <http://www.delph-in.net>, last visited June 4, 2008.
- [9] Bond, F., Oepen, S., Siegel, M., Copestake, A. and Flickinger, D., *Open source machine translation with DELPH-IN*, Proceedings of the Open-Source Machine Translation Workshop at the 10th Machine Translation Summit, 2005.
- [10] Wahlster, W., (Ed), *Verbmobil. Foundations of Speech-to-speech Translation*, Berlin, Germany: Springer, 2000.
- [11] Sells, P., *Lectures on Contemporary Syntactic Theories*, Stanford: CSLI Publications, 1985.
- [12] Sag, I. A., Wasow, T. and Bender, E. M., *Syntactic Theory: A Formal Introduction*, Stanford: CSLI Publications, 2003.
- [13] Riehemann, S. Z., *A Constructional Approach to Idioms and Word Formation*, PhD Dissertation, Stanford University, 2001.
- [14] Riehemann, S. Z., *Type-Based Derivational Morphology*, *Journal of Comparative Germanic Linguistics*, Vol-2, 1998.
- [15] Bird, S. and Klein, E., *Phonological Analysis in Typed Feature Systems*, *Computational Linguistics*, Vol-20, 1994.
- [16] Beesley, K. R., *Finite-state morphological analysis and generation of Arabic at Xerox research: status and plans in 2001*, Proceedings of the Workshop on Arabic Language Processing: Status and Prospects, Association for Computational Linguistics, 2001.
- [17] Buckwalter, T., *Buckwalter Arabic Morphological Analyzer Version 2.0*, LDC catalog LDC2004L02, 2004.
- [18] Smrž, O., *Functional Arabic Morphology. Formal System and Implementation*, PhD thesis, Charles University in Prague, 2007.
- [19] Song, S. and Choe, J., *Type Hierarchies for passive forms in Korean*, HPSG Conference, pp. 250-270, 2007.
- [20] Müller, S., *The passive as a lexical rule*, HPSG Conference, pp. 247-266, 2000.

- [21] Tseng, J., *English prepositional passive constructions*, HPSG Conference, pp. 271-286, 2007.
- [22] Norcliffe, E., *Constructing Spanish complex predicates*, HPSG Conference, pp. 194-213, 2007.
- [23] Tily, H. J. and Sag, I. A., *A unified analysis of French causatives*, HPSG Conference, pp. 339-359, 2007.
- [24] Manning, C. D. and Sag I. A., *Argument Structure, Valence, and Binding*, Nordic Journal of Linguistics 21, pp 107144, 1998.
- [25] Davis, A., *Linking and the Hierarchical Lexicon*, Ph. D. thesis, Stanford University, 1996.
- [26] Davis, A., *Linking by Types in the Hierarchical Lexicon*, Stanford:CSLI Publications, 2001.
- [27] Sag, I. A., *Sign-Based Construction Grammar: An Informal Synopsis*, unpublished manuscript, 2007.
- [28] Marquis, R. C., *Phases and Binding of Reflexives and Pronouns in English*, HPSG Conference, pp. 482-502, 2005.
- [29] Pollard, C. and Xue, P., *Syntactic and Nonsyntactic Long-distance Reflexives*, Syntax and Semantics 33, pp. 317-342, 2001.
- [30] Artiagoitia, X., *Reciprocal and Reflexive Constructions*, A Grammar of Basque, Berlin: Mouton de Gruyter, pp. 607-632, 2003.
- [31] <http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/>
- [32] Islam, M. S., and Ahmed, R., *An HPSG Analysis of Arabic Verb*, Accepted in the 9th International Arab Conference on Information Technology, 2008.
- [33] Islam, M. S., and Ahmed, R., *Nonconcatenative Morphology: An HPSG Analysis*, Accepted in the 5th International Conference on Electrical and Computer Engineering, 2008.

- [34] Islam, M. S., and Ahmed, R., *An HPSG Analysis of Arabic Passive*, Accepted in the 11th International Conference on Computer and Information Technology, 2008.
- [35] Cipollone, D., *Morphologically Complex Predicates in Japanese and What They Tell Us About Grammar Architecture*, OSU Working Papers in Linguistics 56, 2001.
- [36] Attia, M., *Developing a robust Arabic morphological transducer using finite state technology*, 8th Annual CLUK Research Colloquium, 2005.

