Cognitive Aspects of a Statistical Language Model for **A**rabic based on **A**ssociative **P**robabilistic **Ro**ot-**PAT**tern Relations: A-APRoPAT

Bassam Haddad

***Abstract*—Motivated by the nature of Arabic and encouraged by the delivered indications by psycho-cognitive research in retrieval of linguistic constituents of a word, a statistical language model for Arabic based on Associative Probabilistic bi-directional Root-PATtern relations", (A-APRoPAT) was formalized. The basic components of this model are relying on bi-directional probabilistic root-pattern relationships acting as cognitive morphological factors for word recognition in addition to semantic classes capturing textual and contextual root associative network. Considering a root in the mental representation as the highest level of symbolic semantic abstraction for a morphological unit allows the perception of words as a probabilistic applicative process instantiating the most plausible or known pattern to the most conceivable root, in an Associative Root-Pattern Network. As Arabic is known for its highly inflectional morphological structure and its high tendency to pattern and root ambiguity (Pattern Polysemy and Root-Homonymy) this model is assuming bi-directional morphological background knowledge for resolving ambiguities in the form of a probabilistic semantic network. As a major consequence, this paper is stressing the significance of this phenomenon in designing Artificial Cognitive Systems and Cognitive Infocommunications Applications concerned with Arabic interactive systems particularly those related to Arabic natural language understanding and human visual word identification and corrections besides the overall domination of Arabic morphology as a non-linear or non-concatenated processing system in the case of word identification.**

***Index Terms*—Cognitive Informatics, Cognitive Linguistics, Cognitive Infocommunications, Statistical Language Model, Arabic Morphology, Visual Word Identification, Probabilistic Root-Pattern Associative Relation, Root-Homonymy, Pattern Polysemy, Pattern Syntax, Pattern Semantic.**

I. Introduction and Motivation

THIS paper is dealing with some aspects concerned with *cognitive informatics*, *cognitive linguistics* and the closed related *Cognitive Infocommunications (CogInfoCom)* [2] in the context of the *visual word identification problem* in Arabic and, generally in Semitic languages. Within a statistical language model, the domination of the root-pattern aspect in word identification is *reviewed* and *reinvestigated*. As a key result of this research, this paper is proposing considering the overall cognitive domination of the *root-pattern phenomenon* and its *probabilistic associative network* in designing Arabic *artificial cognitive systems* and *Cognitive Infocommunications Applications* such as Human Computer Interactive based natural language understanding and in particular *visual interactive* word recognition, correction, Arabic based Machine Translation and e-learning systems.

In fact, since entering the age of Arabic computational linguistics, discussions about lexical semantic representation were controversial[1]. Classical Arabic linguisticians recognized a long time ago the semantic dimension of the *root-pattern phenomenon* and strongly employed its significance in their theoretical lexical and syntactical analysis. However, after the appearance and implementation of the first computational models, and due to the overall unforeseen computational complexity, and possibly the dominating and relatively advanced Indo-European languages Natural Language Processing Technology, some issues were raised suggesting stem-based or lexeme-based units as an alternative [8], [19] for building simpler morphological model for lexical analysis. According to the laws of ambiguities in [8], ambiguity increases proportionally with lifting the abstraction of lexical units. There are also some reports coming from Information Retrieval discussing the concept of stem-based indexing versus root-based indexing in the context of reducing computational complexity and increasing precision [14], [15].

On the other hand, the current cognitive linguistic research concerned with word identification for Semitic languages and in particular for Arabic and Hebrew is delivering increasingly evidence of the sensitivity of readers to root and pattern during visual word processing. These cognitive models proceed from the assumption, that word recognition can be viewed as a dynamic process operating on the mental lexicon considering basic phonological, semantic and orthographic characteristics processing morphological contents with respect to mental organization [1], [4] [5], [16] and [20].

Since starting our research in Arabic Natural Language Processing (Arabic NLP) aiming at building a comprehensive Natural Language Understanding System, we adopted the *root-pattern paradigm*, motivated by our understanding of the nature of Arabic, and encouraged by the *indications* delivered by psycho-cognitive research in retrieval of linguistic constituents of a word. Relying on these elements, "A-APRoPAT"; (**A**rabic statistical language model based on **A**ssociative relations considering **P**robabilistic Bi-directional

Manuscript received 21 May 2013, revised 03 June 2013.

Bassam Haddad, Department of Computer Science, University of Petra, B.O.BOX 3034, 11181-Amman, Jordan (e-mail: haddad@uop.edu.jo; haddad.bh@gmail.com).

[1]Unfortunately, there is till now no clear research track in Arabic Natural Language Proceeding (Arabic NLP) devoted to Arabic or Multi Lingual Artificial Cognitive Systems. In this context, we believe the effect of the root-pattern paradigm and the word order in the 3-Argument structure; i.e. Verb-Subject-Object or nominal sentences besides the particle semantics need further research in connection to cognitive Informatics.

Root-**PAT**tern relationships) has been developed acting as a *theoretical background* and *conceptual model* for designing artificial cognitive systems concerned with Arabic language. The applications potential of this model are wide-ranging, including supporting Morphological Analysis; Word Sense Disambiguation, Spell-Checking, POS-Tagging and supporting indexing and ranking in root-pattern indexing and semantic based Search Engines, particularly when integrating this view within the *inter-cognitive communicative facility* of such systems.

A. Scope of the Presentation

However, this paper will primarily focus the attention on some cognitive aspects of this model related to root-pattern morphological and organizational structures, emerging from the view that *roots* in Arabic represent *autonomous semantic lexical units*. And on the other hand, patterns represent a phonetic and conceptual unit that can simply be perceived through an instantiation of a pattern with a root, considering the most plausible and associative relationship between a pattern and a root. In addition, within these aspects, the *visual word identification problem* and *non-word correction* will be treated as a representative of the application potential of A-APRoPAT in the context of artificial cognitive system modeling. Details related to application potential, computational model and chains lie beyond the scope of this presentation.

In the following sections, in connection to the word identification problem, some cognitive aspects of the root-pattern paradigm will be introduced, whereas the word recognition process will be reviewed within the main features of a statistical language model. Based on [9], [11] the concept of probabilistic bi-directional root-pattern relations will be further elaborated and investigated and the concept of *root's textual* and *contextual associative relationships* will be introduced.

Empirical analysis toward the model's cognitive justification and applications potential will be presented in the form of an implementation of one aspect of this model, designed for non-word recognition and correction. This model will be conceptually compared in the context of *Non-Word identification problem* to Microsoft® Word Spell-Checker. Finally an overview and discussion about aspects and prospects will be presented.

II. Cognitive Aspects of the Root-Pattern Paradigm

Before elaborating on the fundamental factors involved in the proposed cognitive model for word recognition, an overview of the characteristic of Arabic morphology and its particularity will be introduced to rationalize and simplify the utilized terminology.

A. Particularity of Arabic Morphology

Arabic belongs to the Semitic Languages family, which is widely used in the Arabic & Islamic Worlds; *(Approximately 422 million total speakers)* [21]. Based on [9], [13] Arabic and other Semitic languages belong to a singular *non-concatenative* or *non-linear morphological class* where *root letters* are decisive for forming the majority of words. Roots accordingly represent the highest level of symbolic abstraction for a word[2]. In this context, roots correspond to *ground-morphemes* as the smallest meaningful elements without considering the affixes as is the case in the Indo-European languages. Morphologically, words can be classified into three major lexical classes:

- *Basic Derivative*.
- *Rigid (Non-Derivative)*.
- *Arabized Words*.

Basic Derivative Arabic Words form the overwhelming majority of the Arabic lexical vocabulary. Most of these words can be generated from a *templatic triradical* or *quadriradical* root by adding *consistent prefixes* and suffixes or filling vowels in a predetermined *phonological* pattern form. The patterns have two syntactical categories: *verbal and nominal* word patterns in addition to some semantic content. Non-Derivative words include the lexical non-inflectional word types such as *pronouns*, *adverbs*, *particles* and *stem words*, which cannot be reduced into a known *root* or to *ground-morphemes*. *Arabised Basic Words* consist of words without Arabic origin such as $(/'intarnit/, Internet.)$ Arabised Basic Words, and Non-Derivative Words, do not linguistically evidence a *perceivable root-pattern relationship* [12].

B. Morphological Root-Pattern Factors

Based on [9], [12], whereas a root can be considered as a basic morphological unit carrying the core semantic and meaning of a possible word, a pattern instantiated as a word creates variation on the meaning and on the syntactical category. The exact meaning of a word can not be predicated unambiguously without considering its morphological units; i.e. *the root and its consistent patterns*. Furthermore, the fact that a root is *interwoven in a phonetic pattern* and not necessarily contiguously situated in a word, allows investigating morphological factors *non-linearly*. This view allows the possibility of organizing morphological units in the form of an *associative or semantic relationship* between constituents of Arabic morphology. Moreover, this model is assuming some kind of *uncertainty* among such root-pattern associative relationship.

The concept of a pattern represents a unique syntactical and semantical constituent. It forms *schemes* or patterns or *morphemic types* organizing *consonant* and *vowel* substitution. The root is unalterable giving the basic meaning, while the pattern can be inflected by *infixes*, *prefixes* and *suffixes*, denoting grammatical change forming new words with related meanings.

Most Arabic words can be generated from the templatic *triliteral ground morpheme* $(/f'l/, C_1C_2C_3)$ *(for transliteration, see Appendix A)* or the *quadrilateral ground morpheme* by adding consistent prefixes and suffixes or filling vowels

[2]There are approximately around 7420 Roots, whereas approx. 25-30% are triliteral and 70-75% quadrilateral [3], [17].

in a predetermined pattern form[3]. For each valid Arabic root, there is a certain number of consistent patterns. The degree of consistency, expresses the presence of an associative relationship between a root and a pattern *priming* some word forms. Therefore, a lexical derivative Arabic word can be understood as a result of *applying a substitution* [9] of a root literal to the corresponding consistent pattern literals; i.e. a generative or derivative process. Such a substitution can *cognitively* be regarded as a transformational operation of a root into a pattern word and an instantiation a template with a root letters[4], in the sense of finding the most plausible; i.e. associative or consistent pattern form for a given root and vice versa. By further analysis, this process can be understood as a *bi-directional* and a non-trivial s*earch problem* under *uncertainty* to decide the degree of association or consistency between given roots and the possible patterns, where roots represent the highest conceptual level in the mental representation, and phonetic patterns represent abstract templates for possible instantiations of objects and activities in the real world. To visualize such process of word identification (*see Figure 1*).

Due to historical reasons and difficulties of representing short vowels the Arabic script, the overwhelming written Arabic texts are not vocalized. Considering additionally the fact that a root occurs with many different possibly un-vocalized patterns in most available Arabic texts complicates the process of word identification for the Arabic reader[5]. Such ambiguity is two fold in the sense that for one root there will be many possible un-vocalized patterns, which represents a kind of *pattern polysemy* and for one pattern there might be more than one possible root where each root-pattern relationship might represent different word senses; i.e. some kind of *root homonymy*.

A pattern specifies additionally *semantic* and lexical information about the resulting word though instantiating a root by a pattern. Figure 3 represents a sample of the adopted pattern feature structure capturing lexical syntactic and semantic information on templatic level used in A-APRoPAT model.

In the following, an *inter-cognitive word identification process* will be considered as *a probabilistic applicative process* in the sense of functional programming or uncertain pattern *matching process* problem. This aspect is interesting as this model is operating on abstract relationships between roots and patterns, and the meaning of a word cannot be exactly determined without some kind of substitution of the *root radical* within an adequate or consistent phonetic pattern. Additionally, the uncertainty aspect will be also considered by introducing bi-directional associative relations in form of probabilistic rules expressing an associative relationship on the root-pattern level.

[3]For example, under the hypothetical assumption that $geC_1C_2oC_3en$ were a German pattern, and (FLG) and (FLH), were German roots, then the following words can be instantiated by applying a root to the giving patterns: (/geFLoGen/ flown) and (/geFLoHen/ escaped).

[4]Formal definition is given in [9].

[5]In the context of Arabic script problematic, $An\bar{\imath}s\ Furayh\underline{a}$ wort in 1955 "We are the only nation that needs understanding to read, while all other nations of the earth read to understand" [17].

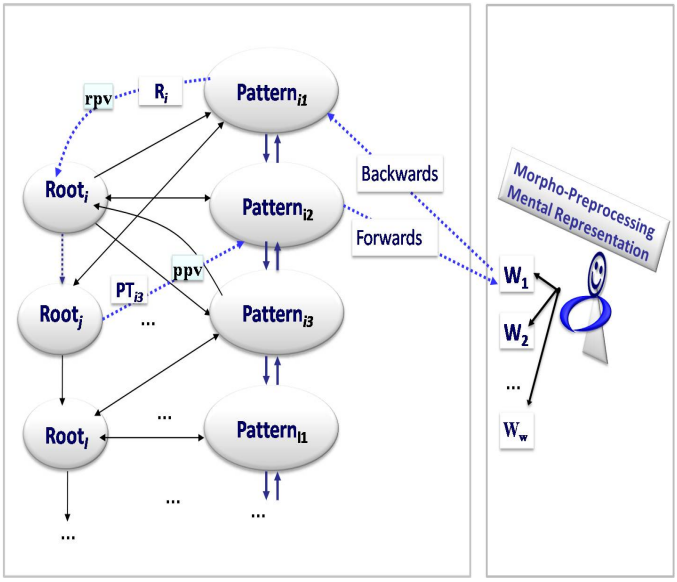

Figure 1. *Visual Word Identification as bi-directional Uncertain Search Problem (from right to left to stress the right to left reading in Arabic).The figure shows that the root R_i was initially sleeted, however the $root_j$ with its pattern PT_{i2} and value were returned in a forwards and backwards dynamic process.*

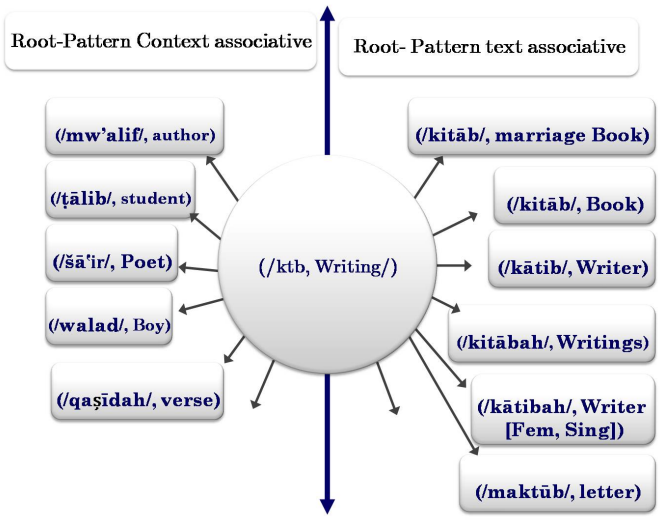

Figure 2. *An example of textual and contextual root-pattern associative relations as priming for the root $(/ktb/, Writing/)$. Predictive Values are not depicted.*

C. Associative Root-Pattern Relationships

In the cognitive science community, there is a general agreement, that there are indeed *morphological factors* having cognitive impact upon the *word recognition process*. This aspect can be evidently observed in many studies concerned with *priming effects* through morphologically related words, in the sense that morphological manipulation affects word recognition and supports the view, that morphological basic units are explicitly represented in the mental lexicon facilitating the word recognition process [4], [5].

In the process of the *empirical* and *corpus* based analysis

toward the *cognitive* and *applied* justification of the (A-APRoPAT) model, an Arabic root might intuitively be a *priming factor* for many different words associated with the ground semantics of the root according to their *frequency of occurrence* and its *contextual network* on the word level. Furthermore, phonetically interrelated pattern word forms might be a *priming factor* for intuitively similar word forms having the same ground semantics of its phonetic pattern form, besides other phonetically similar words containing similar root literals or having a associative relationships. This remark has also been observed in [1], [5] to reassert the *root-pattern dominance* in *reading* of Arabic and Hebrew.

Basically the APRoPAT model differentiates theoretically between two major types of associative root-pattern based relationships forming the elementary units towards building a global network based on *probabilistic contextual bi-directional root-pattern relations* in addition to *semantic classes (see Figures 1 and 2).*

- **Textual-Based Associative Root-Pattern Relations**. This type is concerned with all in a root literals involved associative words. For Example the root word $(/ktb/, Writing/)$ is intuitively associated with the different words having the similar ground meaning of "writing" interwoven in different word patterns such as$(/kitābun/,\ Book)$, $(/kātib/,\ Writer)$, $(/kitābah/,\ Writings)$, $(/maktūb)$, $letter)$, *(see Figure 2)*. In this context, **root homonymy** and **pattern polysemy** might occur, in form of words with distinct root origins however with similar pattern form. On the other hand, *pattern polysemy* can be regarded as a kind of orthographic polysemy due to the un-vocalization on pattern level; i.e. same root but different meaning.
- **Contextual-Based Associative Root-Pattern Relations**. This type is concerned with all words and word patterns involved in the context of the root in form of its contextual semantic network. For example, with the root $(/ktb/,\ Writing/)$, there are word patterns interrelated with "Writing" in context of acting as agent or object or particles, such as $(/tālib/,\ student)$ or $(/qaṣīdah, verse/)$ and others *(see Figure 2).*
- **Semantic Class and Semantic Class Clusters**[6]. A semantic class consists of a certain group of Textual-Based Associative Root-Pattern Relations besides certain group of Contextual-Based Associative Root-Pattern Relations, which are sufficient for representing an abstract concept. Semantic classes are interrelated through their involved root-pattern relationships describing a higher level of abstraction such as an aspect. Figure 5 illustrating the three levels of abstraction involved in A-APRoPAT model. *See also footnote 7.*

Generally speaking these aspects, are related to the *syntagmatic* and *paradigmatic* aspects of a root-pattern probabilistic associative relationship in the sense of considering root-pattern *conceptual and contextual knowledge* in the form of semantic classes or aspects. However APRoPAT model is considering their dimensions on a *higher level of abstraction* based on tamplatic probabilistic bi-directional associative relationships between roots and their possible patterns and their context. Besides capturing such relationships statistically relying on corpus based analysis considering the nature of Arabic and psycho-cognitive research.

In this context a plain root with its ordered radicals represents the *highest level of a symbolic semantic representation*. Moreover, it is associated with certain *phonetic patterns* or *phonetic templates* giving a concrete or image of the real-world by instantiation of such roots with possible associated patterns. A Semantic Class represents accordingly a cluster of involved certain root-pattern relations coming through textual and contextual assignations.

$$\left[\begin{array}{l} (\mathbf{f}a\bar{a}\text{'}i\mathbf{l}un,\ C_1a\bar{a}C_2iC_3un) \\ \\ CAT \left[\begin{array}{ll} \text{Morph} \left[\begin{array}{ll} \text{pref} & [...] \\ \text{suff} & [...] \end{array}\right] \\ \\ \text{Syn} \left[\begin{array}{ll} \text{Type} & \text{NPt[PP]} \\ \text{Case} & \text{Nominative}| \\ \text{Num} & \text{Sing} \\ \text{Gen} & \text{M} \end{array}\right] \\ \\ \text{Sem} \left[\begin{array}{ll} \text{Sem-Cat} & \text{Root-Actor} \end{array}\right] \end{array}\right] \end{array}\right]$$

Figure 3. *Feature Structure of a Pattern Variable (The predicatice values are not depicted).*

The feature structure in *Figure 3* as adopted in this model is an example giving an overview of the possible morphological, syntactic and semantic content, which might be gathered from the pattern $(fa\bar{a}'ilun,\ C_1a\bar{a}C_2iC_3un)$; i.e. *nominal, singular, male, nominative, present participle, beside acting as an actor for the basic semantic of some root.*

Cognitively, this model is assuming some kind of cognitive pre- and post processing in form of backwards and forwards chain within associative probabilistic network for identification a word, concept or an aspect relying on the 3 levels of the model. A cognitive image of word is in terms of A-APRoPAT is therefore a phonetic template with basic potential meaning in associative *uncertain* and /or *imprecise* network. The concrete meaning of a word can be achieved as a probabilistic applicative process in terms of the functional programming.

III. The Fundamentals of Arabic-APRoPAT

Arabic-APRobat is based on the following primary Relations, as adopted from [9]:

- $\vec{\mathcal{R}}_{\mathcal{PT}}$: An associative Relationship from Roots to Patterns defined as *(see appendix)*:

$$\vec{\mathcal{R}}_{\mathcal{PT}} \triangleq \{((r_i, pt_j), \vec{p}pv_{ij}) | (r_i, pt_j) \in \mathcal{R} \times \mathcal{PT}\} \quad (1)$$

where

- $r_i \in \mathcal{RR}$ the set of all Arabic roots.
- $pt_j \in \mathcal{PT}$ the set of all Arabic Patterns.

[6]Some aspects of this model were adopted by **Addlaal** Arabic Search Engine [14].

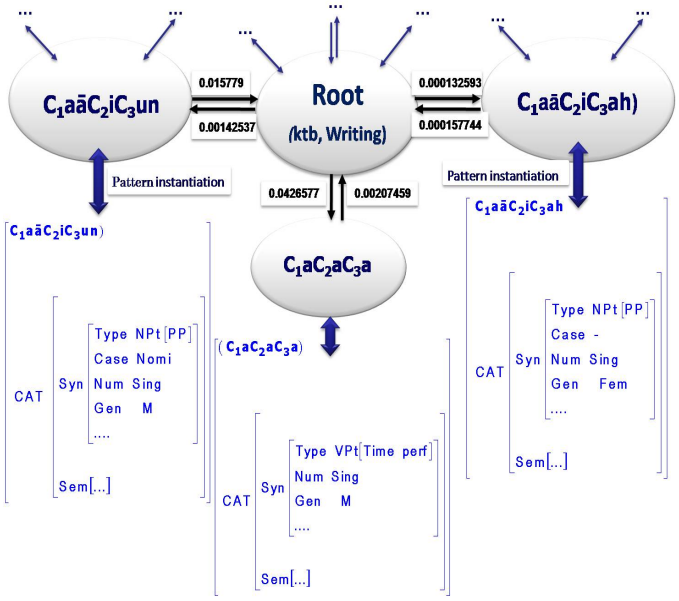

Figure 4. *Example for bi-directional associative root-pattern relations and their possible pattern feature structures. Pattern instantiation is an applicative process [9].*

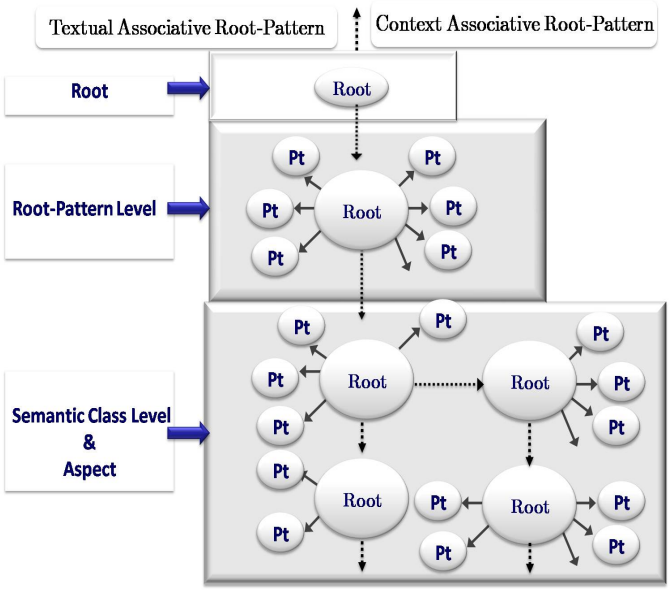

Figure 5. *Root, Root-Pattern and Semantic Class levels as text and context based root-pattern associative relations. Predictive values are not depicted.*

- $\vec{p}pv_{ij} \triangleq P(pt_j|r_i)$ is defined as *Pattern Predictive Value* measuring the degree of probability predicting the associated pattern pt_j when considering the $root_i$.

- $\overleftarrow{\mathcal{PT}}_{\mathcal{R}}$: An associative Relationship from Patterns to Roots defined as:

$$\overleftarrow{\mathcal{PT}}_{\mathcal{R}} \triangleq \{((pt_j, r_i), \overleftarrow{r}pv_{ji}) | (pt_j, r_i) \in \mathcal{PT} \times \mathcal{R}\} \quad (2)$$

where
$\overleftarrow{r}pv_{ji} \triangleq P(r_i|pt_j)$ is defined as *Root Predictive Value* measuring the degree of probability to predict the associated root r_i when considering the pattern pt_j.

- $\overleftrightarrow{\mathcal{RP}}$: Associative Bi-directional Root-Pattern Relationship, which can for each (r_i, pt_j) be established as:

$$r_i \xleftarrow[\vec{p}pv_{ij}]{\overleftarrow{r}pv_{ji}} pt_j \quad (3)$$

$\vec{\mathcal{R}}_{\mathcal{PT}}$ can be interpreted as *uncertain binary forward set of rules* for identification a pattern or a basic derivative word based on known root,

$$r_i \xrightarrow[\vec{p}pv_{ij}]{} pt_j \quad (4)$$

while $\overleftarrow{\mathcal{PT}}_{\mathcal{R}}$ can be interpreted as *uncertain backwards rules* for extracting a root based on known pattern:

$$r_i \xleftarrow[\overleftarrow{r}pv_{ij}]{} pt_j \quad (5)$$

The primary reason for proposing the probabilistic approach, is motivated by the observation, that humans can *promptly identify* a *frequent pattern* or *root*. Furthermore, this model is proposing to measure the degrees of such uncertainty in terms of the values introduced by [9]; i.e. $\vec{p}pv_{ij}$ and $\overleftarrow{r}pv_{ij}$ pattern and root predictive values[7]. To estimate these values statistically, a corpus based approach has been adopted.

A. Acquiring Frequencies of Occurrence

To obtain numerical values for representing the uncertainty between roots and patterns, based on the morphological analysis of a corpus containing 50,544,830 Arabic word-forms in 990 MB text files obtained from the internet and Arabic dictionaries of about 31.5 GB, normalized conditional probabilities were assigned to *6860 Arabic roots* in association with *650 patterns*.

The corpus has mostly been tagged and vocalized to consider some morpho-syntactical features on the pattern level. The data was analyzed by ATW[8] morphological analyzer of Arabic Textware. Vocalization of the most words particulary word endings was essential for computing the frequency of occurrence of patterns in different cases with their instantiated roots. As mentioned earlier, a pattern implies templatic semantic and syntactic information. This aspect is of importance for inferring some information relaying of pattern probabilistic values.

Example

Let $(/ktb/, Writing) \in \mathcal{R}$, $(/maf'\bar{u}l/, maC_1C_2\bar{u}C_3) \in \mathcal{PT}$, then based on the pattern predictive value $\vec{p}pv_{ij} \triangleq P(pt_j|r_i)$, we can establish a binary uncertain relation expressing the probability for predicting the instantiation of the

[7]To visualize the APRoPAT model in terms of medical diagnostic; roots can be considered as signs or symptoms, patterns as intermediate medical entities and a semantic class as a diagnosis. In a probabilistic causal network, a diagnosis occurs by following most plausible weighted paths from different symptoms through their variations; i.e. pattern to the concept. Backwards, means to explain the diagnosis; i.e. the semantic class through the signs backwards; i.e the roots [10].

[8]At present, we are working on lunching a comprehensive project considering root-particle, root-complement associative relationships by adopting the concept of predicative values and refining previously obtained values. We are planning to employ "Petra-Morph" Morphological Analyzer.

pattern $(/maf'\bar{u}l, maC_1C_2\bar{u}C_3)$ with the given root and vice versa; i.e. such:

$$(/ktb/, Writing) \xleftarrow[\vec{p}pv_{ij}]{\overleftarrow{r}_{pv_{ji}}} (maf'\text{`}\bar{u}l/, maC_1C_2\bar{u}C_3) \tag{6}$$

In this context, the identification of basic words can be regarded as probabilistic bidirectional applicative process within a semantic network of associative relationships. Figure 4, shows bi-directional uncertain relationships between the root $(/ktb/, Writing)$ and other possible patterns.The patterns $(C_1a\bar{a}C_2iC_3un)$ and $(C_1a\bar{a}C_2iC_3ah)$ represent examples of two possible cases of root-pattern instantiations. Based on their predictive values some additional syntactic and semantic information can be inferred from the patterns. This example implies also that the word "$(/k\bar{a}tibah/, Writer)$" as feminine is less frequent than the word "$(/k\bar{a}tibun, Writer)$" masculine in the associative network network model.

B. Overview of Application Potential of A-APRoBAT

The significance of the introduced model and computed values i.e. r*oot and pattern predictive values* depends on the application type:

- *Pattern Predictive Relations and Values*, i.e. $\vec{p}pv_{ij} \triangleq P(pt_j|r_i)$ might be interpreted as forward uncertain binary rules in the sense of searching for a pattern compatible with a known root. According to APRoPAT, the root concept depicts the highest level of abstraction and accordingly a patten depicts a forward form of a root. On the other hand *Pattern Predictive Values* support processes involved in generating the most probable word patterns for some possible root, for example for ranking within a correcting process. This aspect can be significant for resolving some ambiguities and in ranking possible candidate corrections.
- *Root Predictive Values*; i.e. $\overleftarrow{r}pv_{ji} \triangleq P(r_i|pt_j)$ might come into effect in the case of generating the most probable roots, within a root-extraction process such as morphological analysis particulary for optimizing the root-extraction process and in particular if the words are strongly deformed or unclear such as postprocessing of OCR systems. Furthermore, it can be regarded as backwards uncertain binary rule in a sense that a root concept depicts a backward process of a pattern.
- *Contextual-Based Associative Root-Pattern Relations*. As this aspect is related to roots contextual network on the root-pattern level, this network can be very useful as additional support for resolving problems involving ambiguity such as word sense disambiguation, Part of Speech Tagging and Information Retrieval.

These aspects might also be extended to different possible applications such as indexing and word sense disambiguation. Giving details about possible algorithms for utilizing these aspects is beyond the scope of this paper. However, based on the main aspect of this model, a hybrid approach for non-word detection and correction has already been implemented and the concepts of the three levels of abstraction has been adopted by Addaall Arabic Search Engine.

IV. Experimental Analysis

The main features of (Arabic-APRoPAT) were introduced in the context of the word identification problem as *an interactive cognitive process*. Based on the fundamental elements of APRoPAT; namely *Associative Bi-directional Root-Pattern Relations*, a program designed to act as *inter-cognitive system* in context of *visual Non-word Recognition and Correction* was implemented. A partly pre-tagged corpus for Arabic was utilized for approximation of root and pattern frequencies. The program was tested and compared with the most popular Arabic Spell-Checker; namely Arabic Microsoft® Word. Humans were asked to decide whether isolated word lists are correct or not, and if not to try to correct them explaining the error type. Most of the persons involved were students at different university levels coming from different Arabic countries. In this experiment, different error patterns were considered such as

- Simple Errors[9].
  Substitution, deletion, insertion and transposition.
- Editing Errors and Boundary Problems.
  Run-on Errors and Splitting Errors.
- Cognitive and Phonetic Mistakes
  E.g. slang or phonetic transcription errors.
- Simple Syntax Errors and Semantic Errors.

The main goal of the experiment was not to evaluate Microsoft®Word's Spell-Checker performance against our system, but to get an overview of the difference between the used lexical analysis compared to APRoPAT concept for word identification and its closeness to act as a computational model simulating an *interactive cognitive process of the word identification*. It was remarkably noticed that Microsoft® Word's Spell-Checker was unable to identify simple cognitive and phonetic type of errors. Even unexpected multiple insertions or some transpositions errors could not be identified (more details are found in [9]). There are indications for the view of adopting a primary linear organizational strategy for the word structures without a *clear ranking strategy* considering *statistical* analysis. We believe that such strategies might not produce natural human expected ranked corrections and results, as it is the case in the implementation strategy employed in APRoPAT model.

However, as a major result of this test, we can conclude that there are strong indications for the presence of the *root-pattern factors* in the mental morphological decomposition supported by test for non-word recognition considering the root-pattern paradigm, which actually *complies* with previous studies for other Semitic languages as explained earlier. The reason for not detecting even simple cognitive error types might be explained by the fact that deep semantic or cognitive patterns of errors can primarily be detected by deep root-pattern reduction analysis utilizing the root morpheme as the basic semantic unit in the context of contextual knowledge base occurring simultaneously in a network of root-pattern relationships.

[9]Details about type of error patterns in Arabic are found in [13].

V. Overview and Discussion

This paper attempted to investigate a widely neglected aspect of Arabic Natural Language Processing; namely the *cognitive aspects of computational linguistics* and *cognitive informatics* in context of *cognitive information and communication*.

Motivated by the nature of Arabic and encouraged by the delivered indications by psycho-cognitive research in retrieval of linguistic constituents of a word, a statistical language model for **A**rabic based on **A**ssociative **P**robabilistic Bi-directional **Ro**ot-**PAT**tern relations", (A-APRoPAT) was formalized. The departure point of this model relies on the root concept as a ground morpheme depicting the highest symbolic level of semantic abstraction with its possible phonetic variations in form of templates or phonetic patterns.

The traditional *two-valued root-pattern paradigm* has been extended to *multi-valued probabilistic root-pattern logic* and in *bi-directional mode*. This model can be viewed as an associative network based on novel probabilistic measures considering uncertainty among bi-directional root-pattern relationships. Furthermore, root-pattern textual and contextual relations have been considered forming semantics classes representing conceptual knowledge about root-pattern associative knowledge necessary for recognition and cognition of words and concepts.

As this model is designed to be comprehensive as possible, the cognitive aspect has been considered for justification of its application potentials. Investigation of all aspects and complexity of this model lies beyond the scope of this paper.

For testing the cognitive elements of this model, a non-word recognition system was implemented considering bi-directional root-pattern relationship to identify words and non-words. The gathered results comply with the current studies of cognitive linguistic research concerned with word identification for Semitic languages and in particular for Arabic and Hebrew, which are increasingly supporting the existence of sensitivity of a human to root and patterns during *visual word processing and communications*. These results are encouraging to review this model within other linguistic problems such as Word Sense Disambiguation; POS-Tagging, supporting indexing and Ranking in Root-Pattern indexing based Search Engines, Machine Learning and others. Aspects related to the semantic class concept and root-pattern indexing will be considered in the intended *APRoPAT Search Engine Project utilizing Petra Morph Morphological Analyzer*.

As preliminary result of this study, the overall domination of such root-pattern phenomena as linguistic image schemata should have its significance and effect in designing artificial cognitive systems such as interactive Arabic natural language understanding systems. This view might be a priming factor for researchers working Arabic NLP to reopen the still controversial issues concerned with *root-pattern* versus *stem-based* approaches. In our opinion, the traditional view of the root-pattern paradigm has been through A-APRoPAT extended offering more promising dimensions for the Arabic NLP. Furthermore, proceeding from the fact that CogInfoCom can be approached from the cognitive linguistics perspective [2], this paper is encouraging researchers of Arabic NLP community to consider their related research within this perspective as a synergic combination of the infocommunications and the cognitive science [18].

On the other hand, we believe that it is worthwhile to consider the principle idea of this model in the context of other Indo-European languages. The motivation behind that is rising from the cognitive value of this model in the context of multilingual computational models, and in particular, when elaborating aspects related to the non-linear morphological contents.

Finally, the wordnet approach [6] is similar on the conceptual level to our approach, however APRoPAT is working on a higher level of abstraction as the involved patterns are considered as template or phonetic images with some initial syntax and semantic besides the fact that associative relationships are statistically captured.

Appendix A
Adopted Transliteration

In this Paper, transcription of Arabic letters is based on DIN and [7]. Long vowels are represented through the letters $(\bar{a})$, $(\bar{\imath})$ and $(\bar{u})$ while short vowels as follows: $(fat\dot{h}a, a)$, $(kasrah, i)$ and $(\dot{d}ammah, u)$.

- A root $r_i \in \mathcal{R}$; the set of all roots, is depicted as 2-arguments, such as: (/Latin transliteration/, Abstract Meaning). E.g. the root $(ktb/, Writing)$.
- A pattern $pt_j \in \mathcal{PT}$; the set of all patterns is also depicted as 2-augments: (/ Latin transliteration/, root radicals template positions), whereas C_1, C_2 and C_3 represent root radicals variables such as in $(/maf`\bar{u}l/, maC_1C_2\bar{u}C_3)$; i.e. $f = C_1$, $' = C_2$ and $l = C_3$.
- Analog a root is represent as $(/f`l/, C_1C_2C_3)$

References

[1] Abu-Rabia S., "The Role of Morphology and Short Vowelization in Reading Morphological Complex Words in Arabic: Evidence for the Domination of the Morpheme/Root-Based theory in Reading Arabic", *Scientific Research*, Vol.3, No.4, pp. 486-494, 2012.

[2] Baranyi P. and Csapo A., "Definition and Synergies of Cognitive Infocommunications," *Acta Polytechnica Hungarica*, vol. 9, pp. 67-83, 2012.

[3] Al-Bawab M., Mrayati M., Meer Alam Y., and Al-Tayyan M. H , *Arabic Verbs in a Computerized Ditionary*: Librairie du Publishers, 1996.

[4] Bentin S. and Raphiq Ibrahim, "Phonological Processing During Visual Word Recognition: The Case of Arabic", *Journal of Experimental Psychology Learning, Memory, and Cognition*, Vol. 22, No. 2, pp. 309-323, 1995.

[5] Bentin S. and Frost R., "Morphological Factors in word Identification in Hebrew", *in Feldman L. (ed), Morphological aspects of language processing*: Hillsdale NJ Erlbaum, pp. 271-292, 1994.

[6] Rodrìguez H, Farwell D., Farreres J., Bertran M., Alkhalifa M., Antonia Martì M, Black W., Elkateb S., Kirk J., Pease A., Vossen P., and Christiane Fellbaum, "Arabic wordnet: Current state and future extensions", *in Proceedings of The Fourth Global WordNet Conference*, Szeged, Hungary, 2008.

[7] Fischer W., *Grammatik des Klassischen Arabisch*: Otto Harrassowitz, Wiesbaden, 1972.

[8] Dichy J., "On lemmatization in Arabic A formal definition of Arabic entries of multilingual lexical databases", *Arabic Language Processing: Status and Prospects, Association for Computational Linguistics (ACL-2001), 10th Conference of the European Chapter*, Morgan Kaufman Publisher, France, 2001.

[9] Haddad B., "Probabilistic Bi-directional Root-Pattern Relationships as Cognitive Model for Semantic Processing of Arabic", *3rd IEEE International Conference on Cognitive Infocommunication 2012*, pp. 279 - 284, Dec. 2012.
[10] Haddad B. and Awwad A., "Representing Uncertainty in Medical Knowledge: An Interval-Based Approach for Binary Fuzzy Relation", *The International Arab Journal of Information Technology*, pp. 63-69, January 2010.
[11] Haddad B., "Representation of Arabic Words: An Approach towards Probabilistic Root-Pattern Relationships", *KEOD-2009, International Conference on Knowledge Engineering and Ontology Development*, Madeira, Portugal, 2009.
[12] Haddad B. and Yaseen M., "A Compositional Approach towards Semantic Representation and Construction of ARABIC", *5 th International Conference, Logical Aspects of Computational Linguistics LACL-2005*, Bordeaux, France, LNAI 3492, Springer-Verlag, 2005.
[13] Haddad B. and Yaseen M., "Detection and Correction of Non-Words in Arabic: A Hybrid Approach", *International Journal of Computer Processing of Oriental Languages*, Vol. 20, Number 4, pp. 237-257, World Scientific Publishing, 2007.
[14] Hattab M. , Haddad B., Yaseen M., Duraidi A. and Abu Shmias A., "Addaall Arabic Search Engine: Improving Search based on Combination of Morphological Analysis and Generation Considering Semantic Patterns", *The second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, 2009.
[15] Moukdad H., "Stemming and root-based approaches to the retrieval of Arabic documents on the Web", *Webology*, Vol. 3, Number 1, 2006.
[16] Rastle K., Davis M. H., Marslen-Wilson W.D., and Tyler L.K., "Morphological and semantic effects in visual word recognition: A time-course study", *Languages and Cognitive Processes*, 15 (4/5), pp. 507-537, 2000.
[17] Sabuni A., *Einfhrung in die Arabisitk*: Helmut Buske Verlag, Hamburg 1981.
[18] Sallai G., "The Cradle of Cognitive Infocommunications," *Acta Polytechnica Hungarica*, vol. 9, no. 1, pp. 171-181, 2012.
[19] Soudi A., Cavalli-Sforza V., "A Computational Lexeme-Based Treatment of Arabic Morphology", *Arabic Language Processing: Status and Prospects, Association for Computational Linguistics (ACL-2001), 10th Conference of the European Chapter*, Morgan Kaufman Publisher, France, 2001.
[20] Verhoeven L., and Perfetti C.A., "Morphological processing in reading acquisition: A cross-linguistic", *Applied Psycholinguistics*, 32, pp. 457-466, 2011.
[21] http://en.wikipedia.org/wiki/Arab_world, 14 June 2013.

Bassam Haddad received both of his B.Sc. degree and his M.Sc. degree in Informatics and Theoretical Medicine from Dortmund University, Germany. He obtained his Ph.D. in Informatics from Vienna University of Technology, Austria.

His main research interests are Natural Language Processing, Cognitive Linguistics, Semantic Processing, Statistical Language Models, Information Retrieval and Machine Learning besides Medical Knowledge Representation and dealing with Uncertainty and Fuzziness.

At present, he is Associate Professor of Computer Science at the Faculty of Information Technology besides his position as President Assistant for Global Partnerships and Relations, University of Petra, Amman, Jordan.