From the Generalized Upper Model Towards an Arabic Upper Model

Husni Al-Muhtaseb

Instructor, ICS Department,
King Fahd University of Petroleum and Minerals,
Box # 952, Dhahran 32161, Saudi Arabia.
E-Mail: husni@ccse.kfupm.edu.sa

Chris Mellish

Reader, AI Department,
University of Edinburgh,
Edinburgh EH1 1HN UK
E-Mail: chrism@dai.ed.ac.uk

Abstract: This work introduces the notion of a computational resource for organizing knowledge developed for natural language realization, the Upper Model. The generalized upper model has been implemented mainly for Latin languages. Would such model be able to support Arabic with minor modifications? A limited number of areas where Arabic and English grammars differ are listed to display possible areas where changes to the upper model may be needed to support Arabic generation. The need of adapting the current upper model to support natural language generation for Arabic is highlighted. Procedures for future research work in the field are described.

I. INTRODUCTION

Arabic has had well-established theoretical studies for more than 1000 years. However, if Arabic is compared with other languages, it has received much less modern computational interest. The aim of this research work is to try to make use of some of the Arabic linguistic theories and adapt them to be used in machine processing.

Given some information in some format, how can we produce a natural Arabic text? The given information which is represented in some internal deep structure should be linked to an interface model which has at its lower level an Arabic sentence generator. In English, there are several models that have been used as interfaces between the information to be communicated and the sentence generator. One of these models is the Generalized Upper Model. This model has been -- and is being -- under use, development, investigation, and enhancement for more than 10 years. The model has proved a significant success as been reported by several scholars. Would this model be able to support Arabic?

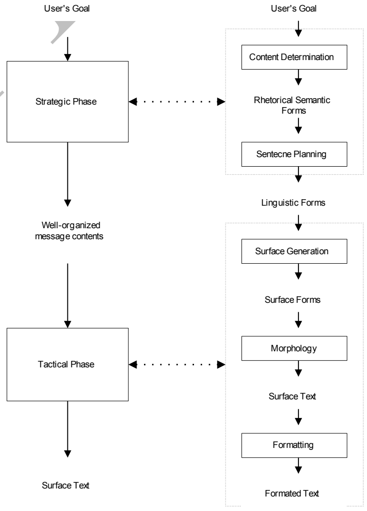

Fig. 1. Natural Language Generation Phases.

An Arabic upper model will provide a reusable- domain-independent interface between any domain knowledge and a realization grammar. Actually, an upper model will also allow the reusability of the grammar. This is very important part for natural Arabic generation and analysis. To adapt the generalized upper model to support Arabic, characteristics of Arabic should be studied. The rest of this paper is organized as follows. The main steps in natural language generation are described in section 2. Section 3 highlights briefly some ideas behind the upper model and some of the developmental stages it went through. In section 4, some differences between English and Arabic are listed. An informal discussion with respect to Arabic and the upper model is presented in section 5. Section 6 is an attempt to present our prediction of the tasks that have to be done, procedures of doing them to adapt the upper model to support natural language generation in Arabic.

II. NATURAL LANGUAGE GENERATION

At least four steps are needed to generate a sentence [1], [2]. The first step is deep content determination which determines the information needed to be communicated. The second step is sentence planning which concerns defining a skeleton or an abstract form for the sentence and the text which will be used. The third step is surface realization where the order of words and syntactic structure are chosen using the output of the previous step. The fourth step is morphology and post-processing where actual inflected words (actual surface structure) are produced. By these four steps sentences are generated from deep structure (internal representation) into the surface structure.

Content determination and sentence planning steps are sometimes considered as a *what-to-say* phase, or *strategic* phase. In this situation, surface generation, morphology, and formatting steps are considered as a *how-to-say* phase, or *tactical* phase [3]. The job of the *strategic* phase is to obtain the needed information and arrange it in a rhetorically coherent manner. The output of this phase is processed by the *tactical* phase to produce a sequence of surface sentences. Two block diagrams of typical system architectures for natural language generation are reproduced from [3] in Fig. 1.

III. THE UPPER MODEL

The Upper Model is a computational resource for organizing knowledge appropriately developed for natural language realization. One of the aims of the Upper Model is to simplify the interface between domain-specific knowledge and general linguistic resources while providing a domain- and task-independent classification system that supports natural language processing [4]. The abstract organization of knowledge - semantic organization - of the upper model is linguistically motivated for the task of constraining linguistic realization in text generation [5]. The upper model has been designed to be a portable, reusable grammar-external resource of information to generate text. It may be considered as an intermediate link between the domain-specific information and the linguistic grammatical core of a text generation system. It has been found that defining the relation between the knowledge concepts of any domain and concepts of the upper model simplifies significantly the task of generation [4].

The upper model can be described as a hierarchy of concepts which is broken into several sub-hierarchies. Concept placement within the hierarchy tells how that concept is expressed in natural language. The principal criterion for attempting to place a new concept within the upper model hierarchy is language use. In general, a concept is a member of a certain class only if this concept is treated by the language as it treats other concepts in that class.

A. The Original Penman Upper Model

The upper model top entity is THING. Originally, (the current hierarchy has more offspring) the THING hierarchy has three offspring: OBJECT, PROCESS, and QUALITY. The authors of the upper model pointed out that one can draw an analogy between these three entities and the linguistic descriptions of noun, verb, and adjective, where OBJECTS are usually nouns; PROCESSES are verbs, and QUALITIES correspond to adjectives. This analogy is useful to demonstrate the connection between the upper model entities and entities (or classes) of linguistic realization.

The OBJECT sub-hierarchy is divided into four subtypes, two of them are according to the consciousness of an entity (CONSCIOUS BEING and NONCONSCIOUS THING) and the other two are according to the decomposability (DECOMPOSABLE and NONDECOMPOSABLE OBJECT). Each of these has its own sub-hierarchy.

The PROCESS hierarchy has been divided into four categories: RELATIONAL, MATERIAL, MENTAL, and VERBAL PROCESSES. Such a categorization follows Halliday's work in [6]. RELATIONAL PROCESSES are the group of processes that relate their participants rather than describing actions of some participants on others. They are of two subtypes: ONE PLACE and TWO PLACE RELATIONS. Each of these has its own sub-hierarchy. The MATERIAL PROCESSES sub-hierarchy contains the intentional and happening actions. It is divided into classes depending upon whether or not the actions can have an actee. These classes are NONDIRECTED and DIRECTED ACTIONS. MENTAL PROCESSES are actions of emotion, cognition, feeling, or decision. The MENTAL PROCESSES sub-hierarchy is divided into two categories: MENTAL INACTIVES and MENTAL ACTIVES. The actor in the latter type is restricted to be conscious-being. The VERBAL PROCESSES sub-hierarchy represents communication actions. It has also two subtypes: ADDRESSEE and NONADDRESSEE ORIENTED.

The QUALITY sub-hierarchy is divided into two sub-hierarchies: The MODAL QUALITIES sub-hierarchy and the MATERIAL WORLD QUALITIES sub-hierarchy. The MODAL QUALITIES sub-hierarchy which represents qualities of wanting, having, or being able to do something, is broken further into sub-hierarchies depending whether upon the quality is condition or not (CONDITIONAL, NONCONDITIONAL) and whether or not the actor is expressed as taking direct responsibility for the process (VOLITIONAL, and NONVOLITIONAL). Qualities that describe things are categorized as the MATERIAL WORLD QUALITIES sub-hierarchy. This sub-hierarchy is further broken into sub-hierarchies depending upon the quality state in gradability (SCALABLE, NONSCALABLE), type of contrast (POLAR, TAXONOMIC), and dynamicness (STATIVE and DYNAMIC). Each of these has its own sub-hierarchy.

It can be noticed that the motivation of breaking an entity into further sub-hierarchies is the language use of the items that are used to realize such entity.

In the next subsection we describe a modified version of the upper model that includes German. This is the merged upper model.

B. The Merged Upper Model

The merged upper model was a result of a detailed comparison of the Penman English upper model and the KOMET German upper model [7]. The purpose of the merged upper model was to serve as the ideational basis for automatic text generation in English and German. The merging criteria which was used was an expansion of the work proposed in [8]. The merging method can be summarized as follows: starting from the topmost entity of the hierarchies of the two models, consider groups of closely related concepts simultaneously. Three alternative operations for each concept are possible in the merging process.

- If two concepts are identical in both models, one of them is chosen.
- If a concept is more specific in one model than a comparable one in the other model, then the more specific concept is considered to be a child for the more general one. In this process the latter concept is extended to include the former one as a more specific concept.
- If comparable concepts hierarchies differ in both models, cross classifications are used.

The merged upper model was used for text generation in English, German, and Dutch within the KOMET project.

One important note to be mentioned here is that the basis of the merging method suggested by Hovy and Nirenburg was that the construction of a merged ontology (model) should be preceded by building an ontology for each language under consideration and organizing the domain entities in terms of that ontology. This information will be used as a guideline for possible adaptation of the upper model to support Arabic generation (see section VI).

It may be worth mentioning here that the differences between the Penman upper model and the German upper model were mainly concerning the hierarchy of PROCESSE types. The Hierarchies of OBJECT and QUALITY can be assumed identical. The Penman upper model PROCESSE hierarchy is more directed toward MATERIAL PROCESSES whereas the German upper model is more directed toward RELATIONAL PROCESSES. Many German relational processes should be defined as material processes to the English grammar generator. The merging solution was to have some overlapping between the two processes types which makes the grammar ambiguous. This solution may produce concepts in the upper model which are not relevant for all languages under considerations.

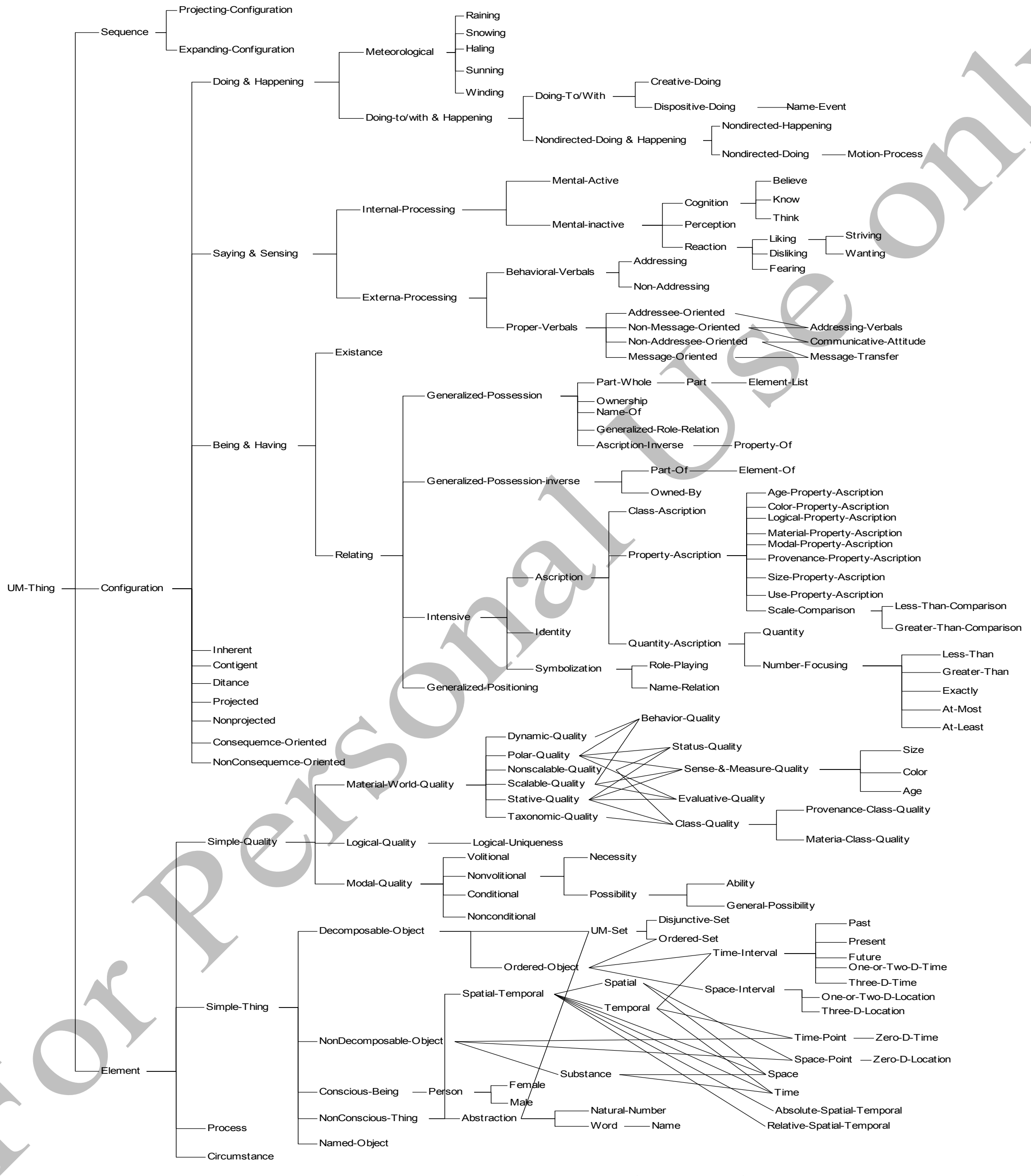

Fig. 2. *UM-Thing Hierarchy.*

The next subsection describes the more generalized upper model that includes Italian.

C. The Generalized Upper Model

Research work similar to the merged upper model has been done to include Italian as a component of the upper model [9]. One main difference between Henschel's work in the merged upper model and Bateman's (and others) work is that there was no comparable Italian upper model that could be taken as a reference for merging. The absence of such model did not allow the principles of Hovy Nirenburg [8] to be fully applied. Modifications - including additions - have only been suggested in cases that are mandatory for Italian. The main generalization process (quoted at length from [9]) was as follows:

> For each sub-hierarchy of the Merged-UM we have individuated a set of relevant Italian linguistic behavior; the behavior for a certain concept then has been

compared to English; if Italian and English/ German behavior were compatible, no modification has been proposed, otherwise some kind of extension has been proposed. The organization for English/ German was then re-evaluated on the basis of the additional information obtained from Italian and a final selection has been made for incorporation in the generalized upper model.

The mandatory extensions and alternations to adapt Italian into the upper model were small according to what was reported and expected by the researchers concerned. Moreover, it has been reported that the majority of modifications could also be applied to English and German.

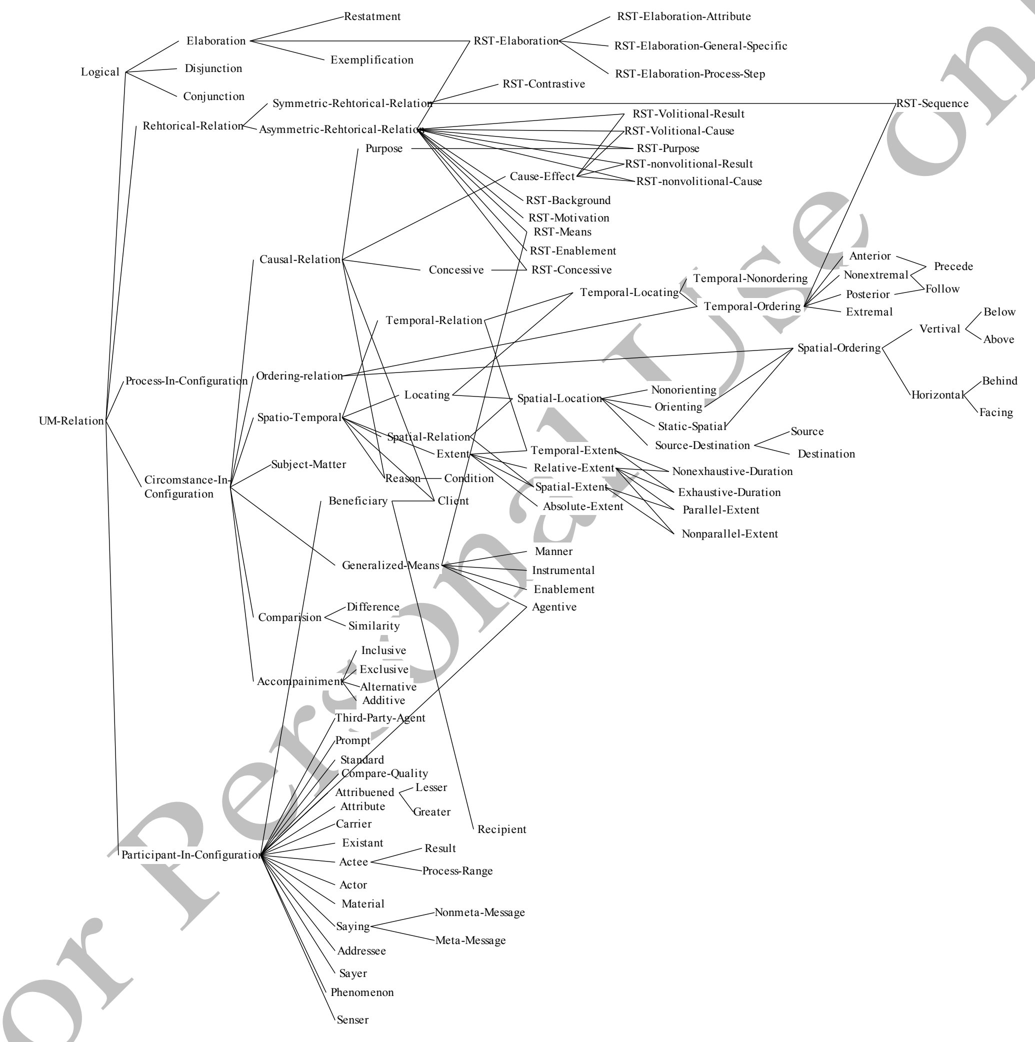

Fig. 3. *Relation Hierarchy.*

A more generalized upper model has been documented in [10] and [11]. This version seems to be more consistent with theory presented by Halliday in [12]. The model of this version has two hierarchies: one is for concepts and the other is for relations. The concepts hierarchy has UM-THING as the top node. UM-THING can be thought as a phenomenon or a situation. It has three main sub-hierarchies which are:

- The configuration sub-hierarchy. A configuration of elements all participating in some situation.
- The element sub-hierarchy. A single object or conceptual element.
- The Sequence sub-hierarchy. A situation where some relations connect various configurations or activities to form a sequence.

The hierarchies of the generalized upper model of [11] are reproduced as Fig. 2 and Fig. 3.

IV. SOME DIFFERENCES BETWEEN ARABIC AND ENGLISH

This section lists some differences between Arabic and English by presenting only Arabic features that look different. It is assumed that the reader has enough knowledge of English to observe the differences. For more details, the reader is advised to refer to [13].

A. Arabic is Categorized as VSO

With respect of word order, Arabic is classified as a VSO (Verb-Subject-Object) language. Linguists used to list methods of showing whether or not a given language could be classified as VSO language. Two of these methods are demonstrated here. For more comprehensive coverage, the reader might refer to [14]. Arabic is an inflectional language where morphological markers may merge with the root of a word affecting its elements, or be affected by its elements.

VSO languages are inflectional languages. A second method to show that Arabic is a VSO language is to check the position of object modifiers. Nominal modifiers should *follow* the noun in VSO languages. This is the case in Arabic. The next two examples may illustrate the situation.

B. Nominal Sentences With No Verbs

Arabic can express a complete meaning in sentences that have no verb at all. Another type of nominal sentence is one which starts by a primate and followed by a verb. The predicate of this nominal sentence is the verbal sentence that comes after the primate.

C. Case Endings

The end-markers of the words are called short vowels or diacritics. There are rules for placing markers on nouns and verbs. These rules depend on the role of the noun (subject, object, reduced, ..), the tense of the verb (past, present, ..) - verbs do not get the reduction end-marker -, the particle used, etc. It is common that end-markers which do not change the shape of the words by adding or deleting letters are not explicitly drawn. The noun <*bâsim*> [باسم], for example, may appear with three different endings. These situations are named: Regularity (nominative) as in <*bâsimun*> [باسمٌ], Opening as in <*bâsiman*> [باسماً], and Reduction (genitive) as in <*bâsimin*> [باسمٍ].

Some end-markers are actually towards the ends of the words but not exactly at their ends. This may appear in plural and dual as in in the word that represent 'the instructors' - <*al-mudarrisûna*> [المدرسونَ], <*al-mudarrisçna*> [المدرسينَ]).

D. Rich Morphology

Morphological markers, particles, personal names, and other pronouns may merge with words affecting their meaning. A simple example can be given to show how rich the Arabic morphology is. *One* word as أنعطكموها which means (Do you want us to give it to you) represents a question that has a verb, an agent, and two patients.

E. Word Derivations

From a single Arabic word, tens of words with possible different meanings can be derived. The denuded original is the base (or source) of derivation. From a denuded original, a past denuded verb (root) can be derived. From the past denuded verb there are up to 15 possible derivations of past augmented verbs. From each of the augmented verbs a confirm verb and an imperative verb can be derived. Moreover, nouns can be derived from each of the past denuded verb, past augmented verbs, and confirm verbs. Some of the derived nouns represent agents, patients, similar qualities, examples of superlative, places, times, instruments, manners, nouns of one act, origins, etc.. For more details the reader may refer to [15] and [13].

F. Personal Nouns

Personal nouns or (pronouns) refer to preceding nouns in sentences. They may be absent (third person), spoken-to (second person), or denoting speakers (first person). Personal nouns may be either prominent or latent. The prominent personal nouns are of two types: connected at the end of words and separated from the words. Latent personal nouns are either obligatorily latent or permissibly latent. An obligatorily latent personal can not be replaced by an apparent noun.

G. The Annullers

Annullers are either deficient verbs or some particles that act similarly to verbs. When one of the annullers is used with a primate and its predicate, it changes their pronunciation and it modifies the time of the described activity, or its state from a probability to an obligation. Particles which are part of the annullers are three groups: <*'inna*> [إنَّ] (indeed) and its sisters. <*lâ*> [لا] (none) of generic negation, and <*mâ*> [ما] (not) and its sisters.

We are not yet sure whether these types of verbs and particles can be mapped to a comparable ones in English.

H. Passive and 'By'

Known transitive verbs are changed to ignored verbs by changing some of the diacritics and/ or adding affixes (infix, suffix, prefix) to the known verbs.

When a sentence is changed to passive by changing the known verb to an ignored verb and making the patient as pro-agent, no place will be left for the agent. Although the agent can be attached to the passive sentence artificially - using some language particles -, It is not common use of Arabic to attach the 'pre-agent' to the passive sentence. Limited number of verbs might accept such attachment.

I. Singular, Dual, and Plural

In addition to singular and plural of the number feature, Arabic has a representation of dual objects. Dual things (and names) have their own rules when syntax and morphology are considered. Different rules are also applied to singulars and different ones to plurals. Some agreements in number (and other features) should be imposed between verbs and names. Rules when to impose agreements are defined.

V. ARABIC AND THE UPPER MODEL

The upper model concepts: THING, PROCESS, and QUALITY as they could be mapped to noun, verb, and adjective are surely valid for Arabic. This may encourage us to assume that a reasonable part of Arabic lies under such concepts. However, when it comes to the basic considerations on which the generalized upper model has been proposed [10] "to motivate sets of distinctions in their lexicogrammatical expression", modification to the upper model to adapt Arabic seems to be necessary.

The classification of Arabic as VSO language may be adapted easily - hopefully - by rearranging words orders of the grammar and without modifying the upper model. When we consider the lexicogrammatical criterion related to Arabic nominal sentences, it seems that either this type of sentences is ignored and mapped, artificially, to several distinct concepts or a necessarily place is to be created to accept such feature.

Case endings situations may be a job for a morphological synthesizer. But some information is needed possibly from the upper model to generate correct end-markers, i.e., number, gender, etc. This information is needed to be examined to assure compatibility. An example for this case is the need to adapt the dual case of number feature in Arabic.

The richness of word derivations of Arabic needs more investigation to decide whether it can get a place in the current upper model or whether it is not directly related to it. A reasonable research work in this area can be found in [16].

The annullers are also spots of investigations. Do they need special classification (and how)? or is it possible to distribute them among the current concepts of the upper model.

VI. CONCLUSION AND FUTURE WORK

The need of the adaptation of the generalized upper model to support natural language generation in Arabic may be done according to the following outline.

A domain needs to be chosen to apply the notion of the upper model. It is good to choose a practical domain that has defined boundaries with limited vocabulary to allow to concentrate more on theoretical issues. Information from the domain should be grouped and studied. The commonly-used grammatical structures should be grouped, analyzed and categorized. Domain's concepts should be identified and classified. Next, two directions could be taken. (1) A generalization of the upper model to support Arabic should be proposed by detailed investigation of the model and Arabic concepts. (2) A limited Arabic systemic grammar should be proposed to accept common structures used in the domain.

With respect to the generalization of the upper model to support Arabic, one or both of the following procedures might be executed.

Procedure 1. This procedure follows the adaptation of Italian into the upper model [9]. For each sub-hierarchy of the generalized upper model a set of relevant Arabic linguistic behavior is to be individuated. The behavior for certain concept is to be compared to English; if Arabic and English are compatible, no modification is to be proposed, otherwise extension should be suggested. Evaluation of whether the suggested extensions are compatible with English should then be studied.

Procedure 2. This procedure is similar to the one suggested in [8]. An Arabic upper model is to be built from scratch, taking into account the Arabic linguistic issues as guidelines. Then the proposed Arabic model is to be merged into the generalized upper model using rules suggested by Hovy [8] and extended by Henschel [7].

Acknowledgments

The first author wishes to thank King Fahd University of Petroleum and Minerals for various supports. Moreover, the Department of AI of University of Edinburgh, where the basis of this work has been started, is acknowledged.

References

[1] Ehud Reiter, Has a consensus NL generation Architecture Appeared, & it psycholinguistically Plausible, *7th Inter. Generation Workshop*, Maine, 1994.

[2] Chris Mellish, Natural language Generation & Technical Documentation, *Saudi Computer Journal*, N. 1, V. 1, 1995.

[3] Ching-Long Yeh, *Generation of Anaphors in Chinese*, 1995, AI Department, Univ. of Edinburgh, Edinburgh, UK.

[4] J. Bateman and R. Kasper and J. Moore and R. Whitney, A general of Knowledge for Natural Language processing: the Penman Upper Model, California, USC, 1990.

[5] J. Bateman, Upper Modeling: A general of Knowledge for Natural language processing, *The Workshop on Standards for Knowledge Representation Systems*, Santa Barbara, 1990.

[6] M A K Halliday, *Introduction to Functional Grammar*, Edward Arnold, London, 1985.

[7] Renata Henschel, Merging the English and the German Upper Model, Darmstadt, Germany, GMD/ Institute fur Integriente Publikation-and Informationssysteme, 1993.

[8] Eduard Hovy and Sergei Nirenburg, Approximatingan Interlingua in a Principled Way, *the DARPA Speech and Natural Language Workshop*, Arden House, New York, 1992.

[9] J. Bateman and B. Magini and F. Rinaldi, The Generalized {Italian, German, English} upper model, *The ECAI94 Workshop: Comparision of Implemented Ontologies*, Amsterdam, 1994.

[10] John Bateman and Renate Henschel and Fabio Rinaldi, The Generalized Upper Model 2.0, GMD/ IPSI Project KOMET, NOTE An experiment in open hyper-documentation, 1995.

[11] John Bateman and Bernardo Magini and Giovanni Fabris, The Generalized upper model Knowledge Base: and Use, *the Conference on Knowledge Representation and Sharing*, Twente, the Netherland, 1995.

[12] M A K Halliday, *Introduction to Functional Grammar*, Edward Arnold, London, second edition, 1994.

[13] Husni Al-Muhtaseb, "The Need for an Upper Model for Arabic Generation", Discussion paper Number 171, Department of Artificial Intelligence, University of Edinburgh, Edinburgh, UK, August 1996.

[14] George Nehmeh Saad, *Transitivity, Causation and Passivization: A semantic - syntactic study of the verb in classical Arabic*, Kegan Paul International, London, 1982.

[15] Antoine El-Dahdah, *A Dictionary of Universal Arabic grammar (Arabic - English)*, Library of Libanon, Libanon, 1992.

[16] S. Al-Jabri and C. Mellish, An Approach to Lexical Choice in Highly Derived Languages, *AISB96 Workshop: Multilinguality in the lexicon*, April 1996.