TOWARDS THE AUTOMATIC GENERATION OF ARABIC TERMINOLOGY

Saad K. Al-jabri[1]

Abstract: in this paper we show that semantic specifications can be described for Arabic derivation. Moreover, semantic interactions that hold between features associated with moulds and meaning representations expressed by roots can be exploited in the process of Arabic lexicalisation. Such a process, with no doubt, will benefit the automatic generation of Arabic terminology. We discuss current approaches to lexicalisation and show that Arabic requires a new framework that is consistent with its nature as a derived language. Our analysis covers linguistic issues as well as computational ones on the linguistic side, we study Arabic derivation from a semantic point of view. On the computational side, we show that semantic aspects of Arabic derivation can be expressed as a semantic taxonomy. Taxonomic organisations are implemented by KR systems that support automatic classification. Such systems are useful when implementing Arabic derivation in which they allow for the KB to be free from redundant information and derivation occurs at classification time. Finally, we demonstrate our approach to Arabic lexicalisation by generating some terminological terms using a generator that is based on semantic principles discussed in this paper.

[1] Data Processing Center-KFSC, Kingdom of Saudi Arabia

I Introduction

In this paper we introduce a new framework for systems that perform Arabic lexical choice focusing on Arabic terminology. **Lexical choice** also called **Lexicalisation,** is the process of mapping semantic representations onto lexical items (words in the language). While native speakers of a language seem to face no difficulties in finding the words to express themselves, computer-based systems need to address the complex character of the task. Lexical choice is at the heart of generation and having good lexicalisation systems is important for systems that will convey ideas in natural languages (e.g., Machine Translation (MT) and Natural Language Generation (NLG) systems). The present-day frameworks for lexicalisation have all been originally developed having English as a target language in mind. English has a morphology which is not as productive in its overall derivation as some other morphological systems such as those found in Semitic languages. This may explain the absence of derivationally motivated approaches to the problem of lexicalisation even for some very productive processes expressed by English. Applying existing approaches, which treat words as isolated items, to the lexicalisation of highly derived languages such as Arabic runs against the spirit of the language and leads to a great redundancy in the knowledge-base.

In Arabic, derivation is a major word-formation process that relates words in the language. Generally speaking, Arabic words are built around consonantal roots which can be linked to core meaning representations. The consonantal roots are modified by derivational processes using derivational affixes. Derivational affixes, in turn are associated with general semantic features. When a derivational process occurs the physical shape of the root will be altered and so as its meaning representation. The interactions between roots and derivational affixes allow us to study derivation not only from a structural point of view but also from a semantic one. Semantic interactions motivated by derivation can be exploited in the process of Arabic lexicalisation. This implies utilising such interactions in building semantic networks that support the mapping between meaning representation and Arabic derived words including terminology, This paper is organised as follows:

- **Section 2** introduces Natural Language Generation and its components focusing on lexicalisation techniques and their

applicability to Arabic as a derived language,

- **Section 3** provides an overview of Arabic derivation and introduces our approach to derivation as a semantic process. This section describes also the derivation of Arabic terminology and their semantic specifications,
- **Section 4** describes the computational model which includes building a semantic taxonomy and describing possible ways for implementing the taxonomy.
- **Section 5** concludes the paper by summarising the ideas presented so far,

2 Lexical Choice in NLG Systems

Natural language generation (NLG) is the process of realizing communicative intentions as text (or speech). Generation is often split into strategic generation (**what to say**) and tactical generation (**how to say it**) [Thompson, 77]. In choosing what to say a generation system will need to consider how the original communicative intention determines the meaning to be conveyed, and then how this meaning can be organized as a sequence of sentence-sized meaning units and how these can be structured. In choosing how to say it a generation system will take the sentence-sized meaning units and will attempt to express them as sentences in the target natural language. The consensus view as to what processes generation looks as follows (McKeown88,Reiter94,Nicolov96]:

Content Determination: The meaning to be conveyed is determined taking into account the communicative goals. The, semantic structure produced might be annotated by rhetorical structure theory relations.

Sentence Planning: Splitting the relatively big semantic structure at the output of the previous stage into units that can be expressed as sentences (or clauses).

Lexical choice: Choosing content words and (abstract) grammatical relations.

Surface realisation: Choosing syntactic structures and introducing closed class words (auxiliaries, prepositions).

Morphology: Performing declination of words.

Synthesising speech/Formatting: Producing speech from the syntactic tree Alternatively, systems producing written output might perform some kind of formatting at this stage.

In this paper we address lexical choice in Arabic that is well-known for its derivational productivity. We first look more closely at the standard notion and techniques

of lexical choice and then discuss how dealing with a productive derivation such as Arabic derivation requires a different model of lexicalisation.

2.1 Lexicalisation Techniques

Lexical choice (lexicalisation) is a process of mapping meaning representations onto lexical items (words in the language). A generation system will need to identify (all) possible words and choose among them the best candidate in a particular situation. Performing lexical choice is non-trivial because the meaning representations are not directly linked to words (i.e., a large number of words may apply in any situation) and choosing the right word requires knowledge not only about the semantics but also about syntax and pragmatics. In general, lexical choice communicates with other processes and decisions in an interactive way and the consequences of choosing a word might be far-reaching and not immediately apparent [Zock9O,Nogier92).

Native speakers of a language seem to face no difficulties in finding the words to express themselves, yet computer-based systems need to address the complex character of the task. At present, researchers have no good insights or indications as to how humans perform lexical choice so easily and current lexical choosers are still far from the performance of humans. Lexical choice is at the heart of generation and having good lexicalisation systems is important for systems that will convey ideas in natural languages (e.g., Machine Translation MT systems and NLG systems).

Early approaches in NLG assumed very simple connections between words and concepts (one-to-one mappings, which have sometimes been referred to as capital letter semantics). A more sophisticated approach was exemplified by the development of **discrimination nets** (also called **(d-nets)** [Goldman75], which map a concept to one of the near-synonymous words that represent possible realisations of that concept. This is achieved by providing for each semantic primitive a decision tree with possible words attached to its leaves. Accordingly, every word sense has associated with it a set of **defining characteristics.** These characteristics are predicates which must be satisfied by the input conceptual representation in order for the input to be realised using that word. For example, the, d-net for the primitive concept INGEST can be related to different verbs such as **eat, drink** and **inhale.** When trying to map a concept to

one of these verbs, the d-net is traversed to determine the realisation of the concept based on a sequence of queries regarding the instance being ingested. Accordingly, the concept will be realised as **eat** if the object is solid or **drink** if the object is liquid and so on. However, the use of highly abstract semantic primitives in d-nets has made them less popular in recent NLG systems. Nevertheless, d-nets have proven to be highly influential for subsequent work in generation [Stede95].

The next generation of lexical choice systems organised lexical knowledge as inheritance hierarchies in which subordinate concepts inherit the properties of their super-ordinates (e.g., the LOQUI generator [Horacek87]). As a result, the primitives are not as abstract as those of d-nets and **inheritance mechanisms** are exploited to reduce the redundancy in the representation. However, the semantic primitives described for concepts remain in general sin-War to those of d-nets. The need to constrain concepts by means of restrictions (motivated by the participant roles of more complex concepts describing different types of situations) encouraged some researchers to explore the possibility of using **frames** in representing concepts expressed by lexical items. Typically, a frame-based system consists of a collection of data structures called frames which represent classes or objects. Each frame has a number of data elements called slots. Each slot contains information about attributes such as values and restrictions on possible fillers [Woods92]. The DIOGENES system provides a **frame** for every lexical item in the Knowledge-Base (KB) in which each frame specifies a concept along with some restrictions on particular roles of the concepts [Nirenburg88). For example, the frame for the word **boy** has the following representation:

Boy:	
CONCEPT-SLOT:	***person***
SEX:	***male***
AGE	***12-15***

Other restrictions on concepts can be described in a similar way. Typically a frame system will include an is-a link (X is-a Y) in terms of pointers to more general frames or frames from which additional slots with default values and other information may be inherited. However, in frame-based systems, there is no formal criterion for when such links should be added to a frame. It is simply up to the designer to decide where a concept should be inserted into the hierarchy. When choosing a word, all frames and their slots need to be examined to filter out unrelated

frames. While more structure is brought to the concept representation by this approach the computational cost resulting from examining all frames is too high. Furthermore, unlike d-nets, despite the use of fine-grained semantic distinctions for each concept there is no guarantee that the system will always come up with an answer,

As an alternative to the above mentioned approaches to lexicalisation, taxonomic knowledge-bases have been introduced to bridge the gap between meaning representations and linguistic resources, Taxonomic knowledge bases for NLG organise world entitics in semantic networks with different levels of abstraction. Concepts appear as classes in structured inheritance hierarchies organised according to their generality (subsumption relationships) where the more specific classes inherit properties from the more general ones. Roles can also be defined to describe relationships between different concepts. As a consequence, there is more structure that can be exploited in lexicalisation [Bateman9l]. For example, the concept expressed by the word **bachelor** can be defined in a taxonomic K.B as follows:

PERSON with attributes **sex:** male, **age-status**: adult, **marital-status**: unmarried,

This concept can be subsumed by a concept expressed by the word **man** which has the following definition:

PERSON with attributes **sex:** male, **age-status:** adult.

The second concept is more general since it contains less information, A typical scenario for language generation present-day is a semantic representation based on some taxonomic knowledge-base which has to be verbalised [Stede95]. In contrast to frame-based systems, taxonomic knowledge-bases define the is-a relationships by an external semantic criterion independent from the data structure [Woods9l]. Some systems utilise such representations to implement lexical choice by means of *automatic classification* (IDAS [Reiter92).[2]

2.2 Morphological Derivation and Current Approaches

Most of the NLG research has concentrated on Indo-European languages and within this almost entirely on English. The present-day frameworks for generation--the systemic approach, the functional unification approach and the classification approach have all been originally developed having English as a target language in mind. Lexicalisation models, as a consequence, have been geared mainly towards English. Such models can be

[2] Automatic classification refers to the ability of a system to automatically classify with respect to an existing taxonomy [Woods 921.

abstractly summarised as mapping semantic, configurations onto lexical items (using some variants of the, techniques described above), the chosen lexical items in the course of surface realisation are augmented with morphological features; after the syntax tree is generated a morphological postprocessor inflects the words. This means that morphological derivation has not been considered in the existing approaches to lexicalisation. For example, present-day generators (e.g., the Penman project [Bateman95]) do not consider the lexical relatedness between verbs and their derivatives such as **write** and **writer** despite the productive nature of this derivation in the English language. In general Current approaches to lexicalisation tend to treat concepts as related entities; words, on the other hand, are viewed as isolated items that just happen to be attached to concepts. These approaches have limitations when they are applied to highly derived languages such as Arabic. For example, in the derived lexicon of Arabic (i.e., a lexicon that contains all words derived from consonantal roots in Arabic); there are clear derivational links between words and ignoring them in generation (and more importantly in lexicalisation) runs against the spirit of the language and leads to a great redundancy in the KB.

3 Arabic Derivation

Arabic derivation forms stems (verbs and nouns) by means of consonantal roots and derivational affixes, A single root can give rise to different derivationally related stems, As a consequence, the majority of Arabic words (verbs and nouns) are built up from a relatively small number of roots. For example, an Arabic root and some of its derivatives are listed in Table 3. 1:

Arabic Lexeme	Transliteration	Derivational mould		Arabic root	English lexeme	English root
كَتَبَ	*kataba*	*fa'ala*	فعلَ	*ktb*	*to write*	*write*
كاتب	*kātib*	*fā'il*	فاعل	*ktb*	*a writer*	*write*
كاتَبَ	*kātaba*	*fā'ala*	فاعلَ	*ktb*	*to correspond*	*corespond*
مكتب	*maktab*	*maf'al*	مفعل	*ktb*	*an office*	*office*
مكتبة	*maktaba*	*maf'alat*	مفعلة	*ktb*	*a libirary*	*library*

Table 3.1 : Some lexemes derived from (ك ت ب k t b)

The sub-regularity associated with Arabic derivation led traditional grammarians, to develop a **morphonological** theory that describes Arabic derivation (morphonology is constituted by the interaction of morphology and phonology [Dressler85]). In this theory, medieval grammarians used notations, which we will call **moulds**, to mediate between words and their morphological shapes[3] A mould is a template that reflects the occurrence of consonants and vowels in a particular word structure. In the mould system, the consonantal root is represented by three or four selected letters (ف ع ل f, l) or (ف ع ل ل f'll) or depending on whether the root is trilateral or quadriliteral (trilateral root consists of three Arabic consonants while a quadrilateral root consists of four Arabic consonants). Vowels and other derivational affixes are copied to the mould form unchanged, For example, the third column in Table 3.1 associates all lexemes in the Table with their derivational moulds.

Moulds were introduced, traditionally, to describe derivation and to account for the productivity of Arabic word-formation. They are widely regarded as a classification system for Arabic derived Words. Arabic trilateral root-based verbs are classified into fifteen moulds. Each mould reflects one instance of a possible triliteral consonantal root modification. These moulds are generally adopted following the work of Arabic Grammarians. The convention in Western studies is to refer to them by their Roman numbers (I - XV) [Wright75]. The derivation of quadrilateral verbs and nouns is also described in traditional work. Moulds for nouns are numerous and less regular than verb moulds. Unlike verb moulds, noun moulds have no conventional numbering system.

Arabic derivation associates verb moulds with semantic features such as causality, intensity, reciprocity, reflexivity and human characteristics. For example, mould II is always associated with intensity while mould IV is associated with causality. Noun moulds are also associated with features to describe objects such as action agents, action patients, tools, instruments, places and machines.

3.1 The Derivation of Arabic Terminology

In this work, we will refer to derivation as the morphological term that describes the process of word-formation by means of the interactions between roots and derivational affixes (including infixes) in Arabic morphology. In particular, it is equivalent to what Arabic grammarians call (اشـــتقاق

lstiqaq). Derivation, in this context, has been a major formation process that incorporates words in the Arabic language. The set of Arabic terminology includes many derived words. Derivation can also be applied to form new ones, For example, the following are some derived terms that have been "recently" introduced in connection with **office supplics** : مفكـرة ***Mufakkirat*** **"agenda",** دباسة **dabbast "stapler",** تقـويم **Taqwym "calendar",** خرامة **kharramat "Perforator and** مقلمة **maqlamat "Pen case"** (76مذكـرة) All these terms and many more are derived from consonantal roots according to well defined derivational rules. The nominal derivation plays an essential role in terminology derivation. Arabic morphology provides a set of moulds that support the derivation of terminological terms under specified Subjects such as *machine place and time.* Table 3.2 lists some **noun** moulds in connection with their semantic features.

Mould	Transliteration	Feature	Example	Transliteration	English
فاعل	*Fā'il*	Action-agent	كاتب	*Katib*	An author
مفعول	*Maf'wl*	Action - patient	مكتوب	*Maktwb*	A document
فعال	*Fa''āl*	intensified-agent	علام	*'Ilām*	a scholar
مفعل	*Maf'il*	an event time/placee	موعد	*Maw'id*	An appointment
مفعل	*Maf'al*	An vent place	مكتب	*Maktab*	an office
مفعال	*Mif'al*	Instrumenrt	مفتاح	*Miftāh*	a key
فاعول	*Fā'wl*	Tool	ساطور	*Sātwr*	a chopper
فعالة	*Fa''ālt*	Machine	حسابة	*hassābt*	a calculator

Table 3.2: Noun moulds and semantic features

3.2 Semantic Interactions in Arabic Derivation

Derived words in Arabic are formed by applying derivational affixes to consonantal roots. Such words are realisations of independent concepts that describe *actions, States, processes and objects. A* consonantal root accounts for a semantic representation that appears in a set of derivationally related words. Roots in non - concatenative morphology are discontinuous morphemes (i.e, they can be interrupted by other

morphemes). It is almost always difficult (if not impossible) to describe precisely the meaning of a consonantal root. We associate consonantal roots with core meanings. A core meaning is defined as follows:

***A core meaning is a semantic representation that appears in a set* of derivationally related words.**

For example, the Arabic words appearing in Table 3.1 share one consonantal root, that is ك ت ب k t b This root is associated with a core meaning that has something to do with the activity of writing.

It is also necessary to consider the semantics of derivational affixes appearing in moulds, These affixes represent morphemes that have been added to the mould in a layered process. The integration of various layers in the mould from result in an integration of different morphemes to describe a set of semantic features. A semantic feature is associated with its mould as a whole unit and cannot necessarily be described by means of individual morphemes involved in the layered process. Accordingly, a semantic feature accounts for another level of semantic representation and can be defined as follows:

A semantic feature is a unit of meaning that is associated with a mould and that can be used to distinguish one concept realised by a word sharing the same consonantal root.

For example, the mould appearing in the first row of Table 3,l describes a general action as a semantic feature. Another semantic feature is associated with the mould in the third row of the same table which describes a **more** specific action: a reciprocal action. **Reciprocation,** in this case, is a semantic feature associated with the **mould** (*فاعل Fā'ala*). Semantic features associated with moulds can be viewed as semantic generalisations in the Arabic word-formation system, Such generalisations usually help native speakers of Arabic to make educated guesses about new words they have never heard before. This would mean that mould semantic features could help in analysing the Arabic word-formation system not only for the existing words but also for potential new words in the language. In fact, it provides a high level representation of semantics that integrates words (including new ones) into the language system by linking their semantic features.

3.2.1 Concept Formation

The above discussion indicates that forming a concept which is realised by a derived word requires two major semantic components, a core meaning and mould semantic features, Semantically, roots are representatives of core meanings in the language. A single root describes a core meaning that does not account for the full meaning of a particular concept. in order to do this, it needs additional semantic features associated with an applicable mould. Moulds, on the other hand, are abstractions, which can say something about the common concept of the meanings of the words that they represent but cannot tell the whole story. Putting together the two semantic aspects (i.e., core meanings and mould semantic features) allows the formation of concepts that describe particular situations through the construction of derivatives. For example, the concepts realised by the Arabic words in Table 3.1 can be represented by means of the semantic interactions that hold between the core meaning associated with (ك ت ب k, t, b), on the one hand and the semantic features associated with their derivational moulds on the other. Thus we are going to view derivation as a parallel process of word and concept formation.

The semantic specifications of derived concepts vary from one class of concepts to another. The semantic variation between classes can be identified by the type of semantic features associated with moulds and the type of core meanings the modify. We classify derived concepts into three main classes, namely, *action, state* and derived *object*. Semantic features that modify core meanings under *action* include causation, intensification, reciprocation and reflexivity. These features are associated with moulds IV, II, III and VII respectively[4]. Semantic features under state modify state core meaning to generate stative concept such as feeling and emotion (e.g ف ر ح *fariha*). Concepts that describe derived object utilise features that are associated with noun moulds and are linked to both action and state concepts. To define domains for each class, core meanings are also classified into action and state core meanings. Such classification is motivated by the appearance of a root in a verb represented by mould I, whether that verb describes an action concept or a stative one. Accordingly, action concepts apply only to action core meanings while stative concepts apply only to state core meanings. Derived objects are linked to both of them.

[4] Not that a feature such as reflexivity could be associated with more than mould.

When forming a concept from semantic descriptions, features that can be associated with moulds are linked to their moulds and representation that can be associated with core meanings are linked to corresponding consonantal roots. This results in providing a mould and a root which are enough to form a word by mapping the chosen root into the selected mould template. For example, a semantic description that includes intensification as a feature and computing as an activity should be linked to mould II (فعل fa''ala) and the root (ح س ب h s b) . The root then can be mapped into mould II template to form the Arabic word (حسب hassaba) "to calculate". Further more. The concept that includes in its description the same concept above and, in addition, a pointer to indicate that **machine derivation** is at focus rather than the action concept, such a description will result in mapping the derivation into mould (فعالة fa''alat) while keeping the root as it was before. This leads to the formation of the Arabic word (حسابة hassabat) "a calculator" as a derived object.

However, since semantic features associated with moulds are too abstract, participant roles such as *actor and* actee may be used to direct semantic descriptions to the appropriate classes.

3.3 Productive Concepts

Since our concerns are going to be with the derivation of Arabic terminology, we introduce here productive concepts which can be defined as follows :

> ***A productive concept is a concept that results from the interactions that hold between core meanings and the semantic features associated with noun moulds.***

According to the above definition semantic interactions that result in forming terminological concepts fall under productive concepts as nominal derivation.

Traditionally, Arabic nouns can be grouped into two classes as regards to their origin: **primitive** and **derivative.** The primitive nouns are all substantive (e.g., رجل *ragul "a man").* The derivative nouns may be substantive such as (مفتاح *miftah* "a key") or what corresponds, in English to adjectives such as (مريض *maryd* ",sick) [wrig\it51]. Moulds for nouns are numerous and less regular than verb moulds. Unlike verb moulds, noun moulds have no conventional

numbering system. The majority of derived nouns are linked to verb moulds and their derivation is described traditionally by means of modified verb moulds.

In this work, productive concepts are regarded as being formed by applying features associated with noun moulds to core meanings which are realised by derivative nouns as productive stems. The most common productive features in Arabic derivation include *action-agent, action-patient, place, instrument and machine* [Al-Rajihi84].

In the following we describe the semantic interaction of the most common productive features that have been used frequently in deriving terminology, namely, *place, instrument and machine.*

3.3.1 Place

Arabic derivational processes provide the means for deriving concepts that describe other aspects of actions such as physical location. *Place* as a semantic feature, is defined as follows:

> ***Place is a semantic feature that interacts with action core meanings to derive, without any reference to a particular place, concepts that describe places in which actions are contained.***

Place, as defined above, is associated with the mould (مفعـل *maf'al)* and is linked to default concepts that are represented by the mould I --- (فَعَـل *fa'ala).* For example, from (كـتب *kataba)* " to write, (سـكن sakan*a)* " to inhabit' and (شرب sariba) "to drink " we could derive (مكـتب *maktab)* "an office", (مسـكن *maskan)* "a house" and (مشـرب Masrab) "a place for drinking" respectively. Arabic derivation provides other moulds associated with the location feature. However, they are rarely used and we do not consider them here.

3.3.2 Instrument

Instrument as a feature is defined as follows:

> ***Instrument is a productive feature that applies to causative actions to describe manual tools that can be used to perform actions without need for external power.***

Instruments are derived from causative actions. However, it is obvious that not all causative concepts can be linked to instruments. The domain for this derivation can be defined by identifying all instrumental actions as one class. However, identifying an action as an instrumental one is a matter of the synchronic usage of the speakers of the

language, and hence, cannot be predicted in advance. Nevertheless, applying this derivation to causative actions can be used to suggest certain derivation realised by words that satisfy the needs for new instrumental terminology.

Traditional studies associate instruments with three moulds that are derived from the causative stems represented by mould I, namely, (مفعـال *mifal*, مفعل *mif'al* and فعالـة *fa"alat*). We, on the other hand, associate instruments with the first two moulds. The remaining mould describes, for us, a different semantic feature which we describe later. The moulds with which we associate instrument are linked to the derivation of verbs under the first mould (mould I). For example, (مفـتاح *miftah*), "a key" is linked to the concept realised by (فتح fataha) "to open". Similarly, (مقـبض *miqbad*) "a Handle") is linked to the default concept realised by (قبض ..*qabada)* "to grasp").

3.3 Machine

Traditionally, **machine** noun derivation is considered to be part of instrument derivation. However, with the number of commonplace machines growing, Modern Standard Arabic (MSA) tends to restrict this derivation to one form which becomes more productive by time, that is (فعالة fa"alt). Below we give a definition for machine as a semantic feature:

> ***Machine is a semantic feature that applies to causative action to describe a machine that involves either external power (such as electricity) or a considerable amount of force when performing an action***

Machine, as we define it, is associated with the mould (فعالـة*fa"ālt*) which we will link to mould II but not to mould I (as suggested by some traditional studies) due to the force property expressed by machine concepts. Examples of this derivation are found in words such as: (سيارة*sayyārat*"a car", طيـارة *Tayyārat*"an "an aircraft", غسالة *gassalat* "a washing machine") which are linked to intensified concepts derived from سير *sayyāra "to walk/move* , طيـر *Tayyāra* "to fly", غسـل *gassāla* "to wash).

4 The Computational Model

in this section we discuss how semantic interactions expressed by Arabic derivation can be organised as a semantic network that

support Arabic lexicalisation. Semantic interactions motivated by derivation as we describe above, can be expressed into two layer semantics. An outer layer dealing with the semantic features associated with moulds and an inner layer dealing with core meaning. The outer layer semantic features, in our domain (i.e, terminology derivation), are *place, instrument and machine.* These features need to interact with action concepts (causative and cause neutral). Action concepts, in turn, are formed by means of interactions between *action* as an outer layer semantic feature and core meanings as inner layer representations, These semantic interactions can be organised as a special type of semantic networks known as taxonomic organisations.

A taxonomy is a semantic network that organises knowledge according to its level of generality. In order to do this, a network defines some links that relate more specific classes (represented by nodes at some level of the network) to more general ones (represented by nodes at higher levels). Accordingly, the more specific classes are said to **inherit** information from the more general classes and the more general classes are said to **subsume** the more specific ones. For example, we could organise our domain in a taxonomy to express the above relationships as in Figure 3. 1.:

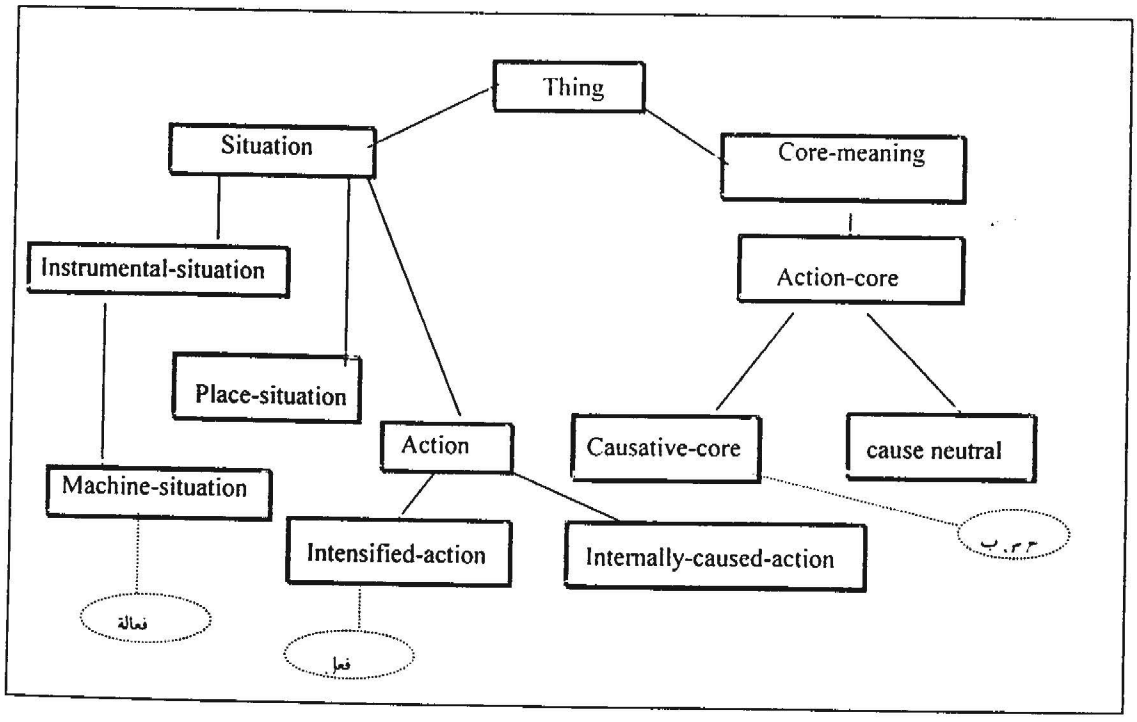

Figure 3. 1: A prototype taxonomy

In Figure 3. 1, the topmost class is thing which subcategorises for two subclasses: Core-meaning and Situation. Core-meaning subsumes inner layer semantic classes while Situation subsumes outer layer classes. Moreover, Situation is a general class that interface our domain with the external world where every derivation in our domain is linked to this class. For example an action concept is subsumed by Situation (e.g, the concept expressed by حسب *hassaba*" to calculate" which is realised by mould فعل fa'ala). Similarly, the concept that is linked to the action concept in the previous example and focuses on machine derivation is also subsumed by Situation (e-g., the concept expressed by حسابة *hassabt which* is realised by mould فعالة fa''alat), as shown in Figure 3.1.

In a taxonomy the subsumption relation permits the assimilation of new concepts into it In Knowledge Representation (KR) systems attempts have been made to allow the automatic insertion of a given description into a structured taxonomy. KL-ONE was the first system to allow the **automatic classification** of new concepts by assimilating them into the taxonomy on the basis of their subsumption relationships [Brachman85]. Once a concept is assimilated in a taxonomy it inherits properties from other classes based on its subsumption relation These properties, in our case, include linguistic properties such as moulds and roots.

4.1 Choice of Formalism

In order to implement lexicalisation based on derivation we looked at various lexical formalisms and explored the usefulness of these formalisms with regard to Arabic derivation. In general, lexical representation formalisms can be classified into unification-based systems inheritance-based systems and those that combine both mechanisms. Basic unification-based formalisms do not provide a proper way to represent non-monotonic inheritance, which is needed to eliminate redundancies in the representation of semantic aspects of Arabic derivation. Systems that are based only on inheritance mechanisms have limitations when used to implement the two-layer semantics. Such systems do not support an automatic way to constraint derivation and do not support **automatic classification.** When looking for a formalism for expressing the semantics of Arabic derivation, we suggest that classification-based systems seem to be the best available choice. Classification-based systems are built around taxonomies and integrate knowledge representation

mechanisms such default and **multiple inheritance,** and structured relations such as **subsumption.** This integration allows for expressive representation of Arabic derivation. The reasoning mechanisms of classification-based systems are based on **classification** and **inheritance**. The knowledge-base is easy to maintain and modify if necessary. In addition, semantic descriptions can be efficiently mapped onto language specific syntactic and lexical representations.

4-.2 Building the Generator

The Generator for Lexemes in Arabic Derivation (GLAD) is implemented as a prototype to map disambiguated semantic descriptions to Arabic derived words [AlJabri97]. GLAD is composed of severed components **a** semantic disambiguator, a **user-interface, knowledge components** and an **automatic classifier** .Of these, the semantic disambiguator is not implemented and the automatic classifier is the I 1 classifier [Reiter92]. The input to the system consists of semantic descriptions describing a particular situation. The input specifies a core meaning and role specification sets for situation participants. The classifier proceeds by classifying role specification sets to identify role relations. After this initial classification, role relations and the core meaning are used to build a general-input class that is interpreted as an II class definition. The general-input class is used in another stage of classification to place the input in the proper place in the situation subtaxonomy. After classifying the general-input class, inference mechanisms are used to reason about the surface realisation of the classified input (using the inherited root and mould). Information resulting from this classification process (the output of the proper generation) is pawed on to the user-interface. The output specifies morphological, syntactic and semantic information for the classified input. Morphological information includes a root and an eligible mould. Syntactic information describes syntactic arguments for the classified concepts and their case-ending marks. Finally, semantic information names the parent(s) of the classified class in the taxonomy.

4.3 Examples

GLAD was originally designed and implemented to generate Arabic derived words making a distinguish between actual and potential words [Al-JABRI9]1. Actual words are those exist in the Arabic lexicon and, in addition, their use is well established by the speakers of the language. Potential

words, on the other hand, are derived from consonantal roots and we cannot judge their establishment in the language use. To do so we need to consult an MSA corpus. Such a corpus is, unfortunately, not available, The new Arabic terminology are potential words.

We tried GLAD with some semantic inputs (expressing the principles stated in this paper) to show the possibility of generating now Arabic terminology from disambiguated semantic descriptions. Some inputs and their results summarised in the following:

Input	Outputs	
Core-meaning : rule-1	Mould : مفعلة	root : س ط ر
Features : instrument- situation, action	Word: مسطرة	
Core-meaning : wash-1	Mould : فعالة	Root : غ س ل
Features : instrument-situation, intensified-action.	Word :غسالة	

Table 4. 1: Example 1

Table 4.1 shows the normal behaviour of the generator when disambiguated semantic inputs are provided to generate actual words. These inputs include a description of a core meaning (GLAD KB us" English verbs to name core meanings that are associated with roots, the number augmented to each verb indicates the corresponding sense as defined in the WordNet [Miller 95]), and a set of semantic features (usually linked to those associated with moulds). The reader should note that the generator utilises a complex knowledge-base which is different in its size from what we have here. Semantic restrictions in the original knowledge-base are well defined in a way the prevents ill-derivation from taking place during classification time. When the first input mentioned above is given to the generator it will be classified tinder causative action and instrument situation at the same time. The mould for instrument will override any previous value and the inference mechanisms will reason about the root. This will be followed by another step from the mechanisms to fill in the mould template using root letters. The result is a derived word that describes an instrument. The same process will repeated for the second input. However, the second input describes a more

specific type of instrument than the first one. This is due to the introduction of intensification in its set of semantic features. Accordingly, it will be classified under machine. The, root in this case will be mapped to the mould that is associated with machine derivation (فعالـــــة fa"ālat).

Input	Outputs
Core-meaning : mince-1	Mould: مفعلة root: ف ر م
Features : instrument- situation, action	Word: مفرمة
Core-meaning : mince-1	Mould: فعالة Root: ف ر م
Features : instrument-situation, intensified-action.	Word: فرامة
Core-meaning : cut-1	Mould: مفعلة root: ق ط ع
Features : instrument- situation, action	Word: مقطعة
Core-meaning : cut-1	Mould: فعالة Root: ق ط ع
Features : instrument-situation, intensified-action.	Word: قطاعة

Table 4. 1: Example2

Examples in Table 4.2 are meant to test the generator with inputs that aim to generate possible terminology for new concepts appear in other languages such as the one realised by the English word **shredder.** To do so, we need to suggest a core meaning that is linked to a consonantal root. Moreover, semantic features need to be mentioned in the input in order to allow the derivation of a possible mould. In Table 4.2 the specified concept is linked to two possible core meanings associated with two roots, namely, (ف ر م *f r m and* ق ط ع *q t '*) which appear in the two Arabic verbs (فرم *farama)* "to mince" and (قطع *qata'a)* "to cut". In addition, we associate each core meaning in the input with two sets of semantic features characterised by the presence/absence of intensification. The output shows four different possibilities

expressed as derived words. However, the concept realised by the English word **shredder** involves considerable amount of force and external power. This means that we should exclude inputs that do not indicate intensification. Accordingly, The correct derivation **can** be achieved through mould فعالة *fa'ãlat* machine derivation. The usefulness of core meanings suggested for this derivation should be left to the speakers' common-sense or, alternatively, should be judged by consulting a modern corpus.

5 Conclusions

In this paper we introduced a new approach to Arabic lexicalisation that is based on derivation. We argued that Arabic has a morphology that is different from that of English and existing lexicalisation techniques are not necessarily useful for Arabic. Derivation in Arabic is a major word-formation process that associates derivational processes with semantic features. These features interacts with meaning representations expressed by Arabic consonantal roots---core meanings. The semantic specification of Arabic derivation can be exploited to states links between semantic descriptions and derived words in their final forms.

The system we proposed in this paper is meant to generate Arabic terminology. Arabic derivation provides the means for mapping technological concepts under specific subjects into certain moulds. We briefly discussed the specifications of semantic features motivated by these moulds. We demonstrated our approach by some examples from a generator that is designed and implemented based on semantic aspects of Arabic derivation.

The arguments introduced in this paper open the door towards a thorough investigation of semantic aspects of Arabic derivation. Such aspects, when carefully studied will benefit the process of generating Arabic terminology and enhance the performance of systems that implement semantic mapping such as NLG and MT systems.

6 References

- [الراجحي 84] عبده الراجحي التطبيق الصرفي، دار النهضة العربية، لبنان 1984
- [مذكور 76] مجموعة المصطلحات العلمية والفنية التي أقرها المجمع، المجلد الثامن عشر، مطبعة الأمانة، مصر، 1976.
- [Al-Jabri97] S Al-Jabri, Generating Arabic Words from Semantic descriptions, Ph.D thesis, Edinburgh Univ., Edinburgh, UK, 1997.
- [Bateman91] J. Bateman, The theoretical status of ontologies in natural language processing, In susanne \peru and Brite Schmitz, editors, proceedings of the workshop on Text Representation and Domain Modelling-Ideas from Linguistics and Al, University of Berlin KIT Report 97, October 9th – 11th, 1991,
- [Bateman95] J. Bateman. R. Hensschel, and F.Rinaldi, The generalized upper model 2.0, GMD/IPSI Project KOMET, Germany, 12 1995.
- [Brachman85] R. Brachman and J. Schmolze. An overview of the KL-ONE knowledge representation system, Cognitive Science, 9-.171 216,1985,
- [Dressler85] W. Dressler, Morphology : the dynamic of derivation, Karoma Publishers, Inc., Ann Arbor, USA, 1985.
- [Goldman75] N. Goldman, conceptual generation, In R. Schank, editor, Conceptual In formation Processing, North Hoand, Amsterdam, 1975.
- [Horacek87] H. Horacek, Choice of words in the generation process of natural language interface. Applied artificial intelligence, I-117-132, Hemisphere Publishing, Washington. D. C., USA, 1987,
- (McKeown88] K. Mckeown and W. Swatout, Language generation and explanation. In M. Zock and G. Sabah editors, Advances in Natural Generation, An Interdisciplinary Perspective, volume 1, Pinter, London, 1988.
- [Miller95] G. Miller, WordNet: A lexical database for English, Communications of the ACM, pages 39-41, November 1995.
- [Micolov96) N. Nicolov, C, Mellish and G. Ritchie. Approximate Generation from Non-Hierarchical Representations. In Donia Scott-, editor, Proceedings of the 8th INTERNATIONAL Workshop on Natural

Language Generation, Herstmonceux Castle, UK, 13-15 June, 1996.

- [Nirenburg88] S. Nirenburg and I. Nirenburg, A framework for lexical selection in natural language generation. In Proceedings of the 12th International conference onComputational Linguistics, pages 471-475, Budapest, 1988 (COLING-88).

- [Nogier92] J. Nogier and M. Zock, Lexical Choice as Pattern Matching, In Timothy E., Nagle Janice A., Nagle Laurie L. Gerholz and Peter W., Eklund, editors, Conceptual Structures: Current research and Practice, Ellis Horwood Series in Workshops, pages 413-436, Ellis Horwood Limited, London, England, 1992.

- [Reiter92] E. Reiter, C. Mellish and J. Levine. Automatic generation of online documentation in the Idas project, In Proceedings of the third Conference on Applied Natural Language processing (ANLP), pages 64 - 71, 1992.

- (Reiter94] E. Reiter, Has Consensus NL Generation Architecture Appeared, and is it Psychologically Plausible, In the Proceedings of the 7" International Workshop on Natural Language Generation, pages 163-170, Kennebunkport, Maine, U.S, 21-24 June 1994.

- [Stede95] M. Stede, Lexicalisation in natural language generation, A survey, Artificial Intelligence Review, 8:309-366, 1995

- [Thompson77] H. Thompson, Strategy and tactics: A model for language production, In Papers from the 13th Regional Meeting, Chicago Linguistic Society, pages 651-668, Illinois, 1977.

- [Wood9l] W. Wood, Understanding subsumption and taxonomy, In J. Sowa, editor, Principles of Semantic Networks, pages 45-94, Morgan Kaufmann Publishers, fNC., San Mateo, California, USA, 1991.

- [Wood92] W. Wood, The KL-ONE family, In F. Lehmann, editors, Semantic Networks in Artificial Intelligence, Pergamon Press, Oxford, 1992.

- [Wright75] W. Wright, A Grammar of the Arabic Language, Cambridge University, UK, 1975,

- [Zock9O] M. Zock. La génération interactive de langage : comment visualiser le passage de l'idée à la phrase., In J. Anis and J. Lebrave, editors, Text et ordinateurs: guistique de Paris X, Nanterre, 1990.

نظام شبكات عصبونات إصطناعية للتعرف الآلي على خصائص نطق اللغة العربية

UN SYSTEME CONNEXIONISTE MODULAIRE POUR LA RECONNAISSANCE DES TRAITS PHONETIQUES DE L'ARABE.

سيد أحمد سلواني * جون كيلان**

ملخص

نعرض في هذه المداخلة نظاما جديدا للتعرف على الكلام العربي الناطق يعتمد في تصميمه على مفاهيم شبكات العصبونات الاصطناعية. يمكن حصر مميزات الطريقة التي انتهجناها في النقاط التالية:

- المنهاج يتلاءم وخصائص اللغة العربية لأنه يأخذ بعين الاعتبار على سبيل المثال الحروف المدودة والمشدودة وذلك بتسخير شبكات مختصة لمعاينتها.

- في تحليل الإشارة الكلامية اعتمدنا على التحليل الحسي المنبثق من معاينة حاسة السمع عند الإنسان مما يؤدي إلى تحسين قابلية التعرف.

- شبكات التخصصة تروض باستقلالية تامة عن بعضها البعض مما يسهل عملية التعميم وبالتالي الاندماج في الأنظمة المعمول بها حاليا في ميدان البحث والتعرّف على المصطلحات المنطوقة مثل برنامج : (Netscape Communicator)

المعطيات التي استقيت من بنك معلومات متكونة من تسجيلات ستة مشاركين (3 رجال و3 نساء). وفي النهاية وعلى ضوء النتائج المحصل عليها بتطبيق النظام الإجمالي على جمل متوازنة حرفيا، نستخلص ونعلق على قدرة النظام في التعرف البحت وكذا على مدى استيعابه لخصائص اللغة العربية مما يؤهله للاستعمال في برامج الترجمة الآلية.

Résumé :

Nous décrivons dans ce papier une approche de reconnaissance automatique de traits phonétiques de l'arabe qui utilise des sous-réseaux de neurones formels. L'analyse acoustique étant effectuée par la technique de prédiction linéaire perceptive (PLP). Un corpus d'apprentissage et de test prononcé par 6 locuteurs algériens est établi afin d'évaluer les performances du système. Nous discutons les résultats obtenus par le système d'identification des macros-classes en nous focalisant sur les traits caractéristiques de l'arabe que sont l'emphase, la gémination et la durée. A l'issue, nous concluons sur les performances relatives des méthodes neuromimétiques en identification pure ainsi que leur capacité à prendre en ligne de compte les problèmes liés à la durée phonologique.

MOTS CLEFS : Reconnaissance de la parole - réseaux de neurones - Analyse perceptuelle - langue Arabe.

* جامعة هواري بومدين للعلوم والتكنولوجيا – الجزائر

** جامعة جوزف فوريي قرونبل – فرنسا