7

Grammar-Lexis Relations in the Computational Morphology of Arabic

Joseph Dichy[1] and Ali Farghaly[2]

[1] *Université Lumière-Lyon 2, ICAR research lab (CNRS/Lyon 2), 86, rue Pasteur, 69365 Lyon Cedex 07 – France*
Joseph.Dichy@univ-lyon2.fr

[2] *Oracle USA, 400 Oracle Parkway, Redwood Shores, California 94065 – USA*
Ali.Farghaly@oracle.com

Abstract: Grammar-lexis rules and relations ensuring correct insertion of major lexical entries (nouns, verbs and deverbals) play an essential part in the computational morphology of Arabic. This chapter, which is based on the experiences of the DIINAR.1 Arabic lexical resource and related software, and on that of the first version of the SYSTRAN Arabic-English MT system, outlines previous approaches of the computational morphology of the language (Section 2): root and pattern (shortly recalled); lexeme-based; machine learning and statistical; stems, based on roots and patterns, and finally, the stem-based approach, including root and pattern as well as grammar-lexis information. The latter, which is the most compliant to the requirements of machine-translation and other high-level applications, is further developed in Section 3. Authors go on presenting the structure of the Arabic word-form and a mapping of rules and relations accounting for grammar-lexis relations operating within the boundaries of that complex unit. In the Word-Formatives Grammar, rules and relations involving the lexical nucleus of the word-form play a crucial part and are formalised in a computational perspective. The stem either coincides with, or is the core of the nucleus, because lexical entries include two overall categories: in the first, stem and entry coincide; in the second, the lexical entry corresponds to a morphological compound encompassing the stem and a lexicalized extension (in most cases, a suffix which is part of the entry). Correct relations between the lexical nucleus and the other formatives included in the word-form are ensured through morphosyntactic specifiers associated to each entry of the lexical database. These relations, which have been included in the DIINAR.1 database, are both finite in number and exhaustive in coverage. They also allow computational morphology and other applications to rely on a good restriction of the generated lexica: only cliticized or affixed formatives that can effectively be associated with a given lexical nucleus are added, and 'illegal' ones are ruled out. In the DIINAR.1 resource, the effective number of inflected word-forms is 7,774,938 (about nine times less than one would obtain through 'blind' generation). A comprehensive mapping of examples is given. Their compatibility with applications going beyond computational morphology is also outlined

A. Soudi, A. van den Bosch and G. Neumann (eds.), Arabic Computational Morphology, 115–140.

7.1 Introduction

The present chapter is fundamentally concerned with the role, which will be shown to be crucial, of grammar-lexis relations in the computational morphology of the written form of Modern Standard Arabic (henceforth 'Arabic'). Computational morphology is the component of the linguistic engineering of the language that deals with the analysis and/or generation of the grammatical and lexical morphemes encompassed in the boundaries of the word-form, the structure of which proves, in Arabic, to be that of a complex unit. The contents of this contribution are based on two experiences in Arabic NLP development, that of the DIINAR.1 Arabic lexical resource and related analyzers and software, and that of the lexical database and analyzers embedded in the SYSTRAN Arabic-English machine translation system.

DIINAR.1 (*DIctionnaire INformatisé de l'ARabe, version 1*), Arabic acronym ***MaçaAliy*** معالي **(***Muçjam Al-çarabiy~aħ Al-Āliy~, /Muçjam al-çarabiyya(t) al-'āliyy/* – معجم العربية الآلي)[1], is a comprehensive Arabic lexical resource of around 120.000 lemma-entries operating at word-form level. It has been completed in close cooperation by IRSIT in Tunis (A. Braham and S. Ghazali), and in France, by the Lumière-Lyon 2 University (J. Dichy) and ENSSIB (M. Hassoun). Main related software are the word-form (or morphological) analyzer developed by M. Ghenima (1998), which was followed by R. Ouersighni's AraParse syntactic analyzer (2001, 2002), R. Zaafrani's Al-Muçal~im (المعلّم) Computer-aided learning system (2002) and R. Abbès's AraMorph morphological analyzer and AraConc concordance software (2004) – all of which have been devised to support the analysis of unvowelled Arabic script, and the generation, when needed, of fully vowelled written words-forms.[2]

The **SYSTRAN'**s Arabic-English MT system is a fully automatic transfer system. A first version has been developed at SYSTRAN's offices in San Diego and Paris between 2002 and 2004 by a team of computational linguists and lexicographers including Jean Sénellart, Ali Farghaly, Dina Abu Qaoud, Mats Attnas and Sylvie Poirier.

Both experiences show that grammar-lexis relations are associated to actual lexical entries, and can, subsequently, only be implemented in a stem-based lexical resource (including root and pattern information), as opposed to a resource founded on pure root-and-pattern combination (Dichy & Farghaly, 2003).

---

[1] Whenever needed, simplified and more traditionally phonological transcription has been added between slashes (//) to the very comprehensive and in many cases original transliteration system reflecting Arabic script introduced in the present volume.

[2] See Dichy, Braham, Ghazali & Hassoun (2002); Abbès, Dichy & Hassoun (2004). Availability: through ELDA, European Evaluation and Language Resources Distribution Agency, 55, rue Brillat-Savarin, 75013 Paris – www.elda.org . Contact: joseph.dichy@univ-lyon2.fr

Section 7.2 begins with a short survey of different approaches to the treatment of Arabic morphology, presenting them from both theoretical perspectives and from a computational viewpoint.

Authors go on (Section 7.3) to present a typology of grammar-lexis relations, which are formalised in a computational perspective. They recall the structure of the word-form in Arabic, focusing on the far less familiar fact that two fields can be distinguished within that unit (presented in Figure 7.2):

[a] the lexical stem, or nucleus formative (NF) – except in word-forms that only include grammatical morphemes –, and
[b] extension formatives (EF-s), which are bound grammatical morphemes.

Rules and relations involved in what can be called a Word-Formatives Grammar (WFG) belong to three general types: [a] NF ↔ EF and [b] EF ↔ EF rules and relations, to which [c] NF_a – NF_b morphological derivation links must be added. Rules and/or relations are typified and exemplified, with the purpose of presenting a mapping of grammar-lexis relations at stake in the computational morphology of Arabic.

7.2 Arabic Morphology: Theoretical and Computational Perspectives

The first module of a lexical resource is based on morphological description of the well formed internal structure of morphemes and words in the language in consideration. Grammar-lexis relations are thus dependent on what constitutes the basic units in the morphology, and how these units interact with other morphological entities to form higher and more complex word-form and syntactic structures. In this section, we give a brief account of Arabic morphology, recalling, from both theoretical and computational perspectives, some of the approaches that have dealt with the complexity of that component of the language.

Arabic morphology received a lot of attention from engineers and computational linguists particularly in the early eighties. Pioneering work on the computational morphology of Arabic goes back to the 1970s (Hlal, 1979, 1985a). The retrieval of the consonantal root from fully inflected words represented a challenge both from a theoretical point of view (Farghaly, 1987, 1994; McCarthy, 1981) and from a computational perspective that has, under certain conditions, proved liable to bring forth crucial theoretical advances and a better coverage of linguistic data, which we will endeavour to illustrate.

7.2.1 Arabic Morphology from a Theoretical Perspective

The notion of the morpheme as a meaningful string of segments delimited by the morpheme boundary symbol "+", and containing no internal morpheme boundary,

is challenged by the facts of Arabic morphology, which exhibits properties that can be recalled, in very short words, as follows:

– Roots are, strictly speaking, built of three or four consonants. Each root dominates a clustering of Arabic lexical morphemes around a semantic field, which can be single, subdivided or multiple.
– Certain changes in nouns, verbs or adjectives based on these consonantal roots yield derivatives. Some vowel and syllabic patterns seem, subsequently, to be associated with a constant set of meanings.
– Traditional treatment of Arabic morphology – especially in computational morphology – sometimes remains taxonomic, abstracting away from the particular root and citing or generating all possible patterns.

These questions have been presented in the preceding chapters. Let us nevertheless recall a few points. McCarthy (1981) revisited the view according to which an Arabic verb of form I, for example, is better analysed as consisting of two separate linguistic units: a consonantal root and a vocalism (Cantineau, 1950a, 1950b). He proposes that each should be assigned to a different tier. Together, they make a prosodic template. McCarthy also mentions the fact that there are certain constraints that apply to the root: some Arabic roots, for instance, may reduplicate the second radical as in *šad~a,* شدَّ 'to pull' and *haz~a,* هزَّ 'to shake', but never the first.[3] (Such facts have been described at length in medieval Arabic linguistic treatises.) Founding their discussion on the facts of Arabic morphology and other languages, McCarthy and Prince (1996) argue that a templatic morphology based on prosodic theory can better accommodate the properties of the non-concatenative nature of Arabic morphology and that of some other languages. Farghaly (1994) suggests that the Arabic lexicon may consist of underspecified entries to represent the discontinuous nature of Arabic morphemes. Farghaly (1987) argues that an adequate description of Arabic morphology has to recognize three levels: (a) that of the root, which is neither pronounceable nor belongs to any grammatical category, (b) that of the stem, which is pronounceable and has to be a member of the word classes of the language, and (c) that of the inflected word, where inflectional affixes are attached observing well defined rules to form the majority of actual Arabic words.

In the same period, many crucial theoretical and descriptive developments founded on other approaches occurred in the computational morphology of the language.

[3] See, for instance, *Al*-suyuwTiy~ (d. around 1505), *Al-muzhir* (المزهر للسيوطي) – a medieval linguistic treatise known by most readers with general knowledge in Arabic grammar or linguistics, the title of which cannot be relevantly translated.

7.2.2 Arabic Morphology in a Computational and Theoretical Perspective

The fact that Arabic word formation involves not only attaching prefixes and suffixes to stems, but also a large number of infixes with many morphophonemic processes, makes recovering the root and analyzing the internal structure of Arabic words a real challenge for both computer processing and linguistic theory and description. Linguists, engineers and computational linguists took up the task of the analysis and/or generation of Arabic words early on. In this section we present a brief description of the main approaches in the treatment of Arabic morphology.

7.2.2.1 The Root and Pattern Approach

The 'root and pattern' approach has already been presented in preceding chapters, and also in Dichy and Farghaly (2003). We will therefore focus very shortly, in this subsection, on historical aspects. The 'discovery' of consonantal Semitic roots by Western Semiticists goes back to the XVIIIth-XIXth cent. French traveller and Orientalist Constantin Volney (Rousseau, 1987). Linguistic discussion of the question including many references, can be found, for Arabic in Dichy (1990, 1993), and in Cassuto & Larcher (2000) for Semitic studies in general.[4] The partly traditional notions of 'root' and 'pattern' should by no means be abandoned, but they need to be limited and submitted to the constraints of formal definition (the set of which is proposed in (Dichy, 2003)). Decisive psycho-cognitive evidence has been given on roots and patterns in Hebrew (Bentin & Frost, 1995; Frost, Forster & Deutsch, 1997), and on the role of roots in the recognition of Arabic written words (Grainger et al., 2003). In the second half of the XXth century (on the whole, after Cantineau (1950a, 1950b)), most linguists and grammarians of Arabic and akin Semitic languages – in the West and in Arab countries alike – came to regard consonantal roots and patterns as basic linguistic components of the morphological description of the languages in consideration. Most researchers and linguistic engineers posited patterns, which are called in Arabic *Âawzaan /'awzān/* أوزان (originally: 'weight, measure, balance, poetic meter'), as presenting formal definitions of well formed Arabic words. These patterns were – and in many projects still are – considered as applicable to any root to generate Arabic lexical entries. D. Cohen (1961) gave a very elegant formulation of this view, which has later been described as a 'neo-Leibnitzian myth' (Dichy, 1993). It is nevertheless crucial to note that some of the forms which could be generated by patterns may have never existed in the Arabic language, and represent, as such, lexical gaps in the Arabic lexicon, which can be used to coin new words as needed, instead of

[4] On roots and patterns in the medieval Hebraic tradition, see, among many others, Zwiep (1996), in modern Hebrew dictionaries, Cassuto (2000); in the Arabic tradition, Troupeau (1984); also: Roman (1999), pp. 198–205 – "Brève histoire de la langue arabe"), which includes a strong refutation of the conjecture on roots as non-ordered consonant triples formulated by Ibn Jinniyy (IVth/Xth century), or as bi-consonantal 'roots' or 'etymons' constituted of non-ordered pairs taken up in the XXth century by G. Bohas.

borrowing foreign words that may violate the morphological and/or phonological rules of Arabic (Fassi Fehri, 1997).

The pioneering work of D. Cohen (1961) introduced a sophisticated representation of Arabic word-form structure, some revisited essentials of which are still in force to-day (Section 7.3 below). Hlal (1979, 1985a), Geith and El Saadany (1987) and many others designed computer systems for the analysis and/or generation of Arabic words relying heavily on the traditional description of Arabic morphology in terms of roots and patterns. The main approach, which has been followed with some variations, was to compile a dictionary of Arabic roots and dictionaries of affixes while maintaining a distinction between prefixes, suffixes and infixes, or to build lexicons of roots and patterns, to which lists of pre- or suffixed elements were added. Continuous look up of elements that could belong to any of these classes is then supposed to yield an analysis of Arabic word-forms.

7.2.2.2 The Lexeme-based Approach

Soudi et al. (2001) propose adopting a lexeme-based morphology, and describe MORPHE, which is a morphological rule compiler for implementation. The lexeme is an abstract concept representing a lexical meaning. Word-forms that share the same lexical meaning are related to a lexeme as members. For example, WORK is a verbal lexeme that includes as members: *work*, *works*, *worked* and *working*. All four word-forms share the lexical meaning ('working'). The variations among them are grammatical, such as past tense versus non past, etc., but not lexical.

An interesting question is: where does the Arabic root fit in a lexeme-based theory? Can we regard the root as a lexeme? The root represents a broad semantic field. In a Lexeme-based model (Aronoff, 1994) all the word forms of a lexeme belong to the same word class whereas the words generated by a particular Arabic root belong to various word classes. Clearly, two different verbs like *kataba,* كتب 'to write' and *Ăiktataba /'iktataba/,* اكتتب 'to enter one's name, to subscribe, to contribute, to invest in' respectively belong to two different lexemes although they are clearly related to the same root. This root also includes the verb *Ăistaktaba /'istktaba/,* استكتب 'to get someone to write', 'to dictate to someone', which partly shares the same meaning as *kataba*, but has a different argument structure. This implies that roots should not be regarded as lexemes à la Aronoff, which raises the question of what is exactly a 'lexeme' in Arabic. One possible answer (Soudi et al., 2001) is that it can, as is the case in English, be defined as an abstract concept covering all the different grammatical forms of a given stem. Thus *katabnaA,* كتبنا 'we wrote' – *yaktubu,* يكتب 'he writes' – *sayaktubu,* سيكتب 'he will write', etc., respectively belong to one lexeme since they all share the same lexical meaning and they only vary in tense, which is a grammatical feature. This otherwise efficient lemmatization procedure nevertheless leaves unanswered the question of the grouping of lexemes sharing a same root in a 'morphological family', or the issue of the derivational role of patterns, as well as pattern-to-pattern derivational links, within a given root. Such a grouping of lexemes had already been outlined

within a given root. Such a grouping of lexemes had already been outlined in Hassoun (1987) and further described in Dichy and Hassoun (1989).

7.2.2.3 The Machine Learning and Statistical Approach

Machine learning approaches to building NLP systems have become very popular in recent years. While rule-based NLP systems are usually time-consuming and require solid linguistic expertise, machine learning techniques are deemed to be fast, inexpensive, and requiring only large corpora. Surprisingly, machine learning techniques produce impressive results in a very short time and without the need for expensive linguistic knowledge (Forster et al., 2003) – although doubts could be raised in the case of languages for which heavy rule-based computational linguistic work has been conducted prior to the use of statistic-based methods. The fact is, one does not witness purely statistical systems, but rather mixed statistical and rule-based approaches (such as Dien et al., 2003). As has been mentioned in the final discussions of the IXth Machine Translation summit conference (New-Orleans, Sept. 2003), purely statistical methods may not at all yield the same results in less studied languages.

For languages like Arabic where solid computational linguistic knowledge and elaborate language resources (lexica, annotated corpora, tree-banks, etc.) are still rare (Nikkhou, 2004), statistical approaches nevertheless came to the rescue when national security needed to deal with millions of documents in Arabic, and little R&D funds, compared with 'big' languages such as English, Spanish, French or German. The underlying assumption here is that linguistic knowledge is present in linguistic data and that machine learning techniques can extract this knowledge going through cycles of training and retraining until it 'learns' the language. This assumption is immediately limited by the fact that researchers usually mention the existence of supervised learning modes, where the training data are annotated, thus facilitating the learning process (for instance, Schafer & Yarowsky (2003)). Effective annotation of corpora needs to be based on heavy previous linguistic development, traditional grammar being, for such a purpose, very far from being state-of-the-art, especially in Arabic, where traditional medieval grammar has not been sufficiently revisited in the light of modern linguistics. One can nevertheless mention unsupervised learning techniques, which are very important when annotated corpora of the language are unavailable.

The recent availability of parallel corpora for Arabic-English prompted many researchers to use machine learning techniques to salvage all kinds of linguistic information. For example, Diab and Resnik (2001) describe how they used a parallel corpus for word sense disambiguation under the assumption that different meanings of the same word in the source language will be translated into distinct words in the target language.

Rogati et al. (2003) followed the unsupervised learning approach for developing a prototype of a non-English Arabic stemmer. Their objective is to build a language-independent stemmer. The model they use is based on statistical machine

translation using an English stemmer and a small parallel corpus for training purposes. Their main approach is to remove prefixes and suffixes from Arabic words. Although they did not remove infixes nor deal with morpho-phonological transformation, they report achieving an improvement of 22–38% on average over unstemmed text and 96% of the performance of a proprietary stemmer built using rules, affix lists and human annotation.

7.2.2.4 The Stem-based Approach

The above approaches aim at analysing and/or generating Arabic word-forms. The problem in any Arabic NLP system, such as tagging, document categorization, automated summarization, speech recognition, machine translation, etc., is that it is not enough just to recognize or generate forms. In NLP programs related to effective application results, important information needs to be associated with each morpheme and lexical entry. There is information coming from the morphological level, such as gender, number, person, mode and tense, definiteness, part of speech (POS), etc. There are also syntactic features, such as Count/Mass nouns, subcategorization frames, what type of a subject or an object a verb takes, etc. One will also have to add semantic information, such as categorizing a noun as referring to human/non human, animate/inanimate entities, or to place and/or time, etc.

The more elaborate the information associated with the lexical entry, the more sophisticated the grammar becomes, and the more powerful the NLP system as a whole turns out to be. In machine translation applications, for instance, such sophisticated linguistic information cannot be done without. It can, on the other hand, never be associated with an Arabic root or with a pattern, because neither Arabic root nor pattern belongs to word classes (the terms refer to linguistic abstraction, not to actual parts of speech). However, combined roots and patterns may form the nucleus of nouns, verbs and adjectives. The linguistic information under consideration, including indispensable grammar-lexis relations for Arabic NLP applications, can only be associated to stems, since a stem, by definition, belongs to a syntactic category, and never to roots or to a mere combination or root and pattern (Dichy & Farghaly, 2003). In the case of a homograph (a very frequent case in standard vowel-free Arabic writing), a given stem could belong to several syntactic categories.

The stem-based approach to Arabic morphology reduces the complexity of Arabic word structure, eliminates large numbers of lexical gaps, and makes it possible to associate relevant and specific morphological, syntactic and semantic features with each entry. Figure 7.1 shows a small subset of the morphological information associated with the lexical entry of an Arabic verb in the SYSTRAN monolingual Arabic dictionary, built as a component of the Arabic-English translator.

7.2.2.5 Stems, Based on Roots and Patterns

One of the most advanced treatments of Arabic morphology using both the root-and-pattern and stem approaches is the work done at Xerox Research Centre in

France by K. Beesley and his colleagues (Beesley, 2001). Beesley's implementation is based on the insights of Karttunen (1994) that morphotactics and variations in morphology can be expressed in regular expressions and can then be compiled into finite state automata which are very efficient, fast and bidirectional. It elaborates on Kimmo Koskenniemi's two-level morphology (Beesley & Karttunen, 2003; Karttunen & Beesley, 2005), on the basis of a partial revisiting of McCarthy's representation (Beesley, 1989), and integrates a first version of Tim Buckwalter's lexicon (Buckwalter, 2002).

The approach should therefore not be described as founded on mere root and pattern combination: in fact, it includes as an essential step the checking of candidate entries generated from root and pattern 'merging', and pre- or suffixes combination, against existing lexical entries, as attested by a reference dictionary such as Hans Wehr (1979), and fully takes into consideration the complexity of Arabic morphology. The Xerox Arabic Lexicon included, four years ago (Beesley, 2001): 4,930 roots, 400 patterns, and 90,000 stems based on roots and patterns. The latter correspond to 70,000 root-pattern intersections on the lower side of the two-level morphological representation, the differences depending on information associated to stems on the higher side (see Beesley & Karttunen (2003)), which clearly shows that, in this approach, morphological and word-form grammar-lexis information is associated to stems. The figure of 90,000 stems also shows that the blind combination of roots and patterns (4,930 roots x 400 patterns = 1,972,000 root-pattern virtual links) has been severely restricted.

[perfect=زَرَع],[imperfect=يَزْرَع],[imperative=اِزْرَع],
[passperf=زُرِع],[passimperf=يُزْرَع]

Fig. 7.1. A sample of SYSTRAN's monolingual dictionary entry of the Arabic verb *zaraça* زرع 'to plant'

7.2.2.6 Stem-based Lexical Resources, Including Root-Pattern and Grammar-lexis Information

Another advanced treatment of Arabic morphology was initiated in France in the early 1980s, in what was known as the SAMIA project[5] (Desclés et al., 1983; Dichy, 1984, 1987; Dichy & Hassoun, 1989; Hassoun, 1987), and has been going on since. It has led to the completion, in collaboration with a Tunisian research centre (IRSIT, now IT.COM), of the DIINAR.1 Arabic lexical resource. Morphological analyzers drawing on this resource have been completed on a parallel basis. The approach can be described as deliberately stem-based, including root and pattern information on the one hand, and a comprehensive coverage of word-form grammar-lexis relations on the other. This makes it closer to the requirements of

[5] SAMIA is the acronym for "Synthèse et Analyse Morphosyntaxiques Informatisées de l'Arabe".

machine translation, such as has been illustrated, in many couples of languages, by the SYSTRAN engines.

In this approach, representations of word-form structures and of word-level grammar-lexis relations are very explicit. This is due to the database structure of DIINAR.1, the subsequent declarative programming of the associated software (Abbès, 2004; Ghenima, 1998; Ouersighni, 2001; Zaafrani, 2002), and also, to the comprehensiveness of the coverage of Arabic morpho-lexical data.

The concepts and methods at stake in that representation of Arabic computational morphology, which is centred on grammar-lexis relations, are presented in some detail in the following section.

7.3 Mapping the Arabic Lexicon: Word-form Structure, Rules and Grammar-lexis Relations in Arabic

7.3.1 Structure of the Word-form in Arabic (Short Recall)

Word-form units feature in Arabic a complex, albeit very regular, structure. Standard word-forms comprise one lexical nucleus and one only.[6] On the right and left sides of that nucleus, specified sets of bound morphemes can be found, in either affixed or cliticized position (Cohen, 1961; Desclés, 1983; Dichy, 1990, 1997; Dichy & Hassoun, 1989).

The structure encompasses:[7]

- **proclitics (PCL),** which consist of mono-consonantal conjunctions (such as *wa-*, وَ- 'and' , *li-*, لِـ 'in order to'), prepositions (i.e. *bi-*, بِـ 'in', 'at' or 'by', *li-*, لِـ "for"), the pre-verb *sa-*, سَـ (indicating the future), the article *Al- /'al/*, الـ etc.;
- a **prefix (PRF)**. The category, after D. Cohen's representation of the word-form (1961), only includes the prefixes of the imperfective, such as *Âa- /'a/*, أ-, morpheme of the 1st person sing., etc.;
- a **stem.** Stems are divided in two general categories (Dichy, 1984):
  - **Type 1 stems**: this first subset consists of major lexical categories that are liable be represented in terms of a PATTERN and of a 3-consonant or 4-consonant ROOT. (Major lexical categories encompass nouns, adjectives,

[6] Poly-lexical entries are, in Arabic, either composed of more than one word-form (e.g. *Al-quruwn Al-wusTý, /'al-qurūn al-wusTā/* القرون الوسطى 'the Middle Ages') or reduced by the morphological system of the language to a mono-lexical unit, e.g. *qarwasaTiy~,* قروسطي 'medieval'. The meaning of the Arabic lexicographical term *naHt,* نحت 'coinage', which describes the phenomenon, refers to the above reduction, which brings the compound to comply with (a) the model of 3-consonant or 4-cons. roots, and (b) the structure of the mono-lexical word-form (Dichy, 2003).

[7] Hebrew word-forms feature similar complex structures (Sampson, 1985), pp. 90–91; for a psycholinguistic approach, see Frost, Deutsch & Forster (2000).

verbs and deverbals.)[8] By convention, the terms 'root' and 'pattern' will be henceforth presented in small capital letters, referring to the formal definition above (Dichy, 2003). A **ROOT** is an ordered triple of consonants (3-C) or, by extension of the system, a quadruple (4-C).[9] A **PATTERN** is, in short words, a template of syllables, the consonants of which are that of the 3-C or 4-C ROOT, with the addition of mono-consonantal affixes (belonging to mono-consonantal roots[10]), such as the *t* 'echo-morpheme' (Roman, 1990). Consider for instance the stem *takab~ar,* تكبَّر 'to be haughty'. This stem can be analysed into the 3-C ROOT /k-b-r/, and the PATTERN /taR^{1}aR^{2}R^{2}aR^3/ (تَفَعَّل), which includes the mono-consonantal root /t/ and the 1st, 2nd and 3rd consonant of the 3-C ROOT (respectively: R^1 = *k*, R^2 = *b*, R^3 = *r*). It is crucial to remember that type 1 stems include *all* the verbs and deverbals of the language (Dichy, 1984).

- **Type 2 stems:** the second subset of stems contains only nouns that cannot be represented in terms of PATTERN and ROOT, such as: */'ismāçīl/,* إسماعيل 'Ishmael', *fiyziyaA',* فيزياء 'Physics'. There are *no verbs* in this category of stems (a corollary of the fact, which has just been mentioned, that all verbs belong to type 1 stems);

– **suffixes (SUF)**, such as verbal inflexions, nominal cases, the nominal feminine ending +*aħ, /a(t)/* +ـَة, etc.;

– **enclitics (ECL)**. In Arabic, enclitics are complement pronouns. Some verbs can have a double ECL, for example: *çal~am+tum-uw-niy-haA, /çallam+tum-ū-nī-hā/,* علّمتمونيها "you [plur. masc.] have taught-me-it" (this /ū/ sequence

---

[8] The term '**deverbal'** refers to what could also be called 'verbal-nominal forms', i.e., nominal forms that include syntactic-semantic verbal features, such as transitivity, etc. These are, in Arabic, the infinitive form, *maSdar,* مصدر, the active participle, *Ăism Al-faAçil, /ism al-fā'il/,* اسم الفاعل, and the passive participle, *Ăism Al-mafçuwl, /ism al-maf'ūl/,* اسم المفعول. Note that other subcategories have been included in deverbals in the DIINAR.1 lexical resource (see § 7.3.3.2, Figure 7.4), following the categorisation of traditional Arabic grammar. This has proven not to be consistent beyond morphological analysis. Concerning the three subcategories above, research conducted in the DIINAR.1 project has shown that traditional Arabic grammatical terminology obscures the fact that the forms in consideration are liable to be *either* deverbals, *or* nouns. Consider for instance the sentence *ÂanaA saAkin fiy ruwmaA, /'anā sākin fī rūmā/* أنا ساكن في روما, 'I'm living in Rome. The active participle *saAkin* (ساكن) admits suffixed plural forms, e.g., *naHnu saAkinuwn* (masc. ساكنون) or *saAkinaAt* (fem. ساكنات) *fiy ruwmaA*, but excludes the 'broken plural' form *suk~aAn,* سكّان which refers to the meaning of 'inhabitant(s). The former is a deverbal, the latter (*saAkin*, plur. *suk~aAn,* ساكن, ج سكّان) has undergone a nominalization process, i.e. has left the deverbal category to become a 'purely' nominal lexical entry (see § 7.3.5.1[b]). Such cases require two distinct entries, each associated with its own grammar-lexis specifiers.

[9] 5-consonant so-called 'roots', included for instance in *Ăiçranfaza, /' içranfaza/,* اعرنفز 'to almost die from cold', which can only be found in ancient poetry or medieval dictionaries, have been neglected.

[10] Mono-consonantal roots in Arabic and Semitic languages have been disclosed by Roman (1990, 1999).

only appears with the plural masculine form of the subject pronoun when an ECL pronoun is attached).

Figure 7.2 (Dichy, 1997) illustrates this structure, in the case of Type 1 stems (conventions in the lower part are explained immediately after):

Conventions (not already encountered here): ## = 'word boundary'; # = 'clitic boundary'; + = 'pre- or suffix boundary'. NF = 'nucleus formative' (referring to the lexical nucleus of the word-form); EF = 'extension formative' (referring to

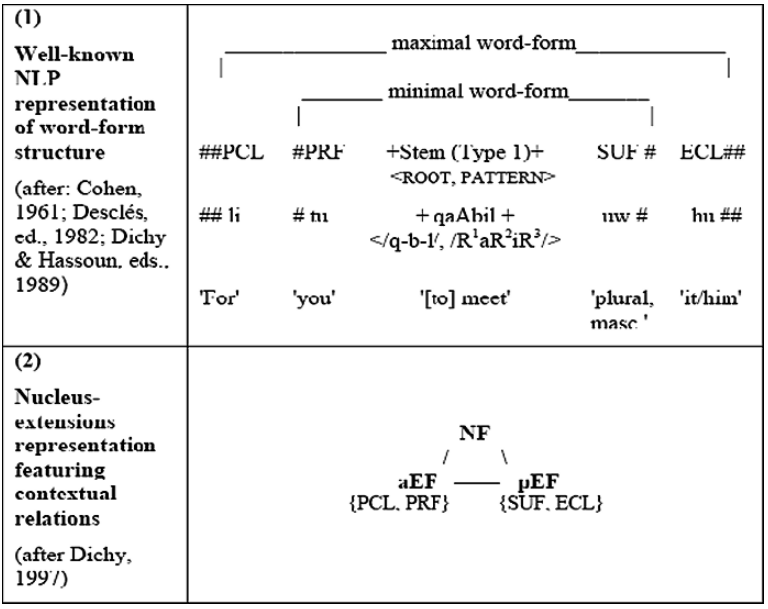

Fig. 7.2. Arabic word-form structure (with ROOT-and-PATTERN stems) – **لتقابلوه**

bound grammatical morphemes); aEF, pEF = 'ante-positioned' or 'post-positioned' EF. The set of aEF-s includes {PCL, PRF}, that of pEF-s comprises {SUF, ECL}.

7.3.2 Word Formatives, Word Specifiers and Word Formatives Grammar

The **Word Formatives Grammar (WFG)** accounts for the rules and relations that ensure correct combination of formatives within the boundaries of the word-form (Dichy, 1987). This grammar includes morpho-phonological transformation rules, and various contextual rules, which will be outlined below (Subsection 7.3.4). Phonological transformations were not accounted for in the morphological analysis of vowelled Arabic words initiated by Cohen (1961) or Hlal (1979, 1985a). They have on the other hand been included in the approach developed in the SAMIA project for the analysis or the generation of vowel-free word-forms,

and the subsequent elaboration of the DIINAR.1 lexical database. One of the postulates of this approach is that linguistic formatives must be specified in terms of morpo-syntactic rules and relations according to the syntagmatic extension of the unit they are inserted in (Dichy, 1987, 1997). Owing to the structure of the Arabic word-form, one is brought to give special attention to rules and grammar-lexis relations, accounting, in short, for insertion rules operating within the scope of that syntagmatic unit. The following concepts and conventions have been subsequently adopted:

– *Word formatives*, i.e., morphemes considered in the frame of the word-form structure, are associated with grammar-lexis **word-specifiers (w-specifiers)**.
– *Sentence formatives* need to be associated with **s-specifiers**, and *text formatives*, with **t-specifiers**.

This can be considered as an overall framework. It could easily be shown that the three types of specifiers above involve different types of phenomena (Dichy, 2005). Specifiers involving word and sentence formatives can be described as **morphosyntactic specifiers**.

7.3.3 Grammar-Lexis Relations in the Processing of Written Arabic

7.3.3.1 Multiple Analyses at Word-form and Sentence Level

Let us recall that, in the morphological analysis of Arabic, the complex operation referred to as the 'segmentation' of the word-form into morphemes (or formatives) is rendered the more difficult because standard writing is 'unvowelled' or 'vowel-free', i.e. bare of *secondary diacritical signs* indicating short vowels (*HarakaAt,* حركات), consonant doubling (*šad~a,* شدّة), diacritical case-endings (*tanwiyn,* تنوين), etc. This has been presented in previous chapters. It has been shown in some details, quite a few years ago (Desclés, 1983; Dichy, 1984), that the resulting homographs entail a high number of potential *existing* analyses for a relevant percentage of word-forms (Abbès, 2004; Ghenima, 1998).[11] This is indeed the case because computational morphology, when it is not included in a syntactic analyzer (Ditters, 1992; Ouersighni, 2001, 2002), deals with word-forms context-free.

Ambiguity due to multiple analyses should subsequently *not* be considered a problem in itself: morphological and morpho-syntactic analyzers aim at assigning word-forms with *all* the analyses that comply with the rules and lexicon of the language, and them only. It is on the other hand necessary to restrict the combination of word-formatives to forms that are 'legal' according to the morpho-syntactic system (including the morphotactics of the writing system) and the lexicon of the language. This is also required to prevent the number of analyses per word-form to climb much higher than allowed by the language and its writing

[11] This could also be tested, in addition to the morphological analyzers drawing on the DIINAR.1 resource (Abbès, 2004; Ghenima, 1998; Zaafrani, 2002), with the morphological analyzer put on the Internet by the Xerox European Research Centre (Beesley, 2001).

system. Non existing analyses should therefore be ruled out: the vowel-free word-form *'çlnt,* أعلنت, for instance, should not be analysed as **Âaçluntu, */'açluntu/,* أعلُنْتُ or **Âaçlunat, */'açlunat/,* أعلُنَتْ (no meaning in both cases), but – among other forms – as *Âaçlantu, /'açlantu/,* أعلَنْتُ or *Âaçlanat*, */'açlanat/,* أعلَنَتْ 'I' or 'she declared publicly'.

7.3.3.2 Restricting Generated Lexica

Ruling out what could be described as 'morphological noises' is an equally crucial issue when it comes to restricting the number of entries of a lexical resource.

Let us consider a few figures:

1) In the DIINAR.1 lexical resource, the number of combined proclitics and suffixes which are effectively in use in Modern Standard Arabic, and that of prefixes and enclitics is shown below (Abbès, Dichy & Hassoun, 2004):

Comments:

(a) Prefixes do not combine (see § 7.3.1 above).

(b) Enclitics may combine in doubly transitive verbs, which seldom occurs in present-day Arabic, where one of the complements is usually preceded by a preposition; e.g.: Ancient Arabic *manaH+tu-ka-hu,* منحتكه 'I have given_you_it' is currently realized as *manaH+tu-hu la-ka,* منحته لك 'I have given_it to_you'.

Proclitics (combined)	64
Prefixes	8
Suffixes (combined)	67
Enclitics	13

Fig. 7.3. Number of EF-s in DIINAR.1

(c) In the above numbers, extension formatives (EF-s) combinations have been restricted to effective use. Ancient Arabic proclitic combinations, such as *Âa-fa-bi-ka-Al-, /'a-fa-bi-ka-'l/* أ/ف/بـ/كـ/الـ 'interrogative-then-by-such_as-the (generic article)', or *bi-ka*, 'by-such_as', and a few others, have not been included.

(d) In Figure 7.3, suffixes that only include secondary diacritics, (traditionally called 'vowel-signs', *HarakaAt,* حركات), i.e. basic case-endings in nouns and a subset of mode markers in verbs, have not been included. This ensures a 'lower-hypothesis' interpretation of the calculations presented in the demonstration below.

2) The number of lemma-entries belonging to major lexical categories is the following:

Comments:
In the DIINAR.1 lexical resource, following traditional Arabic grammar, adjectives have been included in nominal stems, and two morphological subcategories have been added to deverbals. These are, as shown in Figure 7.4: (a) 'analogous adjectives' (صفات مشبهة) and (b) 'nouns of time and place' (أسماء المكان والزمان). Both categorisations, which remained acceptable in the context of computational morphology, have proved inconsistent when extending grammar-lexis relations to syntactic features. Clearly, (a) are adjectives and (b) are nouns. This bias can be corrected in the related analyzers and generators, through modifying the (sub)category in specifiers associated with lexical entries.

Let us now undergo a bit of *ab absurdo* reasoning:

On the basis of Figure 7.3, blind combination of bound grammatical morphemes would give:

$$64 \times 8 \times 67 \times 13 = 445{,}952 \text{ 'virtual' extension formatives (EF-s).}$$

Nouns, including adjectives	29,534
[Broken plural nominal forms (جموع التكسير), included in the number of nouns above]	[9,565]
Proper names (limited prototype) (أسماء الأعلام)	1,384
Verbs	19,457
Deverbals (مشتقات اسمية) :	
- infinitive forms (مصادر)	23,274
- active participles (أسماء الفاعل)	17,904
- passive participles (أسماء المفعول)	13,373
- 'analogous adjectives' (صفات مشبهة)	5,781
- 'nouns of time and place' (أسماء المكان والزمان)	10,370
[Total number of deverbals]	[70,702]
Subtotal of lemma-entries	121,077

Fig. 7.4. Number of lemma-entries in the DIINAR.1 Lexical resource

Unconstrained combination with the total number of stems in Figure 7.4 leads to a generated lexicon of 'virtual' word-forms of:

$$445{,}952 \text{ EF-s} \times 129{,}258 \text{ stems} = 57{,}642{,}863{,}616 \text{ 'virtual' word-forms.}$$

Limiting the figures to inflected forms, the combination would still yield:

$$8 \text{ prefixes} \times 67 \text{ suffixes} \times 129{,}258 \text{ stems} = 69{,}282{,}288 \text{ 'virtual' forms.}$$

In the DIINAR.1 resource, the effective number of inflected word-forms is 7,774,938 (Abbès, Dichy and Hassoun, 2004, which includes a breakdown according to lexical categories), i.e. 11.22% of the 'virtual' figure above.

As for the lexical nuclei of word-forms, knowing that the amount of ROOTS in DIINAR.1 is 6,546, with an estimated number of 400 patterns, the figure would be:

6,546 ROOTS × 400 PATTERNS = 2,618,400 ROOT-PATTERN virtual links.

This comes against 119,693 lexical *existing* lemma-entries (and 129,258 existing stems including 'broken plurals', proper names being for obvious reasons left out). We have seen in § 7.2.2.5 that the Buckwalter-Xerox lexicon includes 90,000 entries, based on 4,930 ROOTS and 400 PATTERNS, the blind combination of which would have led to 1,972,000 ROOT-PATTERN virtual links.

Remarkably – knowing that sources and research contexts did in fact differ –, the ratios of overall entries per ROOT in the two lexical resources are next to identical:

– DIINAR.1 lexical database: 119,693 entries / 6,546 ROOTS = 18.28
– Buckwalter-Xerox lexicon: appr. 90,000 entries / 4,930 ROOTS = 18.25.

One is therefore brought to the conclusion that the 'virtual results' above are not only absurdly enormous, they are also blurred: for lack of explicit decision procedures, there would be no way in which a given analysed or generated 'form' could, or not, be deemed part of the language. Restricting generated lexica through rules involving grammar-lexis relations associated to actual lexical entries is therefore necessary both for computational generation and analysis, and in the building of efficient lexical resources. Let us now consider the general types of rules and relations that are valid within the boundaries of Arabic word-forms.

7.3.4 General Types of Rules and Relations in the Word-Formatives Grammar

7.3.4.1 The Three Types of Contextual Relations Involved in the WFG

Rules involving word-formatives (NF and EF-s) are based on three types of relations (Dichy, 1987): **⇒ 'entails', ≠> 'excludes', ** 'is compatible with'** or **'admits'**, the third of which is attached to the opposed pair of the first two as an 'elsewhere' relation of a special kind, directly connected to ambiguity in language analysis processes. In generation, all 'compatibility' (or 'admit') relations can in fact be rewritten in terms of either 'entail' or 'exclude' rules. 'Compatibility' relations are mostly useful in the formalization of recognition rules, when ambiguity is at stake. They express relations that only appear in analysis.[12]

It is essential to note that automatic analysis and generation of linguistic data are *not* to be considered as reverse processes (Desclés, 1983; Dichy, 1984, 1990, 1997).

---

[12] Developments on ambiguity in Arabic NLP have been presented in Dichy (1990, 2000). Statistics on ambiguity in 'unvowelled' written Arabic are given in Abbès (2004). For a general reference on ambiguity in Arabic, see Arar (2003), and A. Farghaly's contribution on "Lexical Ambiguity in Arabic Machine Translation Systems" in the same volume.

7.3.4.2 Rules Related to the Two General Fields of the Word-form Structure

Grammar-lexis relations are connected to the **word-Formatives Grammar (WFG)**, which accounts for the rules and relations involved in Arabic word-form structures (Dichy, 1987, 1990). In the well-known representation recalled in the upper half of Figure 7.2 (see § 7.3.1 above), two general types of word-formatives can be distinguished:

– Formatives pertaining to the bound grammatical morphemes of the language, and encompassed within the boundaries of the word-form, are called **EF-s (Extension Formatives)**.
– The lexical nucleus of the word form (except in word-forms that only include grammatical morphemes) is called a **NF (Nucleus Formative)**.

The WFG includes, accordingly, three types of rules and/or relations, which are directly attached to the triangle featured in the lower part of Figure 7.2. By convention, in 'PCL → SUF', the arrow '→' can be read either as 'determine' (either ⇒ 'entails', or ≠> 'excludes') or 'are compatible with' (**), with reference to the three types of rules mentioned in § 7.3.4.1. The two-headed arrow '↔' is read in the same way, with the addition of 'reciprocity' ('and vice-versa').

Types of rules and relations involved are:

[a] **EF ↔ EF contextual rules and relations**, such as PCL → SUF rules, e.g.:

PCL = Prep. {*bi#, li#*} ⇒ SUF = {*+i, +in,+a, +n, +iyna, +iy, +ayni, +ay*}

which can be phrased as: '*if* the proclitic is a preposition (i.e., a member of the set between braces), *it follows* (or: *this entails*) that the suffix is one of the indirect (or genitive, *majruwr* مجرور) case suffixes', the set of which is listed between braces. The selection of the correct case-ending in the list is performed through different types of morphological and syntactic rules.

[b] **NF ↔ EF rules and relations**. A simple example involves the major lexical category to which the NF belongs, such as PCL → NF category. The above rule, for instance, needs to be completed by the following:

PCL = Prep. ⇒ NF = Noun.

[c] **NF – NF relations** are morphological derivation links, which are, in a great number of cases, not rule-predictable. Consider, in nouns, singular – 'broken plural' links, for instance: sing. *kitaAb,* كِتاب – plur. *kutub,* كُتُب 'book(s) vs sing. *sinaAn*, سِنان – plur. *Âasin~aħ*, */'asinna(t)/,* أسِنَّة 'spearhead(s)', sing. *HimaAr,* حِمار – plur. *Hamiyr,* حَمير *Humur,* حُمُر and *ÂaHmiraħ, /'aHmira(t)/,* أحْمِرة 'donkey(s)'. In these nouns, the PATTERN of the singular is /$R^1iR^2āR^3$/, فِعال ; the PATTERNS of 'broken plurals' appear to be, sometimes /$R^1uR^2uR^3$/, فُعُل, sometimes /'a$R^1R^2iR^3$a(t)/, أفْعِلة, and sometimes another 'broken plural' form, including, in some cases, a suffixation plural form, e.g.: *qiTaAr,* قِطار 'train', shows two plural forms, *quTur,* قُطُر , which pertains to the

/R[1]uR[2]uR[3]/, فُعْل PATTERN of 'broken plurals', and *qiTaAraAt,* قِطارات, which is constructed with the suffix +*aAt,* ـَات.

7.3.5 Rules of the WFG and Grammar-Lexis Specifiers

Rules and relations involved in the Word Formatives Grammar entail the need, for the entries of a lexical database, to be associated with grammar-lexis relations. The latter are essentially attached, in computational morphology, to types [b] and [c] above. As mentioned above, they are called in the SAMIA-DIINAR.1 approach, word-specifiers (w-specifiers).

7.3.5.1 The Two Types of NF ↔ EF Contextual Rules and Relations

A crucial point in the overall structure of the Arabic lexicon, which seems to have been widely overlooked, appeared in the elaboration of the DIINAR.1 resource. Grammar-lexis relations concerned with the nucleus are liable to involve, in addition to compositional combinations of nucleus and extension formatives, non-compositional ones, i.e. 'frozen' or 'lexicalized' combinations (Dichy, 1984, 1990, 1997). Let us consider these two types:

[a] NF ↔ EF compositional relations, and related w-specifiers

Compositional NF ↔ EF combinations are, on the whole, easy to grasp, although they include a few tricky aspects (see § 7.3.5.3). A simple example is that of: Stem → SUF rules, e.g.:

$$\text{Stem} = \text{diptote} \Rightarrow \text{SUF} = \{u, a, i\}$$

This rule can be rephrased as: 'a stem whose declension is diptote (*mamnuwç mina AlS~arf, /mamnūç mina S-Sarf/,* ممنوع من الصرف) entails case-endings belonging to the listed set'. An additional rule restricts the occurrence of SUF /i/ with diptote nouns and adjectives to construct-state syntactic structures (*ĂiDaAfaћ, /'iDāfa(t)/*, إضافة), e.g.: *min maçaAlimi Al-çaASimaћ, /min maçālimi Al-çāSima(t)/,* من معالمِ العاصمة 'from the monuments of the capital'. Other suffixes are accounted for in different rules.

Many other examples could be given: nouns with 'broken plurals' often exclude (≠>) suffixed plural forms; intransitive verbs exclude ECL complement pronouns; verbs that only admit non-human complements exclude a subset of the ECL-s, such as *-hum*, 3rd person plur. masc. or *-ki*, 2nd person sing. fem., which can only refer to human entities.

[b] **NF ↔ EF non-compositional lexicalized relations, and related w-specifiers**

In Arabic, as in other Semitic languages of the same family, in addition to ROOT and PATTERN derivation, one finds lexical derivation by means, essentially,

of suffixation.[13] With Type 1 stems (§ 7.3.1) featuring ROOT and PATTERN combination, the two means of derivation are liable to add up. Letting aside, in the present contribution, *phrasal compound* expressions in which a given syntactic structure is frozen (as in *majmaç çilmiy~,* مجمع علمي 'Science Academy', *jamç Al-maçluwmaAt,* جمع المعلومات 'data gathering', or *ÂaHaATa fulaAnAã çilmAã bi-, /'aHāTa fulānan çilman bi-/,* أحاط فلانًا علمًا بـ 'to inform someone of'), one can distinguish between two types of lexical entries based on strict morphological means (Dichy, 1997):

[1] **'Simple' lexical entries** coincide with the stem (or nucleus), e.g.: *maktab,* مكتب "office", "bureau". The entry can, as expected, be inserted in a word-form, such as *wa-bi-maktab-i-naA,* وبمكتبنا "and by our office" ("and-by-office-genitive case /i/-of us").

[2] **Morphological compound entries** (as opposed to *phrasal compounds*) feature a 'lexical freezing' of the combination of a given nucleus with a given extension formative. The morphological compound is coded in the database as a full entry of its own (a w-specifier is, in addition associated with the stem, in order to account for occurrences that have not undergone a 'lexical freezing' of the NF – EF relation). The lexical entry *SuHuf-iy,* صُحُفي 'journalist', for instance, is a compound entry:

(a) it does not coincide with the stem, or nucleus, it encompasses. The stem is, here: *SuHuf,* صُحُف (otherwise meaning 'sheets', 'papers'), which features a combination of ROOT /S-H-f/ and PATTERN /R^1uR^2uR^3/ (فُعُل);

(b) it includes on the other hand the extension formative *+iy~* (*yaA' Al-n~isbaħ, /yā' al-nisba(t)/,* ياء النسبة 'relative adjective or noun' morpheme). An analogous example, going as far back as the IInd cent. of Hijra/VIIIth cent. c.e., is *kutub+iy~,* كُتُبيّ , 'librarian' (in the medieval meaning of the word).

– A given word can be either a frozen morphological compound, or result from composition: *jaAmiçaħ, /jāmiça(t)/,* جامعة can be analysed either as the morphological compound *jaAmiç+aħ, /jāmiç+a(t)/,* meaning 'mosque', which is linked in the lexicon with the 'broken plural' *jawaAmiç,* جوامع or as the feminine of the active participle *jaAmiç+aħ, /jāmiç+a(t)/,* meaning 'collecting', i.e. 'she who collects', or 'compiles' or 'brings together'. This is a very frequent phenomenon.

– Morphological compounds can, of course, also be inserted in a word-form, e.g. *wa-bi-SuHuf+iy~+i-naA,* وبصُحُفيِّنا 'and by our journalist' (= 'and_by_ journalist_genitive case /i/_of us').

[13] In this representation, affixed elements included in PATTERNS, such as /ma/ in *mawçid,* موعد 'promise', 'pledge' (or 'appointment', 'date') are *not* considered as prefixes or suffixes (following Cohen (1961) and Desclés (1983), as well as traditional Arabic grammar) – see § 7.3.1.

Another example is found in proper names, such as *(Al-Kuwayt,* الكويت in which the lexicalized EF is the proclitic article *Al-*.

The above distinction is methodologically crucial. It gives additional interpretative evidence to the demonstration presented in § 7.3.3.2, according to which derivation of 'virtual' lexical entries from ROOT and PATTERN is no sufficient basis for the Arabic lexicon. But the issue goes much further: ROOT and PATTERN derivation is complemented by 'external' derivation, i.e. by the lexicalization of the NF – EF combination, which is only found in nouns (Dichy, 1984), and cannot be predicted by rules, because the process described above only occurs to answer the need, for the lexicon of a given language, to build a new entry, when a newly encountered entity requires nomination. This is correlated to the non-compositional nature of the NF – EF relation. It follows, in a computational perspective, that morphological compounds can only be recognized or generated with the help of a lexical resource.

7.3.5.2 Morphological NF – NF Derivation Links, and Related W-Specifiers

Another type of relation is NF – NF linking combinations, also called, in Semitic studies, 'internal' derivation, because these links feature a variation in PATTERN, the ROOT remaining constant. Such derivations have to be encoded as w-specifiers, whenever the morphological link is not strictly rule-predictable, which occurs in a majority of stems (Dichy, 1987, 1990; Hassoun, 1987).

This is the case, for instance, in a wide number of singular ↔ 'broken plural' links in nouns or adjectives (exemplified in § 7.3.4.2[c]), as well as in most 'perfective' ↔ 'imperfective' links (*maADī, /mādin/* ↔ *muDaAriç, /muDāriç/,* ماض ↔ مضارع), in verbs of 'simple' PATTERNS (*Al-fiçl Al-mujar~ad,* الفعل المجرد). In the same subcategory of verbs, the verb ↔ infinitive form link (*fiçl* ↔ *maSdar,* فعل ↔ مصدر), and other similar ones (such as verb ↔ analogous adjective, *fiçl* ↔ *Sifaħ mušab~ahaħ, /Sifa(t) mušabbaha(t)/,* فعل ↔ صفة مشبّهة) also need to be encoded lexically.

Restrictions on conjugation paradigms, such as the passive or the imperative, which can be described in terms of semantic rules (based on features such as agentivity, and human/non human complements, etc., cf. Ammar & Dichy (1999), pp. 17, 19–20), also pertain to NF ↔ NF links. In a lexical resource, they need to be encoded as w-specifiers, because the entries are not actual linguistic signs (with 'signifiant' and 'signifié' features), but mere chains of characters associated with a set of linguistic specifiers (Dichy, 1997).

7.3.5.3 Morphological Derivation Links Including a Morphological Compound

In many cases, *morphological compound entries* also feature a *morphological derivation link*. For example: the morphological compound *madras+aħ, /madras+a(t)/,* مدرسة, 'school' is associated with the broken plural form *madaAris,* مدارس. By contrast, the analogous entry *maktab+aħ, /maktab+a(t)/,* مكتبة 'library',

has a suffixed plural form *maktaba+At,* مكتبات. This is due to the fact that the expected broken plural *makaAtib,* مكاتب is already associated in the lexicon with *maktab,* مكتب "office", "bureau" (which is a 'simple' lexical entry – see § 7.3.5.1[b] above).

7.4 Conclusion: Exhaustive Coverage of Morphological Features, and the Question of what Lays Beyond

The third section of this contribution has presented the main types of grammar-lexis relations that can be observed at word-form level in Arabic, and produced evidence for the crucial need for a comprehensive mapping of rules and relations involving the lexical nucleus of words in the computational morphology of the language.

To make this presentation clearer, three questions remain to be answered: (1) Are the rules and relations in consideration finite in number? (2) What is the actual extent of their coverage of word-form structures, relations and rules? (3) And what lays ahead, beyond word-level analysis? These issues are crucial in a computational perspective, because they are concerned with, respectively, the *feasibility* of the task of associating a whole lexicon with w-specifiers, the *reliability* of the software drawing on the lexical resource in consideration, and the compatibility of the results achieved in computational morphology with further developments involving sentence and text analyses.

7.4.1 Finiteness in Number of W-Specifiers, and the Feasibility of their Association with the Entries of an Entire Lexical Database

Arabic EF ↔ EF rules belong to finite sets for obvious reasons: EF-s are finite in number, and a finite set of relations are at stake, because EF-s belong to the grammatical morphemes of the language.

Altough lexical NF-s belong to open sets, the **finiteness of NF ↔ EF w-specifiers**, i.e. of the number of w-specifiers to be associated with the entire set of lexical entries, can be demonstrated as follows (Dichy, 1997):

(a) Because EF-s belong to finite sets, it follows that a finite set of features constricting NF ↔ EF relations can be established.
(b) This finite set of features, which corresponds to morpho-syntactic w-specifiers, can in turn be associated with every entry of a lexical database, i.e., to each NF.

In other word, both morphological grammar-lexis relations and the task of assigning every entry of a lexical resource with specifiers operating at word-level can be demonstrated as limited. Though lengthy, the task can subsequently be performed in a limited period of time.

7.4.2 Finiteness *and* Exhaustiveness in the Coverage of Data at Morphological Level, and the Reliability of Resulting Lexical Resources and Analyzers

Because grammar-lexis relations are liable to be embedded in finite sets of w-specifiers, it follows that the coverage of data at word (or morphological) level can be exhaustive. In other terms, *finite* sets of w-specifiers can produce *exhaustive* coverage of data within the boundaries of the word-form. This ensures a very high level of reliability in the software and analyzers drawing on a language resource such as DIINAR.1.

The above demonstration naturally goes beyond the mere case of the Arabic language. The criteria of finiteness *and* exhaustiveness in linguistic sets of features have been introduced by Mel'čuk (1982) in the general context of morphological description, and later taken up by him in lexicography. On the other hand, the demonstration also partly explains, in our view, why morphological and morphotactic rules and relations accounting for word-form generation and/or analysis, can be implemented very effectively using finite state transducers (Beesley & Karttunnen, 2003; Karttunnen, 1994).

7.4.3 Beyond Computational Morphology: Grammar-lexis Relations at Sentence and Text Levels

We now come to our conclusive remark, which deals with the usefulness of morphosyntactic specifiers included in rules and relations operating at word-form level in sentence and text analysis. Grammar-lexis relations at sentence level (s-specifiers), and furthermore at text level (t-specifiers) will, for obvious reasons, require different approaches. In the related resources, contextual relations need to be established between categories and sets of morphemes on the one hand, and sets of denotative and referential semantic categories and features on the other. The mapping of grammar-lexis relations already performed in the finite domain of computational morphology can nevertheless be considered a substantial progress towards lexical resources needed at sentence and text-level, owing the high degree of complexity of word-form structures in Arabic, compared to, say, French or English. W-specifiers already include a number of semantically related syntactic features needed at higher levels of analysis, such as (in)transitivity (including related prepositional structures) in verbs, type of plural according to the lexical subcategory a given nominal form belongs to and many other lexical categorisations and descriptions, which have been exemplified in various sections of this chapter.

References

Abbès, R. (2004). *La conception et la réalisation d'un concordancier électronique pour l'arabe*. Thèse de doctorat en sciences de l'information, ENSSIB/INSA, Lyon.

Abbès, R., Dichy, J. & Hassoun, M. (2004). The Architecture of a Standard Arabic lexical database: some figures, ratios and categories from the DIINAR.1 source program. In *Proceedings of the* COLING-04 *Workshop on Computational Approaches to Arabic Script-based Languages* (pp. 15–22), Geneva.

Ammar, S. & Dichy, J. (1999). *Les verbes arabe.* Paris: Hatier. Fully Arabic version, with specific introduction: *Al-'afālu l-çarabiyya,* الأفعال العربية (same publisher and year).

Arar, M. (2003). *Dāhiratu l-labsi fī l-çarabiyya* [The phenomenon of ambiguity in Arabic,ظاهرة اللبس في العربية]. Amman: Dār Wā'il.

Aronoff, M. (1994). *Morphology by Itself: Stems and Inflectional Classes*. Cambridge, MA: MIT Press.

Beesley, K. (1989). Computer Analysis of Arabic Morphology: A two-level approach with detours. In Comrie, B. & Eid, M. (Eds.) (1991), *Perspectives on Arabic Linguistics III: Papers from the Third Annual Symposium on Arabic Linguistics* (pp. 155–172). Amsterdam: John Benjamins.

Beesley, K. (2001). Finite-state morphological analysis and generation of Arabic at Xerox research: Status and plans in 2001. In *Proceedings of the ACL-01 Workshop on Arabic Language Processing: Status and Prospects* (pp. 1–8), Toulouse, France.

Beesley, K. & Karttunen, L. (2003). *Finite State Morphology.* Stanford, CA: CSLI Publications.

Bentin, S. & Frost, R. (1995). Morphological factors in visual word identification in Hebrew. In Feldman L.B., (Ed.), Morphological aspects of language processing (pp. 271–292). Hillsdale, NJ: Erlbaum.

Buckwalter, T. (2002). *Buckwalter Arabic Morphological Analyzer Version 1.0.* Linguistic Data Consortium, Philadelphia. LDC catalog number LDC2002L49 and ISBN 1-58563-257-0.[14]

Cantineau, J. (1950a). La notion de 'schème' et son altération dans diverses langues sémitiques. In *Semitica, 3,* 73–83.

Cantineau, J. (1950b). Racines et schèmes. In *Mélanges offerts à William Marçais.* Paris : Maisonneuve.

Cassuto, P. (2000). Le classement dans les dictionnaires de l'hébreu. In Cassuto, P. & Larcher, P. (Eds.), *La sémitologie, aujourd'hui* (pp. 133–158).

Cassuto. P. & Larcher, P. (Eds.). (2000). *La sémitologie, aujourd'hui.* Travaux du Cercle linguistique d'Aix-en-Provence n°16, Publications de l'université de Provence:

Cohen, D. (1961). Essai d'une analyse automatique de l'arabe. *T.A. informations*. Reprod. in Cohen, D. *Études de linguistique sémitique et arabe* (pp. 49–78). The Hague/Paris: Mouton.

Desclés, J.-P., dir. (1983). (H. Abaab, J.-P. Desclés, J. Dichy, D.E. Kouloughli, M.S. Ziadah). *Conception d'un synthétiseur et d'un analyseur morphologiques de l'arabe, en vue d'une utilisation en Enseignement assisté par Ordinateur*. Rapport rédigé à la demande du Ministère des Affaires étrangères.

Diab, M. & Resnik, P. (2001). An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 255–262), Philadelphia, PA.

Dichy, J. (1984). Vers un modèle d'analyse automatique du mot graphique non-vocalisé en arabe. Presented at the Conference on "Communication entre langues européennes et

[14] Retrieved December 16, 2006, from http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49

langues orientales", Montvillargenne, Oise. Revised version in Dichy, J. & Hassoun, M. (Eds.), (1989), pp. 92–158.

Dichy, J. (1987). The SAMIA Research Program, Year Four, Progress and Prospects. In *Processing Arabic Report 2* (pp. 1–26). T.C.M.O., Nijmegen University, Netherlands.

Dichy, J. (1990). *L'écriture dans la représentation de la langue : la lettre et le mot en arabe.* Doctorat d'État, Université Lumière Lyon 2, Lyon.

Dichy, J. (1993). Deux grands 'mythes scientifiques' relatifs au système d'écriture de l'arabe. In *Savoir, images, mirages,* Journées d'Études arabes, Special issue of *l'Arabisant* (pp. 32–33). Paris: Association Française des Arabisants.

Dichy, J. (1997). Pour une lexicomatique de l'arabe : l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot. *Meta 42,* 291–306. Presses de l'Université de Montréal.

Dichy, J. (2000). Morphosyntactic Specifiers to be associated to Arabic Lexical Entries - Methodological and Theoretical Aspects. In *Proceedings of ACIDA 2000* (Vol. 'Corpora and Natural Language Processing', pp. 55–60), Monastir, Tunisia.

Dichy, J. (2003). Sens des schèmes et sens des racines en arabe: le principe de figement lexical (PFL) et ses effets sur le lexique d'une langue sémitique. In Rémi-Giraud, S. & Panier, L., dir., *La polysémie ou l'empire des sens* (pp. 189–211). Lyon: Presses Universitaires de Lyon.

Dichy, J. (2005). Spécificateurs engendrés par les traits [±animé], [±humain], [±concret] et structures d'arguments en arabe et en français. In Béjoint, H. & Maniez, F. (Eds.), *De la mesure dans les termes*, Actes du colloque en hommage à Philippe Thoiron (pp. 151–181). Lyon: Presses Universitaires de Lyon.

Dichy, J. Braham, A., Ghazali, S. & Hassoun, M. (2002). La base de connaissances linguistiques DIINAR.1 (DIctionnaire INformatisé de l'ARabe, version 1). In Braham, A. (Ed.), *Proceedings of the International Symposium on the Processing of Arabic,* Université de la Manouba, Tunisia.

Dichy, J. & Farghaly, A. (2003). Roots and Patterns vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical database centred on Arabic be built? In Proceedings of the *IXth MT Summit Workshop on Machine Translation for Semitic Languages: Issues and Approaches* (pp. 1–8), New Orleans.

Dichy, J. & Hassoun, M. (Eds.) (1989). *Simulation de modèles linguistiques et Enseignement Assisté par Ordinateur de l'arabe – Travaux SAMIA I.* Paris: Conseil International de la Langue Française.

Dien, D., Kiem, H. & Hovy, E. (2003). BTL: a Hybrid Model for English-Vietnamese Machine Translation. In *Proceedings of the IXth MT Summit* (pp. 87–94), New Orleans.

Ditters, E. (1992). *A Formal Approach to Arabic Syntax: The Noun phrase and the Verb Phrase.* Ph.D. dissertation, Catholic University of Nijmegen, Netherlands.

Farghaly, A. (1987). *Three Level Morphology*. Paper presented at the Arabic Morphology Workshop, Linguistic Summer Institute, Stanford, CA.

Farghaly, A. (1994). Discontinuity in the Lexicon: A Case from Arabic Morphology. In *International Conference on Arabic Linguistics*, The American University in Cairo, Cairo, Egypt.

Fassi-Fehri, A. (1997). *Al-Maçjama wa-t-taxTīT – NaḌarāt jadīda fī qaDāyā l-luγa l-çarabiyya* [Lexicography and language planning. Arabic Language matters reconsidered, المعجمة والتخطيط – نظرات جديدة في قضايا اللغة العربية]. Casablanca, Morocco: Al-Markaz al-thaqāfiyy al-çarabiyy.

Forster, G., Grandrabur, S., Langlais, P., Plamondon, P., Russel, G. & Simard, M. (2003). Statistical Machine Translation: Rapid Development with limited Resources. In *Proceedings of the IXth MT Summit* (pp. 110–117), New Orleans.

Frost, R., Deutsch, A. & Forster, K.I. (2000). Decomposing morphologically complex words in a non linear morphology. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*, 751–65.

Frost, R., Forster, K.I. & Deutsch, A. (1997). What can we learn from the morphology of Hebrew? A masked priming investigation of morphological representation. *Journal of Experimental Psychology: Learning, Memory and Cognition, 23*, 829–856.

Geith, M. & El-Saadany, T. (1987). Arabic morphological analyzer on a personal computer. Presented at the Arabic Morphology Workshop, Linguistic Summer Institute, Stanford, CA.

Ghenima, M. (1998). *Analyse morpho-syntaxique en vue de la voyellation assistée par ordinateur des textes écrits en arabe*. Ph.D. dissertation, ENSSIB/Université Lyon 2.

Grainger, J., Dichy, J., El-Halfaoui, M. & Bamhamed, M. (2003). Approche expérimentale de la reconnaissance du mot écrit en arabe. In Jaffré, J.-P. (Ed.), *Dynamiques de l'écriture: approches pluridisciplinaires*. *Faits de langue*, 22, 77–86.

Hans Wehr. (1979) A dictionary of modern written Arabic. 4th edition, edited. by J. Milton Cowan. Wiesbaden, Harrassowitz.

Hassoun, M. (1987). *Conception d'un dictionnaire pour le traitement automatique de l'arabe dans différents contextes d'application.,* Ph.D. (thèse d'État), Université Lyon 1.

Hlal, Y. (1979). *Méthode d'apprentissage pour l'analyse morphosyntaxique (expérimentée dans le cas de l'arabe et du français).* Ph.D. dissertation, Université Paris-Sud, Centre d'Orsay.

Hlal, Y. (1985a). Morphology and syntax of the Arabic language. *Arab School of Sciences and Technology: Informatics* 4C, 1–8.

Hlal, Y. (1985b). Morphological analysis of Arabic speech. In *Workshop Papers Kuwait/Proceedings of Kuwait Conference on Computer Processing of the Arabic Language* (Section 13, pp. 273–294).

Karttunnen, L. (1994). Constructing Lexical Transducers. In *Proceedings of COLING-94*, (pp. 206–411), Tokyo, Japan.

Karttunnen, L. & Beesley, K.R. (2005). Twenty-five years of finite-state morphology. In Arppe, A., Carlson, L., Lindén, K., Piitulainen, J., Suominen, M., Vainio, M., Westerlund, H. & Yli-Jyrä, A. (Eds.), *Inquiries into Words, Constraints and Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday (2005)*. CSLI Studies in Computational Linguistics ONLINE, pp. 71–83. Copestake, A. (Series Ed.). Stanford, CA: CSLI Publications.

McCarthy, J. (1981). A Prosodic Theory of Nonconcatenative Morphology. *Linguistic Inquiry*, *12*, 373–418.

McCarthy, John J. & Prince, Alan S. (1996). Prosodic morphology. Technical report 32, Rutgers University Center for Cognitive science.

Melčuk, I. A. (1982). *Towards a Language of Linguistics: A System of Formal Notions for Theoretical Morphology.* München: Wilhem Fink Verlag.

Nikkhou, M. (Ed.) (2004). *NEMLAR International Conference on Arabic Language Resources and Tools,* Cairo. Paris: ELDA.

Ouersighni, R. (2001). A major offshoot of the DIINAR-MBC project: *AraParse*, a morpho-syntactic analyzer of unvowelled Arabic texts. In *ACL-01 Workshop on Arabic Language Processing: Status and Prospects* (pp. 66–72), Toulouse, France.

Ouersighni, R. (2002). *La conception et la réalisation d'un système d'analyse morpho-syntaxique robuste pour l'arabe: utilisation pour la détection et le diagnostic des fautes d'accord.* Ph.D. dissertation, ENSSIB/Université Lyon 2.

Rogati, M., McCarley, S. & Yang, Y. (2003). Unsupervised Learning of Arabic Stemming Using a Parallel Corpus. In *41st Annual Meeting of the Association of Computational Linguistics* (pp. 391–398), Sapporo, Japan.

Roman, A. (1990). *Grammaire de l'arabe.* Paris: P.U.F., coll. "Que sais-je?".

Roman, A. (1999). *La création lexicale en arabe, ressources et limites de la nomination dans une langue humaine naturelle.* Presses Universitaires de Lyon.

Rousseau, J. (1987). La découverte de la racine en sémitique par l'idéologue Volney. *Historiographia Linguistica, 14*(3), 341–365.

Sampson, G. (1985). *Writing systems.* Stanford University Press.

Schafer, C. & Yarowsky, D. (2003). A Two-Level Syntax-Based Approach to Arabic-English Statistical Machine Translation. In *Proceedings of the IXth MT Summit Workshop on Machine Translation for Semitic Languages: Issues and Approaches* (pp. 45–52), New Orleans.

Soudi, A., Cavalli-Sforza, V. & Jamari, A. (2001). A Computational Lexeme-Based Treatment of Arabic Morphology. In *ACL-01 Workshop on Arabic Language Processing: Status and Prospects* (pp. 155–162), Toulouse, France.

Troupeau, G. (1984). La notion de 'racine' chez les grammairiens arabes anciens. In Auroux, S., Glatiny, M., Joly, A., Nicolas, A. & Rosier, I. (Eds.), *Matériaux pour une histoire des théories linguistiques,* pp. 239–245. Presses Universitaires de Lille.

Zaafrani, R. (2002). *Développement d'un environnement interactif d'apprentissage avec ordinateur de l'arabe langue étrangère*. Ph.D. dissertation, ENSSIB/Université Lyon 2.

Zwiep, I. E. (1996). The Hebrew linguistic tradition of the Middle Ages. *Histoire Épistémologie Langage, 18*(1), 41–61.