

Written Communication and the Mental Lexicon

Agglutinative Word-Form Recognition in Arabic

Joseph Dichy

Abstract. This paper is concerned with the overall structure of the mental lexicon in an agglutinative language such as Arabic. Due to that feature, as reflected by the writing of the language, a great number of nouns, adjectives and verbs do not appear as nude word-forms and include formatives on their right and/or left side. This implies that the recognition of the lexical stem of a word-form is a complex process, which is accounted for by a *word-formative grammar*. This is not without consequences for the structure of the mental lexicon, both in general and in that language.

It well-known today that the mental lexicon cannot be reduced to a mere sense/form set of features. In addition to meaning (or sense) and to phonic or graphic form, one must add a third fundamental aspect, which relates the lexical entry to its context. This is what Igor Melcuk (1982) calls the *syntactic*, which includes, e.g. gender in languages such as Latin, French or German... Verbs can be transitive or intransitive but also accept—or not—human (or quasi human) subjects or complements (Dichy, 2005).

In this paper, we consider the *syntactic* of Arabic lexical entries within the frame of the written word-form, the features of which are called *morpho-syntactic specifiers*. These include, at first, the indication of the major lexical category of the stem, due to the fact that some proclitic tool-words are related to nouns (typically, cliticised prepositions) and others to verbs (e.g. cliticised conjunctions). Complement pronouns being also cliticised, the transitivity or intransitivity of verbs must also be taken into consideration. These are basic examples. In this paper, a complete list of Arabic word-form specifiers is presented. In the DIINAR database, these have been added on a systematic basis to all the 129,000 lexical entries.

The paper illustrates the fact that morpho-syntactic specifiers are necessary in Arabic for the recognition of word-forms. They subsequently need to be added to the meaning and form of lexical entries in the mental lexicon of the language.

Joseph Dichy  0000-0002-9123-7358
Canadian University Dubai
E-mail: joseph.dichy@yahoo.fr

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 209–228. <https://doi.org/10.36824/2024-graf-dich>
ISBN: 978-2-487055-10-0, e-ISBN: 978-2-487055-11-7

1. Introduction

The structure of the mental lexicon is a crucial aspect of grapholinguistics. It is well-known that the mental lexicon needs to register both the acoustic forms of lexical entries and their written realisations, e.g.,

The English written word *read* is realised phonetically as either [ri:d] or [red] and corresponds, respectively, to either the present-tense or the preterit of that verb.

It is less known, though, that many other pieces of information are to be added for the reading process to operate, particularly in an agglutinative language such as Arabic, where:

a verb can be transitive or not, accept human subjects or human complements, or not (Dichy 1990, chapter X), and require, or not, a plural subject (Fassi-Fehri 2012), a noun can be masculine/feminine (or neutral), countable or not, collective/singulative...

Such morpho-syntactic factors directly affect the structure of the Arabic written word-form. One also needs to mention other types of information associated with lexical entries and affecting word-form recognition in the process of syntactic group or sentence understanding, such as collocations, lexical functions (Mel'cuk et al. 1984; 1995), phrasal relations, etc.

How do such relations operate? Elaborating descriptions or databases including a seemingly unlimited number of relations associated with every item of the lexicon seems at first sight extremely difficult to achieve. Natural languages encompass, in addition to limited inventories of grammatical items, an unlimited series of lexical or terminological items. The view according to which one cannot really deal with unlimited inventories can, however, be challenged.

One knows that a small number of syntactico-semantic features allow the insertion of lexical items in syntactic structures (Chomsky, 1972...). The author has shown (Dichy, 1997; 2005) how in Arabic and other languages, a *limited number* of morpho-syntactic and syntactico-semantic features restrict subject-to-verb and verb-to-complement relations in argument structures associated with verbs, e.g.,

In English, the verbs *to think* or *to believe* require a human grammatical subject (or an assimilated cognitive agent); *to read* requires a non-human grammatical object-complement¹. So do their respective Arabic equivalents, *fakkara*, *ʔiʔaqada*, and *qaraʔa*.

1. Except in the phrase *I read you*, where this verb takes another meaning, which entails considering it as another entry in a lexical database. English grammars often disregard the [±human] feature, which must, in the author's view, be added to the well-known [±animated] one. In Arabic, only the former is systematically relevant.

A phrasal verb such as *to put off* has a different meaning depending on the [$\pm$ human] feature of their complement (*to put off someone* vs. *to put off something*). In Arabic, the verb *raġiba* has two opposite meanings depending on the associated preposition, *raġiba fī* signifying “to desire” or “wish to,” and *raġiba ʿan*, to “dislike” or “detest”.

Recalling known approaches to these questions² is beyond our scope here. The essential point is that grammar-lexis relations associate unlimited—or non-finite—items, such as the entries of the lexicon of a given language, to limited—or finite—ones,³ such as grammatical categories, relations, and constraints. The fact that a finite number of relations projected on the non-finite entries of the lexicon ‘manage’ linguistic utterances is a general characteristic of natural languages, which makes it possible for either production or recognition to operate.

This general characteristic has been illustrated by the author in the recognition of Arabic word-forms. As a contribution to language theory and to grapholinguistics, we present here the set of morpho-syntactic information associated to the entries of the Arabic lexicon. These morphosyntactic pieces of information are needed for producing or recognising written word-forms.

We consider indicating how a set of information similar to that observed in Arabic operates in Biblical Hebrew or in Aramaic/Syriac word-forms in further works. The written word-forms of this family of Semitic languages can be described as *agglutinative*. In these languages, grammatical morphemes are ‘agglutinated’ in a very structured and consistent way around a single lexeme, according to what can be described as a word-formative grammar (Desclés et al. 1982; Dichy 1990). Arabic and Biblical Hebrew (as opposed to Modern Hebrew) only feature monolexical entries, a subsequent majority of which can be analysed into a Semitic consonantal root and a morphosyntactic pattern. There are a few compound entries in Arabic, but their very small number and the fact that they are only found in nominal or adjectival entries allow one to consider the phenomenon as marginal. The complex structure of the word-form, which includes the root and pattern structure, offers a good explanatory hypothesis of the mono-lexical structure of Arabic entries.

2. Let us quote, though, the semantic and cognitive approach of Desclés (1994...) and M. Abraham (1995), and, for Arabic, in a different theoretical frame, A. Fassi-Fehri (2012...).

3. Melchuk (1982, 1984); for Arabic see Dichy (1997), Dichy & Farghali (2003) and the DIINAR (DIctionnaire INformatisé de l'ARabe) data-base (Dichy, Braham, Ghazali & Hassoun, 2002).

1.1. Structure of the Agglutinative Written Word-Form in Arabic

In the agglutinative graphic word-form structure of Arabic, one encounters a number of grammatical lexemes and a single lexical nucleus, according to the example below:

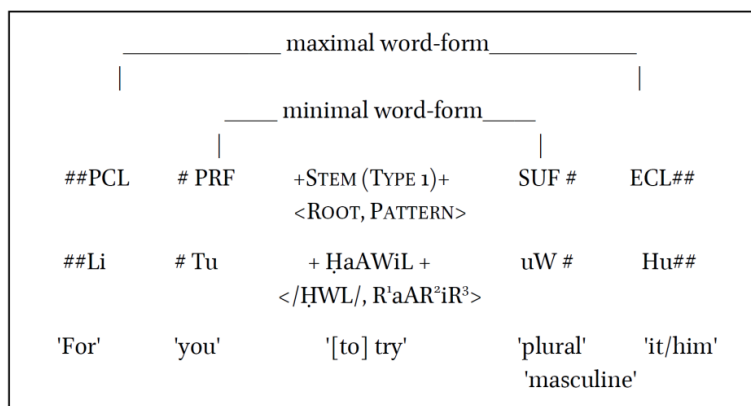


FIG. 1. Scheme of the graphic Arabic word-form in Arabic

Word-Form Codes Used in the Above Table and Below

1. Word formatives

STEMS or *nuclei* are either analysable into ROOTS and PATTERNS (type 1), as in the example above, or not (type 2), as with loan-words or a great number of proper names.

PCL = proclitic(s), PRF = prefix, SUF = suffixe(s), ECL = enclitic(s);

2. Word-formative boundaries

+ = strong word-formative boundary, or lexical nucleus boundary,

= weak word-formative boundary or clitic boundary, occurring after a PCL or before an ECL.

3. ROOT and PATTERN codes

R¹, R², R³, or R1, R2, R3, respectively 1st, 2nd and 3rd radical consonant in 3-consonantal roots. R⁴ or R4 is added in 4-consonantal roots. A PATTERN can be represented (a) either with these numbered radical consonants, to which other elements, such as short or long vowels, are added or (b) with the traditional Arabic grammar code, using letters /f/, /ʕ/ and /l/, and repeating /l/ in 4-consonant ROOTS, e.g., in 3-consonant ROOTS: *makātib*, “offices”: (a) maR1āR2iR3 and (b) mafā’il and in 4-consonant ones: *ʿarqala*, “to hinder, impede” or “obstruct”: (a) R1aR2R3aR4a and (b) faʿlala.

4. Transcription codes only used in the above table

Capital letters transcribe Arabic graphemes written in the body of word-forms, while *lower-case letters* transcribe diacritical short vowel signs.

The frame presented in Fig. 1, primitively introduced by David Cohen (1961), has been taken up and developed in (Desclés et al. 1982; Dichy 1990) and so forth.

For the lexical nucleus to be correctly inserted into word-forms, it must be associated with *morpho-syntactic specifiers* (introduced in Dichy 1990 and 1997), which account for relations between the nucleus and the other formatives situated right and left of it in the graphic word-form. They need to be associated with every verb, noun and adjective of the language in the speaker's mental lexicon, or in a lexical database, such as the one designed and achieved in the second part of the 1990's for the DIINAR database (DIctionnaire INformatisé de l'ARABE)/معالي (معجم العربية الآلي)—between researchers located in Lyon and a high-level Arab research centre in Tunis⁴. These specifiers are limited in number: they account for relations between lexical nuclei and grammatical lexemes included in graphic word-forms. The fact that grammatical lexemes belong to closed lists accounts for the limited number of specifiers, which need in turn to be added to every lexical entry.

Psycholinguistic research on the recognition of written Hebrew (Frost et al., 1997, Frost et al., 2000) and written Arabic (Grainger, Dichy et al., 2003) has shown that the reading process in these languages is primarily based on contour recognition. This is highly in accordance with the formal analysis introduced here. It shows that the Semitic word-form reading process is based on closed-list information associated with every lexical item included in the mental lexicon of readers. These information items are called morpho-lexical specifiers.

2. Arabic Word Formatives and Associated Specifiers

2.1. Word-Form Nuclei Categories

In the DIINAR database, we have considered the category of verbs on one side and that of nouns and adjectives on the other, following the medieval Arabic grammatical tradition. Although nouns and adjectives need to be distinguished syntactically and semantically, we have found

4. *Institut Régional des Sciences de l'Informatique et des Communications* ("The Regional Institute of Computer Science and Tele-Communications"), Tunis, Profs. Salem Ghazeli and Abdelfattah Braham.

that the nouns/adjectives category remains sufficient for the recognition of word-forms. Adverbs, such as *ʾayḍan*, “also, too” or *laylan*, “nightly, by night” are included in the noun category.

2.1.1. *Lexical Nuclei*

Main Features

The main feature of the lexical nucleus is that it is either:

- unrelated to the root and pattern derivation system, such as nouns pertaining to the following categories:
 - proper names (generally borrowed, including anciently), e.g., *Butrus* (from the Latin *Petrus*, *Peter* in English), *Maryam*, *ʾIsmāʿīl*, *Yū-nescū* (UNESCO), *Sūriya*, “Syria,” or
 - loan-words, eg.: *banāna*, “banana,” *makīna*, “machine,” *mūbāyl* (“mobile phone”),
- representable in terms of root and pattern, such as:
 - nouns/adjectives, e.g., *karīm*, “noble and generous,” ROOT /K-R-M/, PATTERN *faʿīl*, or R1aR2iR3, where R1 = /K/, R2 = /R/, R3 = /M/,
 - verbs, e.g., *rasīla*, “to be long and flowing” [hair], which can be analysed into ROOT /R-S-L/, and PATTERN *faʿīla* or R1aR2iR3a.

Number of Radicals in ROOTS

I have shown (Dichy, 1990), that the Arabic language only features 3-consonant and 4-consonant roots, the definition criterion in Semitic languages being that roots must be subject to semitic-morphological derivation with simple and augmented patterns, e.g.,

- the 5-consonant word *barnāmaj*, “roster,” or “program” has a 4-consonant plural, *barāmij*, featuring an ad-hoc 4-consonants ROOT /B-R-M-J/, which allowed, in modern times, the verb *barmaja*, “to program”;
- *safarjal* “quince” has for a plural *safārij* (ad-hoc 4-c. ROOT /S-F-R-J/);
- the 2-consonant word *yad* “hand” has 3-consonants plurals: *ʾayādin* and *ʾaydin* (ad-hoc 3-consonants ROOT /Y-D-Y/); medieval dictionaries also mention *ʾayyada* and *yadda*, “giving a hand, helping and supporting” (ad-hoc ROOTS /ʾ-Y-D/ and /Y-D-D/).

Also note that *all* verbs and verb-derived nouns and adjectives pertain to the Semitic root-and-pattern frame, which, on the other hand, only includes a subset of nouns and adjectives (Dichy, 1990).

Consistency of Nucleus Representation

In this representation, consonants belonging to the pattern and added to the root are considered as part of the lexical nucleus, whether they appear between the ROOT consonants, e.g.,

- *wāw* inserted before the last ROOT consonant, as in *durūs*, “lessons,” or whether they appear at the beginning or the end of the PATTERN, e.g.,
- /a&/⁵ added at the end of *kabāma&*, “dullness” or “lassitude, weakness”; *ʾistiḳāma&*, “honesty”;
- *madraj*, “course,” “route,” “path,” “road” ...⁶

This representation has the advantage of avoiding splitting the word-form nucleus, and of featuring it, for the sake of consistency, as a full entry of the mental lexicon or of a lexical database. Morpho-syntactic or syntactico-semantic specifiers are associated with that entry as a whole.

2.1.2. *Tool-Words or Void Nuclei*

Tool-words can also appear in the position of a word-form nucleus, e.g.,

- *wa#inna#bum*, where “#” is for clitic boundary, *inna* is a conjunction, and *#bum* is the plural masculine complement pronoun.

Some word-forms do not include a nucleus, e.g.,

- *La#bum* “for them,” *bi#bi* “by him or it”.

When the ECL pronoun is syntactically marked as a direct-case pronoun, an epenthetic nucleus appears, e.g.,

- *ʾanā wa#ʾiyyā#ka*, “me and you” [as one]; Koran I,5: *ʾiyyā#ka naʿbud+u*, “Thee [only] we worship”.⁷ Another example is the use of *miṭl*, “like” or “likeness” in the phrase-word *ka#miṭl+i#bi* “similar to” (Koran 42/11: *laysa kamiṭlibi šayʾun*, “There is naught to the likeness of Him”).

5. By convention (after D. Cohen 1961), the *tāʾ marbūʿa* is represented as *a&*. There are two distinct representations (Desclés et al., 1982, Dichy 1990):

- without a suffix boundary as in the example above, where it belongs to the pattern, or
- with a strong boundary symbol “+,” i.e., *+a&* when it is a feminine gender suffix, e.g., *jamil+a&*, “pretty,” feminine, or a singulative SUF (*ʾism al-wāḥid*), as in: *naml+a&*, “ant,” plur. *naml+āt*.

6. For the analysis of these two pattern ‘infixes’ (*m+* and *+a&*) identified as 1-consonantal roots, see A. Roman (2011).

7. Current Arabic grammars present the sequence *<ʾiyyā+complement pronoun>* as an “isolated direct case pronouns” (*ḍamāʾir munfaṣila manṣūba*), which is doubtful, since isolated pronouns are always of the subject case.

In this paper, we do not present the tables of compatibility between tool-words and proclitics and/or enclitics (ECL), due to the fact that every tool-word category has its own compatibility rules. In fact, these tables, which I devised for the DIINAR database and associated analysers, allow the building of a generated lexicon for Arabic natural language processing (NLP), which only includes extant items.

2.2. Prefixes (PRFs)

Prefixes are, after D. Cohen (1961/70), strictly verbal and only appear in the conjugation of the imperfect or the imperative. These are, for the imperfect:

- *ʾa+* or *ʾu+*. This PRF is for the first-person singular;
- *na+* or *nu+*, first-person plural;
- *ta+* or *tu+*, which is for the second person masculine/feminine, as well as for the third person feminine;
- *ya+* or *yu+*, which is for the third person masculine/feminine.

These prefixes are in fact personal pronouns indicating the first, second, or third person (masc. or fem.). Only *ʾa+* and *na+* indicate a difference in number (singular vs. plural). The selection of the vowel *a* or *u* above depends on the morpho-syntactic pattern of the verb. Vowel *u* appears with patterns *faʿala*, *ʾafʿala* and *fāʿala* or with the passive voice.

For the imperative:

- *ʾu+* or *ʾi+* is an epenthetic *hamza* appearing at the beginning of the imperative verbs pertaining to the simple pattern and in some augmented patterns, with a variation of the vowel according to a well-known grammar rule (due to apophony).

2.3. Suffixes (SUFs)

SUFs are divided into nominal/adjectival and verbal ones. In the same way as prefixes, their presence in the word-form is morphologically necessary: all lexical entries have suffixes, the minimal suffixes in nouns being the case ending, and in verbs, gender and number desinences. This is formally represented by the “+” strong boundary symbol.

2.3.1. Verbal SUFs

These suffixes are related to conjugation paradigms, which have been modeled in Ammar, and Dichy (1999/2008), in accordance with the representation devised for the conjugation of verbs in various languages

by the French grammarian Bescherelle. His basic idea is that a small number of verb models (ranging from 88 to over 200) can allow the conjugation by analogy of all the verbs of a language, provided the rules of the language are modeled in such a way as to produce a strict formal classification of verbs in relation to conjugation. (In today's terms, this is described as a taxonomy, i.e., a formal classification where a given item appears in one slot and one only.) Regarding the Arabic language, I have designed a 2-dimensional table featuring horizontally the simple and augmented verb patterns, and vertically the 33 types of 3-consonantal roots, according to the number of the transformation causes they include, such as a radical consonant *hamza*, *waw* or *ya*, two identical consonants in second and third position ($R_2 = R_3$), and two or more of these features within a given root.⁸ Another table accounts for 4-consonantal roots.

The conjugation of Arabic, complex as it may seem with certain verbs, is regular (with a very small number of exceptions). There are, subsequently, two ways of answering any question related to it:

- by applying the rules, which requires a certain level of calculation, and is likely to occur as a remediation strategy, when the speaker encounters a difficulty;
- by analogy, through imitating the conjugation of a similar verb. This is the idea brought forth by Louis-Nicolas Bescherelle in the 19th century, which is applied in the Bescherelle series of conjugation reference books for many languages, published in France today by Hatier. It includes Ammar & Dichy's *Bescherelle des verbes arabes* (1999/2008). The 125 verb models identified in this work are numbered, and the volume includes a lexicon of verbs, each entry of which is associated with the number of the model according to which it can be conjugated.

Human memory may well operate by analogy, with the possibility for speakers to resort to calculation when they encounter a difficult point. Whatever it is, Ammar & Dichy's analogical models (1999/2008 and 2021)⁹ perfectly fit conjugation systems assisted by a database.

8. See Dichy 1993; Ammar & Dichy 2008, p. 34-36; 2021, chapter 1. By contrast, other conjugation presentations include hundreds of models, while not covering the whole conjugation of the language (see Bayyūmī et al. 1989; Osman 1995; Reig 1983; Sheidlin 1978).

9. The former volume (1999/2008) includes a lexicon of 10,000 verbs, and a short grammar; the latter (2021) has a reference lexicon of just under 20,000 verbs and includes many grammatical remarks in relation with the verb models, and a detailed theoretical presentation of the analysis that led to a complete coverage of Arabic conjugation.

Conjugation Morphosyntactic Specifiers

In the DIINAR database and its analysers¹⁰, the numbers associated with every verb and referring to the 125 analogical models of Ammar & Dichy (1999/2008) are considered as morphosyntactic specifiers related to conjugation. A pattern-matching generation program based on these models has been designed and implemented by R. Zaafrani (2002).

Other specifiers related to conjugation are concerned with the passive voice, the imperative, or the absence of active and/or passive participles:

- Some transitive verbs can only be conjugated in the passive voice at the third person. This occurs when the verb is transitive to non-human complements, e.g.,
 - *waba'a* "to prepare" [luggage], *hayya* "to live" [something, an experience].

In such verbs, the non-human complement cannot "speak" (1st person) or "be addressed" (2nd person) in the passive voice.

- Some intransitive verbs, or prepositional transitive verbs, can only be conjugated in the passive voice at the third person masculine singular, e.g.,
 - *furiḥa* "to be glad," "to rejoice"; *wali'a* "to be madly in love (with)," "to be passionately fond (of)," followed in Arabic by the preposition *bi*#.

In some verbs, in the passive voice, the first and the second person are strictly excluded for semantic reasons, and the 3rd person expresses 'impersonal passive' phrasing, meaning, word-for-word:

- *furiḥa*, "One has rejoiced," i.e., "People rejoiced"; *fatātun kaṭīran mā wuli'a biḥā*, "a maiden many-a-time [men] fell in love with"...
- Some verbs are both intransitive and non-agentive, which excludes the passive voice and the imperative, e.g.,
 - *saruwa* "to be of noble descent"; *yamuna* "to be blessed by luck"; *ra'ufa* "to be of merciful nature"; *qabuḥa* "to be ugly".
- In addition to the preceding constraint, some strictly non-agentive verbs do not have an active and/or passive participle, e.g.,
 - *ṣaḡura* "to be small," for which one can neither say *ṣaḡīr* in this meaning of the verb, nor **maṣḡūr*. Instead, one uses the analogous adjective (*ṣifa muṣabbaba*) *ṣaḡīr*.
 - Other analogous adjectives are related to some of the above examples, e.g., *ra'ūf* "merciful (by nature)" for *ra'ufa* or *qabīḥ* "ugly" for *qabuḥa*.

10. See the doctoral dissertations co-supervised by the author of Ghenima (1998), Zaafrani (2002), Ouersighni (2002), Raheel (2010).

2.3.2. Nominal/Adjectival SUFFs

Nominal/adjectival suffixes are of five general types:

- gender and number suffixes, such as:
 - *+a&* (feminine singular), *+āt* (feminine plural), *+āʾ* (feminine singular in some adjectives) *+ūna* (masculine plural), *+ū* (masculine plural, elided *na*)...;
- case-ending suffixes, such as:
 - *+un* (subject case, indefinite), *+u* (subject case, indefinite, *n* elided), *+an* (direct case, indefinite, written with *alif*), *+a* (direct case, indefinite, written with *alif*, *n* elided), *+in* (indirect case, indefinite), *+i* (indirect case, indefinite, *n* elided), etc.;
- compound gender-number and case-ending suffixes, such as:
 - *+ūna* (masc. pl., subject case) and *+īna* (masc. pl., direct and indirect cases); *+āni* (dual number, subject case) and *+ayni* (dual number, direct and indirect cases); *+āt+un* (fem. pl., subject case) and *+āt+in* (fem. pl., direct and indirect cases), etc.
- the suffix *+iyy* of the relative noun-adjective (*yāʾ* ‘*al-nisba*’);
- the compound suffixes *+iyya* and *+iyyāt* are used in modern Arabic for abstract notions, e.g.,
 - *silmi+iyya*, “pacifism”; *lisān+iyyāt*, “linguistics”.

Morphosyntactic Specifiers Related to Nominal/Adjectival SUFFs

These suffixes are subject to lexical or grammatical constraints, and require specifiers associated with lexical entries:

- Indication of gender and number suffixes must be given systematically, e.g.,
 - *mudarris*, “teacher,” masc. plur., *mudarris+ūna/mudarris+īna*; the feminine is *mudarrisa&un*, fem. plur., *mudarris+āt+un/mudarrisāt+in*.
- A great number of plural forms are not built with suffixes but require a change in the pattern of the noun or adjective. These are traditionally referred to as ‘broken plural’ forms. When this occurs, it requires that, in the database, plural and singular forms be registered together and that the relation between the singular and the broken plural be ‘arrowed’, e.g.,
 - *kurs+iyy*, “chair,” broken plural, *karās+in*; *safīr*, “ambassador,” plural *sufarāʾ*; adjective *karīm*, “noble and generous,” plural forms *kuramāʾ* and *kirām*...
- ‘Arrowing’ also applies to the feminine and plural of many adjectives, such as:
 - *ʾazraq* “blue,” fem. sing. *zarqāʾ*, pl. *zurq*.

- In some cases, the plural is in modern Arabic built with either a suffix or a change in the pattern (broken plural), e.g.,
 - *jamīl* “beautiful, pretty,” masc. pl. for humans *jamīl+ūna*, fem. *jamīl+āt+un*, shared broken plural *jumalā*’.

This requires indications of suffixes, broken plurals, or feminine forms featuring a change in pattern (as with *ʿazraq* above) that cover all cases. These relations are, of course, limited in number.

- Case endings vary with a subset of nouns and adjectives, known as diptotes (*mamnūʿ mina al-ṣarf*), e.g.,

- *ʿaswad*, “black,” fem. *sawda*’, do not take all the case endings.

This requires a specific morphosyntactic specifier, due to the fact that not all diptotes are predictable, e.g.,

- the proper noun *Nibād* is a diptote when it is feminine, and not when it is masculine; *ʿašyā*’ “things”, is lexically coded as a diptote.

- The relative suffix *+iyy* does not appear with all nouns or adjectives, due to usage. This means that its occurrence with a given entry needs to be recorded, either as an additional lexical entry or with an associated specifier. E.g.:

- *bayrūt*, “Beirut” allows *bayrūt+iyy*, “related to, or inhabitant of Beirut”.

- On the other hand, *Dubay*, “Dubai” or *Jibūṭī*, “Djibuti,” which end with a *y*, do not accept the suffix *+iyy*. This entails the use of suppletive forms, such as *ʿabl Dubay* or *jibūṭī*, “inhabitants of Dubai or Djibouti”; the same can be said, for other reasons of *Abū Dabī* or *Rās al-ḥayma*...

- The suffix *+iyy* may entail a modification of the noun it is suffixed to, e.g.,

- *qabīla*’, “tribe” becomes *qabal+*, when *+iyy* is suffixed: *qabal+iyy*, “tribal”.

- In contrast, *ṭabīʿa* “nature” is not affected in its structure by this suffix: *ṭabīʿ+iyy*, “natural”.

- Some proper nouns feature a more radical change in their structure, e.g.,

- *al-quds*, the Arabic name for “Jerusalem,” has a relative noun-adjective *maqdis+iyy* (built after the ancient Islamic name of the city *Bayt el-Maqdis*). The adjective *quds+iyy* being already taken in the lexicon, with the meaning of “sacred,” in contemporary Arabic, one says *ʿabl* or *sukān al-quds*, “people or inhabitants of Jerusalem [Arabic name]”.

- The city of Casablanca is called in Arabic *al-dar al-bayḍā*’, the relative noun-adjective being *bayḍāw+iyy*.

In the same way as with broken plurals, an ‘arrowing’ relation must, when applicable, link the stem of the noun with and without the suffix *+iyy*.

- When the noun ends with /+a&/, the introduction of +iyy entails the deletion of that marker, e.g.,
 - *jāmi*‘a&, “university” becomes *jāmi*‘+iyy, “academic,” “related to a university”.

This allows the <noun+iyy> compound, shaped like a *nūn* to have a feminine form <noun+iyy+a&>, avoiding a ‘doom loop’ such as:

*<noun+a&+iyy+a&+iyya...>.

2.3.3. *Proclitics (PCLs)*

PCLs are of various types. Basically, they are 1-consonant tool-words (the article *al*# being a specific case, featuring an /l/ consonant preceded by an epenthetic hamza).

The main difference between prefixes and suffixes, which are included in minimal words (see Fig. 1), is that PCLs are added to them without being morphologically necessary to the structure of the word-form. This is formally represented by the # weak boundary symbol.

PCLs are divided into three general categories, the first of which is common to all major category nuclei, the second and the third being respectively related to nouns and verbs.

PCLs are submitted to a strict order of occurrence in the word-form, as well as to a number of grammar rules. Both fall under the word formative grammar (Desclés & al, 1982; Dichy 1990, chap. X, §3.4).

Two features are related to the readability of word-forms and facilitate word-form recognition:

1. PCLs belong to a limited set of consonants: /ʔ/, /w/, /f/, /k/, /b/, /l/, /ʔā/, /s/, /l/, to which one must add the article /ʔal/.
2. They appear according to the following strict order:
 - Position 1: /ʔa#/, interrogative
 - Position 2: /wa#/, /fa#/, coordinating conjunctions;
 - Position 3: /la#/ corroborative conjunction (*lām al-tawkîd*)
 - Position 4: /ka#/, comparative particle
 - Position 5: /bi#/, /li#/, prepositions
 - Position 6 and 7: /ʔal#/, definite article
 - Position 7: /li#/, energetic particle, which loses its vowel /i/ when combined with the coordinating conjunctions *wa#* or *fa#* and produces: /wa#l#/, /fa#l#/; /sa#/, future marker; /wa#/, *wāw rubba*; /ʔa#/, /ʔā#/ vocative particles.

The above list includes PCLs that one does not expect to occur in contemporary Arabic. It is nevertheless essential to retain them, the system needing to be sturdy enough to analyse both modern and classical corpora.

2.3.4. *PCLs Shared by All Major Categories*

Four PCLs appear at the beginning of a word-form (positions 1-3) according to the above-mentioned strict order.

The first shared category of PCLs is the interrogative tool-word *hamza*, which always comes first when occurring in a word-form.

The second category in the word form order of PCLs is that of the 1-consonant coordinative conjunctions, i.e., *wa#* “and,” *fa#* “and then,” and of the corroborative conjunction *la#*, which can only be translated into English through the addition of an adverb such as “surely,” “indeed,” or “certainly”.

2.3.5. *Verbal PCLs*

In position 7, one finds 2 verbal PCLs,

- the energetic particle *li#* usually followed by a jussive imperfective: e.g., *li#na+ktub*, “let us write”. One must also add the compound PCLs /*wa#l#*/ and /*fa#l#*/
- the future marker *sa#*, e.g., *sa#na+ktub*, “we shall write”.

2.3.6. *Nominal PCLs*

These include:

- the comparative tool-word *ka#* “as, like”;
- the 1-consonant prepositions *bi#* “with, along, by...,” *li#* “to, for...”
- the article *al#*;
- 1-consonant tool-words, mainly found in ancient poetry, *wa#*, *wāw rubba*, meaning “many-a,” e.g.,
 - *wa#layl+in ka#mawj+I l#baḥr+i...*, “Many-a night alike sea-waves...,”
- and the vocative particles /^ʔa/, /^ʔā/, e.g.,
 - ^ʔa#*fātima mabl+an...*, “O Fatima, slowly...” (examples from Imruʔu al-Qays’s Ode).

2.3.7. *Category Selection*

PCLs mentioned in the two paragraphs above select the major category (noun, verb) they can appear with. Verbal PCLs also select the conjugation paradigm (imperfective, jussive, and imperative).

These selections pertain to the word-formative grammar.

2.4. Enclitics (ECLs)

These are the attached complement pronouns. Verbal ECLs can only appear with transitive verbs, and highlight the necessity of adding a morpho-syntactic specifier indicating that:

- either the verb is only transitive to non-human complements, e.g., *qara'a*, “to read” seen above. This excludes pronouns referring to humans, such as the masculine plural *#bum*;
- or that the transitive verb admits human complements, in which case all ECLs can appear.

2.5. Double Enclitics

Some verbs admit two direct complements, e.g.,

- *ʔaʕtā*, “to give” which admits *ʔaʕtayt+u#ka#bu*, “I gave it to you,” or the performative marriage formula: *zawwaj+tu#ka#bā*, “I give her in marriage to you”.

Sequences of double enclitics are restricted for euphonic reasons: one cannot find, e.g., **ʔaʕtayt+u#ba#ba*. These restrictions are also included in the word-formatives grammar.

3. Lexicalized formatives (Dichy, 1997)

Some formatives attached to nominal nuclei present what can be described as a cluster structure usually occurring in proper nouns, e.g., *al#Tayyib* (masculine example); *al#Zabrāʔ* (feminine example). Another example is that of *al#Babr+ayn*, the name of Bahrein, originally meaning “the two seas”.

4. Conclusion

Language specifiers restrict the number of items in the generated lexicons of the language. Let us take, as an example, the DIINAR database. The number of *extant* entries or nuclei is 131,087, which generates 7,764,937 *attestable* inflected minimal word-forms, instead of between 45 and 70 million virtual ones with a blind generation system¹¹ (Dichy et al., 2004; 2005). The breakdown is the following:¹²

11. I do not present here a critical view of ancient and even current research in that domain.

12. Column 3 stands for “GENERATED LEXICA of minimal words (inflected forms)”.

TABLE 1. Number of entries in the generated lexica of DIINAR, for minimal word-forms

	Stems or nuclei	GEN. LEX.
Verbs	19,457	3,050,715
Deverbals = immediate verb-derived nominal forms, such as: مصادر، أسماء الفاعل والمفعول، صفات مشبهة، إلخ	70,702	2,909,772
Nouns and adjectives (including broken plurals)	39,099	1,781,316
Proper names (limited prototype)	1,384	11,403
Tool-words (provisional figure)	445	11,731
Total	131,087	7,764,937

In this database, the morpho-syntactic specifiers accounting for the relations between the nucleus of the word-form designed by me (Dichy, 1990; 1997) and listed in the present paper, have been added to each of the entries. These specifiers restrict the number of minimal words (i.e., of entries with their prefix and/or suffixes) generated from the database to 7,764,937. All forms included in the generated lexica of DIINAR are *extant* words. On the contrary, an unconstrained combination of all word formatives with the total number of nuclei in the figure above would lead to a generated lexicon of 'virtual' maximal word-forms amounting to 57,642,863,616! Limiting the figures to minimal word-forms, the combination would still yield:

$$8 \text{ prefixes} \times 67 \text{ suffixes} \times 131\,087 \text{ stems} = 70,262,632 \text{ 'virtual' forms}^{13}$$

This shows the efficiency of the representation implemented in the DIINAR dB in restricting the analyses to real Arabic data.

It is essential to note that restricting general lexica is crucial in the natural language processing of languages (NLP), because the related software needs it to:

- operate correctly in recognition, i.e., to recognize extant word-form, rejecting others, even if they can be generated virtually;
- produce correct word-forms included in the mental lexicon, which implies that the morphosyntactic specifiers presented in this paper should be considered and integrated.

13. The consistency of these figures can be easily explained based on the analysis of the word-form and of its constituents (Dichy, 1990). Complementary word-analysis software was developed through 5 doctoral dissertations supervised or co-supervised by me between 1998 and 2010 (see note 10).

The above description shows the consistency of the constraints on the structure of word-forms in Arabic and partly explains the process of readability in unvowelled writing and the related mental lexicon in that family of Semitic languages.

References

- Abraham, Maryvonne (1995). "Contribution méthodologique à la constitution d'un dictionnaire informatique des verbes : analyse sémantico-cognitive des verbes de mouvement et d'activité". PhD thesis. Paris: EHESS.
- Ammar, Sam and Joseph Dichy (1999). *Bescherelle des verbes arabe*. Édition revue et actualisée 2008. Paris: Hatier.
- (2021). *20001 verbes arabes. Dictionnaire de conjugaison, précédé des principes de classification et de modélisation de conjugaison*. Beyrouth: Librairie du Liban.
- Bayyūmī, Ḥasan, Kalfat Khalīl, and Aḥmad al-Shāfiʿī (1989). بيومي، حسن، خليل كلفت، أحمد الشافعي. معجم تصريف الأفعال العربية [Dictionary of Arabic Verb Conjugations]. دار إلياس العصرية: القاهرة [Cairo].
- Braham, Abdelfattah and Salem Ghazali (1998). "ابراهيم، عبدالفتاح وسالم غزالي". قاعدة البيانات المعجمية العربية أو مشروع معجم العربية الآلي (معالي—DIINAR)، حصيلة وآفاق [Arabic Lexical Database or Automated Arabic Dictionary Project]. In: المجلة العربية للعلوم [Arab Journal of Science] 32, pp. 14–23.
- Chomsky, N. (1972). *Studies on Semantics in Generative Grammar*. Trad. fr. *Questions de sémantique*, Paris, Le Seuil, 1975. La Haye: Mouton.
- Cohen, David (1970). "Essai d'une analyse automatique de l'arabe". In: *Études de linguistique sémitique et arabe*. Originally published in *T.A. informations*, 1961. Paris: Mouton.
- Deroche, François (2004). *Le livre manuscrit arabe. Préludes à une histoire*. Paris: Bnf.
- Desclés, Jean-Pierre (1994). "Relations casuelles et schèmes sémantico-cognitifs". In: *Langages* 113. *Les relations actanciels. Sémantique, syntaxe, morphologie*, pp. 113–125.
- Desclés, Jean-Pierre et al. (1982). *Conception d'un synthétiseur et d'un analyseur morphologiques de l'arabe, en vue d'une utilisation en Enseignement assisté par Ordinateur*. Tech. rep. Rapport rédigé sous la direction de J.-P. Desclés pour le Ministère français des Affaires étrangères (sous-direction de la Politique linguistique).
- Dichy, Joseph (1990). "L'écriture dans la représentation de la langue : la lettre et le mot en arabe". 2 vol. Doctorat d'État. Université de Lyon.

- Dichy, Joseph (1993). "Knowledge-system simulation and the computer-aided learning of Arabic verb-form synthesis and analysis". In: *Processing Arabic Report* 6/7, pp. 67–84, 92–95.
- (1997). "Pour une lexicomatique de l'arabe : l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot". In: *Meta* 42.2, pp. 291–306. URL: <http://www.erudit.org/revue/meta/1997/v42/n2/002564ar.pdf>.
- (2005). "Spécificateurs engendrés par les traits [$\pm$ ANIME], [$\pm$ HUMAIN], [$\pm$ CONCRET] et structures d'arguments en arabe et en français". In: *De la mesure dans les termes. Hommage à Philippe Thoiron*. Ed. by Henri Béjoint and François Maniez. Presses Universitaires de Lyon, pp. 152–181.
- (2022). "Semitic Writings and Short Vowels: Alternative Hypotheses in a Renewed View of the Analytics of Writing". In: *Proceedings of Grapholinguistics in the 21st Century, 2022*. Ed. by Yannis Haralambous. Vol. 9. Grapholinguistics and Its Applications. Brest: Fluxus Editions, pp. 377–390. DOI: <https://doi.org/10.36824/2022-graf-dichy>.
- Dichy, Joseph and Ali Ferghali (2003). "Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical database centred on Arabic be built?" In: *Proceedings of the Workshop on Machine Translation for Semitic Languages: Issues and Approaches*. IXth Machine Translation Summit, New Orleans, 23–27 September 2003, pp. 1–8.
- Dichy, Joseph and Mohamed Hassoun (2005). "The DIINAR.1-«معالي» Arabic Lexical Resource, an outline of contents and methodology". In: *The ELRA Newsletter* 10.2, pp. 5–10.
- Dichy, Joseph et al. (2002). "La base de connaissances linguistiques DIINAR.1". In: *Colloque international sur le traitement automatique de l'arabe*. Ed. by Abdelfattah Braham. La Manouba-Tunis: Université de La Manouba, pp. 45–56.
- Fassi Fehri, Abdelkader (2012). *Key Features and Parameters in Arabic Grammar*. The Netherlands/Philadelphia: John Benjamins.
- Frost, R., A. Deutsch, and K. Forster (2000). "Decomposing morphologically complex words in a non linear morphology". In: *Journal of Experimental Psychology: Learning, Memory and Cognition* 26, pp. 751–765.
- Frost, R., K. Forster, and A. Deutsch (1997). "What can we learn from the morphology of Hebrew? A masked priming investigation of morphological representation". In: *Journal of Experimental Psychology: Learning, Memory and Cognition* 23, pp. 829–856.
- Ghenima, Malek (1998). "Analyse morpho-syntaxique en vue de la voyelisation assistée par ordinateur des textes écrits en arabe". PhD thesis. ENSSIB/Université Lyon.

- Grainger, Jonathan et al. (2003). "Approche expérimentale de la reconnaissance du mot écrit en arabe". In: *Dynamiques de l'écriture : approches pluridisciplinaires*. Ed. by Jean-Pierre Jaffré. Vol. 22, pp. 77–86.
- al-Hāḡib, Ibn (1975). "ابن الحاجب. الشافية في علمي التصريف والخط" [al-Shāfiya on the Two Sciences of Morphology and Orthography]. In: رضي الدين [Commentary on al-Shāfiya]. Ed. by Nūr al-Ḥasan et al. 4 vols. بيروت [Beirut]: [House of Scientific Books].
- Jinnī, Ibn (1985). "ابن جني. سر صناعة الإعراب، تحقيق حسن هندراوي" [The Secret of the Craft of Desinential Inflection]. Ed. by Hassan Hindāwī. 2 vols. دمشق [Damascus]: دار القلم [House of the Qalam].
- Mel'čuk, I., A. Clas, and A. Polguère (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.
- Mel'čuk, I. and et al. (1984). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I*. Presses de l'Université de Montréal.
- Mel'čuk, Igor (1982). *Towards a Language of Linguistics, A System of Formal Notions for Theoretical Morphology*. München: Wilhelm Fink Verlag.
- Osman, Nabil (1995). *Konjugation Lexikon Arabische Verben*. München: Max Hueber Verlag.
- Ouersighni, Riadh (2001). "A major offshoot of the DIINAR-MBC project: AraParse, a morpho-syntactic analyzer of unvowelled Arabic texts". In: *ACL 39th Annual Meeting, Workshop on Arabic Language Processing: Status and Prospect*. Toulouse, pp. 66–72.
- (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". PhD thesis. ENSSIB/Université de Lyon.
- Raheel, Saeed (2010). "Apprentissage artificiel et fouille de données multilingues : le cas des documents arabes". Doctorat en Sciences de l'Information et de la Communication. Université de Lyon.
- Raheel, Saeed and Joseph Dichy (2010). "An Empirical Study on the Feature's Type Effect on the Automatic Classification of Arabic Documents". In: *Computational linguistics and intelligent text processing, Proceedings of the 11th international conference, CICling 2010*. Ed. by Alexander Gelbukh. Iași, Romania, March 2010. Berlin/Heidelberg/New York: Springer-Verlag, pp. 673–686.
- Reig, Daniel (1983). *La conjugaison de la langue arabe*. Paris: Maisonneuve et Larose.
- Roman, André (2011). *Grammaire systématique de la langue arabe*. Paris: L'Harmattan.

-
- Sheidlin, Raymond (1978). *201 Arabic verbs fully conjugated in all the forms*. New York: Barron's Educational Series.
- Zaafrani, Riadh (2002). "Développement d'un environnement interactif d'apprentissage avec ordinateur de l'arabe langue étrangère". PhD thesis. ENSSIB/Université de Lyon.