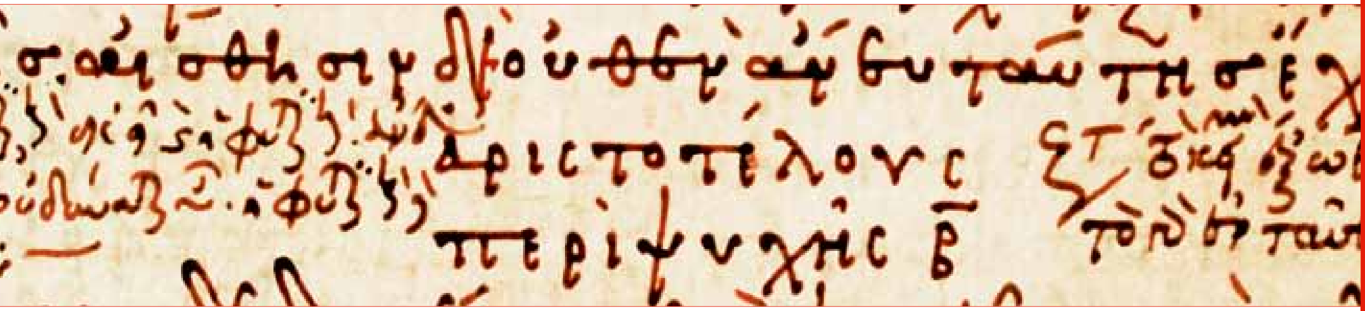
ISSN 2239-012X

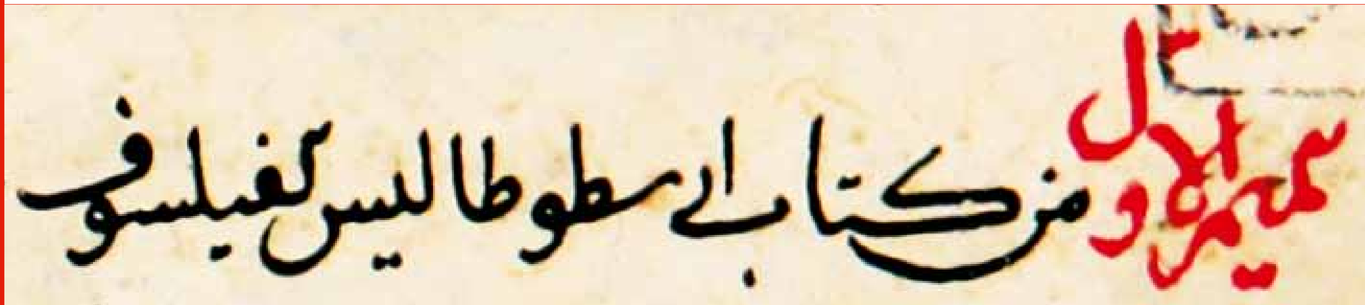


With the support of the European Research Council

Pacini Editore

Studia graeco-arabica

Studia graeco-arabica



3

2013

Studia graeco-arabica

The Journal of the Project
Greek into Arabic
Philosophical Concepts and Linguistic Bridges
European Research Council Advanced Grant 249431

3

2013

Published by
ERC Greek into Arabic
Philosophical Concepts and Linguistic Bridges
European Research Council Advanced Grant 249431

Web site: http://www.greekintoarabic.eu
Service Provider: Università di Pisa, Area Serra - Servizi di Rete di Ateneo

ISSN 2239-012X

Studia graeco-arabica cannot be held responsible for the scientific opinions of the authors publishing in it.

Registered at the law court of Pisa, 18/12, November 23, 2012.
Editor in chief Cristina D'Ancona.

Cover
Mašhad, Kitābḫāna-i Āsitān-i Quds-i Raḍawī 300, f. 1v
Paris, Bibliothèque Nationale de France, grec 1853, f. 186v

The Publisher remains at the disposal of the rightholders, and is ready to make up for unintentional omissions.

Publisher and Graphic Design
Pacini
Editore
Via A. Gherardesca
56121 Ospedaletto (Pisa) - Italy

Printing
Industrie Grafiche Pacini

Studia graeco-arabica

3

2013

G2A Web Application

Istituto di Linguistica Computazionale "Antonio Zampolli"
Consiglio Nazionale delle Ricerche - Area della Ricerca di Pisa

Part II. Lisān al-ʿarab as a source of lexical and morphological knowledge

Ouafae Nahli, Emiliano Giovannetti

1. Introduction

The following sections illustrate a part of the study on the morphology of the Arabic language which is carried on within the framework of the ERC project *Greek into Arabic. Philosophical Concepts and Linguistic Bridges* ADG 249431. We used the Arabic lexicographic encyclopaedia *Lisān al-ʿarab* and, thanks to the regularity of its structure, we developed a system for the extraction of morphologically labelled word sequences, to be exploited for morphological analysis purposes.

2. Elements of Arabic morphology

We would like to present some peculiarities of the Arabic lexicography that we used (together with Arabic morphologic characteristics), in order to pinpoint some structured regularities of the Arabic lexicographic encyclopedia *Lisān al-ʿarab*. This allowed us to perform automatic extractions of lemmas and of their respective irregular forms.

In the dictionaries of the Arabic language, the order of the words is not established from the first letter of each word, as for many other languages. Instead, every word is sorted on the basis of its radical consonants. Arabic lemmas are represented using canonical forms:

- nouns in their masculine or feminine singular form;
- adjectives in their masculine singular form;
- verbs in the third person masculine singular form of the perfect.

When words are not formed through standard production rules, they must be explicitly included in the dictionary. For example:

• the triliteral verb (or basic verb), derived from triliteral roots $C_1C_2C_3$ (where ‘C’ stands for consonant), presents a very complicated morphology:

- the middle vowel of the perfect verb has a semantic value;[1]
- the middle vowel of the imperfect verb varies according to phonetic parameters[2] and it is necessary to indicate it in the dictionary;
- the verbal noun (the maṣdar) can be formed according to different patterns and some primitive verbs have several verbal nouns, depending upon the different meanings of the basic verb. For this reason, the verbal noun is usually listed together with the verb in the dictionary.

• There are no general rules for the formation of the “broken” plural, but a series of schemas that casually can be applied to nouns and adjectives. Moreover, some singular nouns have several broken plurals. The word أَسَد /*’asad*/ “lion”, for example, has the plurals “lions”: آساد /*’āsād*/ - آسُد /*’āsud*/ - أُسُود /*’usūd*/ - أُسُد /*’usud*/ - أُسْد /*’usd*/ - أُسْدان /*’usdān*/. For this reason the broken plurals are explicitly indicated in the dictionary.

[1] Muḥammad ibn al-Ḥasan al-Astarābādī, *Šarḥ šāfiyya Ibn al-Ḥāǧib*, ed. by M. Nūr al-Ḥasan, Dār al-kutub al-ʿilmiyya, I-IV, Beirut 1975-1982, I, p. 67-75.

[2] al-Astarābādī, *Šarḥ šāfiyya Ibn al-Ḥāǧib*, I, p. 114-38 Nūr al-Ḥasan.

3. *Lisān al-ʿarab*

The Arabic lexicographic encyclopedia *Lisān al-ʿarab* (لسان العرب "The Language of the Arabs"), by Ibn Manẓūr (1232-1311), follows an alphabetical order according to the last radical consonant. All the roots sharing the same last radical are sorted on the basis of the standard alphabetical order. For instance, the root عبد, transliterated "E-b-d",[3] (where 'E' stands for the first, 'b' for the second, and 'd' for the third radical consonant), is found under the letter 'd', with the internal ordering determined by the consonants 'E' and then 'b'. One of the peculiarities of *Lisān al-ʿarab* is the rigid structure through which every lemma and its relative irregular forms are organized into sequences along the whole dictionary. Examples of sequences we have detected are:

verbal sequence: perfect - imperfect – masḍar 1 – masḍar 2, ...
وعَبَدَ اللهَ يَعْبُدُه عِبادَةً ومَعْبَداً ومَعْبَدَةً

nominal sequence: noun (or adjective) - broken plural (or broken plurals)
وَرَجُلٌ عَابِدٌ مِنْ قوْمٍ عَبَدَةٍ وَعُبُدٍ وَعُبَّدٍ وَعُبَّادٍ

By isolating and analyzing these sequences it is possible to distinguish between verbs and nouns and to identify the inflection of lemmas and all the irregular forms.

4. *"Form Extractor" system*

The digital version of *Lisān al-ʿarab* (henceforth DL, *Digital Lisān*) was downloaded from the Internet. It is composed of 29 Microsoft Word documents, each one corresponding to a letter. All the files were converted in plain text format (TXT); then, they have been assembled in a single file. *DL* basically includes a set of roots: each roots is followed by a description that contains the derived forms of the root. Roots are formed by a number of letters which are comprised between one (the letters of the alphabet) and six (sextiliteral roots).

The "Form Extractor" system, developed in Java, allows for the automatic extraction of the roots and the relative irregular forms contained in the *DL* resource. As we have mentioned above, the description of each entry (i.e. of each specific Arabic root) follows a rigorous schema. For this reason, *DL* relies heavily on the use of regular expressions.

After a preliminary work of correction of some errors typical of *DL*, all the roots have been extracted and classified on the basis of the number of their consonants: triliteral, quadriliteral and quintiliteral roots (sextiliteral roots, very few in number, have been analyzed one by one). Within the set of triliteral roots we have also singled out "irregular" roots, like those which contain a *hamza* and/or a *yāʾ* and/or a *wāw*, as well as does in which the second consonant is the same of the third.

We focused on the extraction of forms which derived from triliteral roots. This process has been divided into two steps. First, using a battery of regular expressions, all the roots and the respective derived forms have been extracted. Although the extraction of roots proved to be an easy task, the detection of derived forms was far more challenging and required the construction of complex regular expressions, some of which have not yet been fully developed.

[3] To facilitate computer processing of the DL resource, we have adopted the transliteration system developed by Tim Buckwalter (see Nahli, "Computational Contribution", Part I, n. 12).

See an example of rules for the extraction of derived forms (and its instantiation for root ‘Ebd’) in Table 1, where, for example, a proclitic (if present) can be a particle (or a combination of particles) compatible with the verb or with the noun.

Tab. 1. First step: rule instantiated for the extraction of the forms of the root عبد (transliterated ‘Ebd’).

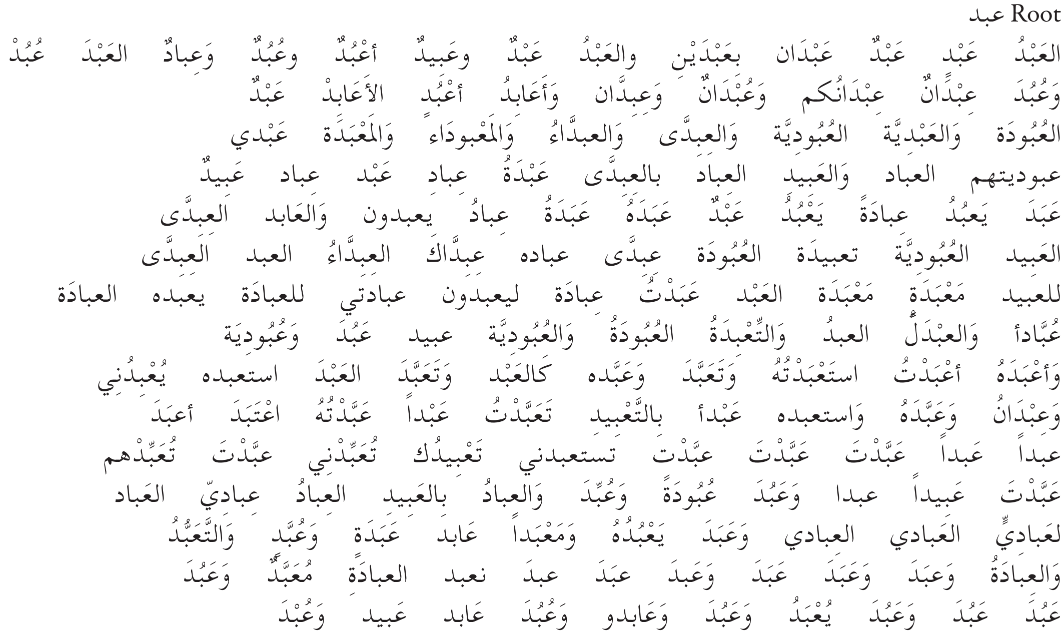
[(proclitic)*(prefix)*]**C1**(infix)**C2**(infix)**C3**[(sufix)(enclitic)*]

⇓

[(proclitic)*(prefix)*]**E**(infix)**b**(infix)**d**[(sufix)(enclitic)*]

Table 2 shows a screenshot of the first step, resulting in a ‘match’ of the instantiated expression of Table 1. The root ‘Ebd’ is followed by all its derived forms which feature in its description.

Tab. 2. Screenshot of the extraction of the root عبد “Ebd”.



Root عبد

العَبْدُ عَبْد عَبْدٌ عَبْدَان بعَبْدَيْن والعَبْدُ عَبْدٌ وعَبيدٌ أعْبُدٌ وعُبُدٌ وَعِبادٌ العَبْدَ عُبُدْ
وَعُبُدَ عِبْدًانٌ عِبْدَانُكم وَعُبْدَانٌ وَعِبِدَّان وَأَعَابِدُ أعْبُد الأَعَابِدْ عَبْدٌ
العُبُودَة وَالعَبْدِيَّة العُبُودِيَّة وَالعِبِدَّى وَالعِبدَّاءُ وَالمَعْبوداءً وَالمَعْبَدَة عَبْدي
عبوديتهم العباد وَالعَبيد العبادَ بالعبدَّى عَبْدَةُ عِباد عَبْد عباد عَبيدٌ
عَبَدَ يَعبُدُ عبادَةً يَعْبُدُ عَبْدٌ عَبَدَهُ عَبَدَةُ عِبادُ يعبدون وَالعَابد العبدَّى
العَبيد العُبُودِيَّة تعبيدَة العُبُودَة عبدَّى عباده عِبدَّاك العبدَّاءُ العبد العَبدَّى
للعبيد مَعْبَدَة مَعْبَدَة العَبْد عَبَدْتُ عِبادَة ليعبدون عبادَتي للعبادَة يعبده العبادَة
عُبَّادأ وَالعبْدَلُّ العبدُ وَالتِّعْبدَةُ العُبُودَةُ وَالعُبُودِيَّة عبيد عَبُدَ وَعُبُودِيَة
وَأعْبَدَهُ أعْبَدْتُ استَعْبَدْتُهُ وَتعَبَّدَ وَعَبَّده كَالعَبْد وَتعَبَّدَ العَبْدَ استعبده يُعْبِدُنِي
وَعِبْدَانُ وَعَبَّدَهُ وَاستعبده عَبْدأ بالتَّعْبيد تَعَبَّدْتُ عَبْداً عَبَّدْتُهُ اعْتَبَدَ أعبَدَ
عبداً عَبداً عَبَّدتَ عَبَّدْتَ عبَّدْت تَستعبدني تَعْبيدُك تُعَبِّدْنِي عبَّدْتَ تُعَبِّدْهم
عَبَّدْتَ عَبيداً عبدا وَعَبُدَ عُبُودَةً وَعُبِّدَ وَالعبادُ بالعَبيد العبادُ عِبادِيّ العَباد
لعَبادِيٍّ العَبادي العبادي وَعَبَدَ يَعْبُدُهُ وَمَعْبَداً عَابد عَبَدَة وَعُبَّد وَالتَّعَبُّدُ
وَالعبادَةُ وَعبدَ وَعَبَدَ عَبَدَ وَعبدَ عبدَ عبدَ نعبد العبادَةً مُعَبَّدٌ وَعَبُدَ
عَبُدَ عَبُدَ وَعَبُدَ يُعْبَدُ وَعَبُدَ وَعَابدو وَعُبُدَ عَابد عَبيد وَعُبْدَ

In the second step we applied a ‘second level’ of regular expressions to the previous output, with the aim of extracting specific sequences of words.

In this case, it is possible to associate a specific label to each regular expression (the so-called “named capture groups”) so that, when a battery of different expressions is applied to the output of the first step, the system can indicate each match. As a result, the system can extract specific word sequences labeled with morphological information for each root.

Table 3 shows the extraction of triliteral perfect verbs derived from root عبد ‘Ebd’ through a regular expression of second level.

Tab. 3. Example of extracted sequence of the primitive verbs derived by the root عبد ‘Ebd’.

Ebd عبد (Root)
عَبُدَ (Perfect_Verb) عُبُودَةً (masḍar) عُبُوديَّةً (masḍar)
عَبَدَ (Perfect_Verb) يَعْبُدُهِ (Imperfect_Verb) عبَادَةً (masḍar) مَعْبَداً (masḍar) مَعْبَدَةً (masḍar)
عَبدَ (Perfect_Verb) عَبَداً (masḍar) عَبَدَةً (masḍar) عَابدٌ (adjective)
عَبِدَ (Perfect_Verb) يَعْبَدُ (Imperfect_Verb) عَبِدٌ (adjective)

See another example in Table 4, which refers to the extraction of all the derived verbs through another regular expression of “second level”.

Tab. 4. Example of extraction of derived verbs.

Ebd عبد (Root)
عَبَّد (Verb-II)
أعْبَدَ (Verb-IV)
تَعَبَّدَ (Verb-V)
استَعْبَدَ (Verb-X)
Eqd عقد (Root)
عَقّد(Verb-II)
عَاقَدَ (Verb-III) مُعَاقدَة (masḍar III) عِقَاد (masḍar III)
أعْقَدَ (Verb-IV)

5. *Conclusions*

A lexicon lies at the core of most morphological analyzers of Arabic. The validity of a morphological analyzer is strictly dependent on the quality and coverage of its lexical database. For this reason, we have focused on the acquisition, structuring and storage of Arabic lexical resources and on their application for Natural Language Processing tasks.

The strictness of morphology and grammar of the Arabic language, which is masterfully represented in the *Lisān al-ʿarab*, encouraged us to engage in the challenging task of mining from such an old and remote source. The sequences extracted as described above can be used in various ways to support the morphological analysis of Arabic, either as standalone resource, or to enrich quantitatively and qualitatively the extant resources, both for classical and modern Arabic.

The work conducted so far on *Lisān al-ʿarab* will contribute to the creation of a lexicon of classical Arabic to be used for the analysis of ancient texts. In order to create lexical resources of modern Arabic, however, it is necessary to discard the obsolete forms and make sure that the acquired lexical items reflect the modern use of the language. Moreover, for modern Arabic it is necessary to take into account the statistical information about word frequencies extracted from corpora of modern Arabic belonging to different domains.[4] This extractive approach can be applied with relative ease to other dictionaries, both of classical and modern Arabic, adapting the set of regular expressions depending on the structure of the internal representation of the dictionaries.

[4] M.A. Attia - P. Pecina - L. Tounsi - A. Toral - J. van Genabith, “Lexical Profiling for Arabic”, in I. Kosem - K. Kosem (eds.), *Proceedings of eLex* (10-12 November 2011), Institute for Applied Slovene Studies, Ljubljana 2011, p. 23-33.

Finito di stampare nel mese di settembre 2013
presso le Industrie Grafiche della Pacini Editore S.p.A.
Via A. Gherardesca • 56121 Ospedaletto • Pisa
Tel. 050 313011 • Fax 050 3130300
www.pacinieditore. it

AZIENDA CON SISTEMA QUALITÀ CERTIFICATO
PACINIeditore
★ ISO 9001 ★