

# **Words Lost and Found**

## **The Diachronic Dynamics of the Arabic Lexicon**

by

**Petr Zemánek and Jiří Milička**

**2017**

**RAM-Verlag**

# Studies in Linguistics

ISBN: 978-3-942303-45-3

© Copyright 2017 by Petr Zemánek, Jiří Milička

© Copyright 2017 by Charles University, Faculty of Arts

© Copyright 2017 by RAM-Verlag, D-58515 Lüdenscheid

RAM-Verlag

Stüttinghauser Ringstr. 44

D-58515 Lüdenscheid

[RAM-Verlag@t-online.de](mailto:RAM-Verlag@t-online.de)

<https://ram-verlag.deu>

## Acknowledgement

This publication has been prepared and published with the financial support of the grant provided by the Czech Science Foundation (GAČR), no. 13-28220S, entitled “Structure of culture: Arabic and Islamic culture in the Prism of the Corpus Linguistics”, implemented at the Faculty of Arts, Charles University

---

# Contents

<b>1. Introduction</b> .....	<b>1</b>
<b>2. Background: methods and data</b> .....	<b>3</b>
The data used .....	4
Lemmatization .....	10
<b>3. On the history of Arabic</b> .....	<b>13</b>
The assumed conservative character of Arabic .....	13
The position of the Arabs .....	16
Historical phases .....	17
Pre-Islamic phase .....	21
The beginnings of Classical Arabic .....	23
The ritual side .....	24
The administrative part .....	25
The Greek connection .....	25
The subsequent centuries .....	27
<b>4. The ways of lexical innovation in Arabic</b> .....	<b>33</b>
Some inner developments within the lexicon .....	35
<b>5. Frequencies of synsemantic words</b> .....	<b>39</b>
Conjunctions .....	39
Particles .....	44
Prepositions .....	51
Summary for conjunctions, particles and prepositions .....	60
<b>6. Words found, words lost</b> .....	<b>65</b>
Words found .....	65
Words lost .....	80
The role of dictionaries .....	82
<b>7. Key words across the centuries</b> .....	<b>86</b>
<b>8. Conclusion</b> .....	<b>137</b>
<b>Appendix</b> .....	<b>142</b>
<b>Index</b> .....	<b>160</b>
<b>References</b> .....	<b>164</b>

## Transcription

Arabic grapheme	Transcription	Characteristics
أ، ء	ʔ	hamza, glottal stop
ب	b	voiced bilabial occlusive
ت	t	non-voiced alveolar occlusive
ث	t̪, th	non-voiced interdental fricative
ج	ɡ̟	voiced postalveolar affricate
ح	ħ	non-voiced pharyngeal fricative
خ	ħ	non-voiced velar fricative
د	d	voiced alveolar occlusive
ذ	d̪, dh	voiced interdental fricative
ر	r	voiced alveolar fricative trill
ز	z	voiced alveolar fricative
س	s	non-voiced alveolar fricative, sibilant
ش	ʃ	non-voiced postalveolar fricative, sibilant
ص	ʂ	non-voiced alveolar pharyngealized fricative, sibilant
ض	ɖ	voiced alveolar pharyngealized fricative
ط	ɸ	non-voiced alveolar pharyngealized occlusive
ظ	ʒ	voiced interdental pharyngealized fricative
ع	ʕ	voiced pharyngeal fricative
غ	ɡ̠	voiced velar lateral trill
ف	f	non-voiced labiodental fricative
ق	q	non-voiced postvelar (uvular) occlusive

ك	k	non-voiced velar occlusive
ل	l	voiced alveolar lateral fricative
م	m	voiced bilabial nasal occlusive
ن	n	voiced alveolar nasal occlusive
ه	h	voiced glottal fricative
و	w	voiced bilabial fricative, semivowel
ي	y	voiced palatal fricative, semivowel

**Remark:**

Due to technical reasons, in some charts and graphs a simplified transcription is used. In the legends, a correct transcription is provided.

# 1 Introduction

This current volume examines the Arabic lexicon from the point of view of changes on a chronological axis. Our approach is both lexical and dynamic, as there are several points that address the dynamic behavior of the lexical items and their subsystems. The lexicon is regarded as a wide variety of views, not just a list of words, therefore, we examine the lexical lists from several points of view, with only one connected with words as such. The composition of the words, especially of the phonemes (in our case, consonants), consisting of the sum of the words are also observed. In addition, various samples of words, such as chosen types of parts of speech, and morphological classes, such as plurals, are taken into consideration.

At the same time, we work with quantitative data. Based on a corpus of Arabic that covers all the historical phases of Arabic (cf. below), we were able to acquire data from the exact historical phases of the use of Arabic in written records, and we have used this data as the basis of our considerations.

We only consider such lexemes that exhibit a frequency of at least 20 occurrences in the whole corpus. This threshold is set up arbitrarily, but takes into consideration the time span and number of chronological divisions, so that there is still a chance that the distribution of the occurrences is not only singulative and concentrated only into one work or in titles by only one author.

The behavior of the sets of lexemes is investigated by applying the hypothesis that there ought to be some general patterns that enable to see the type of changes that have occurred in Arabic throughout the ages.

We are interested in relations between frequency and stability, which are often posited as those that influence the possible survival of a given lexeme in the lexicon. At the same time, the usage of highly frequent words can have variations across time. This will also be one of our points of interest.

The loss of words or creation of new words is a process that occurs in all languages. Arabic is certainly no exception, although, in the case of classical Arabic, there are many signs of an increased conservative character.

The question of the characteristic words for a given century was also investigated: a keyness metric based on the *Relative Risk* was used to obtain the keywords for each century. The word distribution in each century was compared with the rest of the corpus and the most overrepresented words were retrieved. For this purpose, we have used the *Average Reduced Frequency* (ARF) of the words instead of the plain word frequency, which was

## *Introduction*

necessary to sort out words that are locally frequent but marginal from a broader point of view, i.e. word types that are only prominent in one or two texts and scarce in the rest of the sample. A detailed description of these metrics can be found in Chapter 7.

Some of the data we obtained is available in the appendix. Namely, there are words identified as loans, followed by an overview of names identified as Turkic.

Authors' contribution: Jiří Milička prepared the raw data for the study, designed, implemented and described the method of keyword extraction, prepared the lemmatization and the metric. Petr Zemánek checked the raw data and formulated the vast majority of the text of the study.

We are grateful to a number of students and other co-workers who took part in the preparation of the data, namely Adéla Provazníková, Martina Hainová, and Michal Lázníčka.