



Ain Shams University
Faculty of Al-Asun
Department of English



A Corpus-Based Descriptive Approach to Build a Bilingual Lexicon of the Egyptian Colloquial Arabic Words on Social Media Platforms and Their English Equivalents

A thesis submitted in the fulfillment of the requirements of MA in Translation

By

Bacem Abdullah Essam

Supervisors

Dr. Mostafa Mahmoud Aref

Professor of Computer Science

Faculty of Computer Science and Information Sciences,

Ain Shams University

Dr. Fayrouz Fouad

Lecturer in Linguistics

Department of English Language,

Faculty of Al-Asun,

Ain Shams University

2018

DEDICATION

TO

My Wife

ACKNOWLEDGMENTS

Essentially acknowledging the formidable effort my supervisors exerted, I love to thank **Prof. Mostafa M. Aref** and **Dr. Fayrouz Fouad** for their professional guidance and lenient instruction. My gratitude is extended to my examiners: **Prof. Mohsen Rashwan** and **Associate Prof. Nagwa Younis** for spending valuable time and effort in reviewing my work and suggesting useful comments for improving it.

I would love to give remarkable credit to the two computational linguists: **Professors Nicoletta C. Zamorani** and **Sara Goggi** (*National Research Council, Pisa, Italy*) for their valuable tips as regards detailing and standardizing the methodology. Equally important, ineffable gratefulness is aired to **Dr. Eshrag Refaee** (*Heriot-Watt University, Edinburgh*) for enriching my corpus by sending me her own annotated and automatically-extracted Arabic corpora and to the American and British translators for reviewing the colloquial translation.

List of Abbreviations

Abbreviation	Stands for
ALP	: Arabic Living Project
AmE	: American English
AO	: Arabic Ontology
AWN	: Arabic WordNet
BrE	: British English
CCM	: Cross-cultural matching
CECA	: Contemporary Egyptian Colloquial Arabic
FWs	: Functional Words
LIWC	: Linguistic Inquiry and Word Count
NLP	: Natural Language Processing
POS	: Part of Speech
SF	: Semantic Field
SL	: Source Language
SR	: Semantic Relations
SS	: Semantic Similarity
ST	: Source Text
SUMO	: Suggested Upper Model Ontology
TE	: Translation Equivalence
TL	: Target Language
TT	: Target Text
UGC	: User Generated Content
WN	: WordNet

List of Tables

Table	Page
Table 1.1 Examples of the Prominent Translation Techniques	17
Table 2.1 CECA Word specification from the collected wordlist	51
Table 2.2 The most frequent FIXED functional words within the studied CECA corpus.	54
Table 2.3 The most frequent INFLECTIONAL functional words within the studied CECA corpus.	55
Table 2.4 Representation of the unified Arabic hypernyms for some CECA words	57
Table 2.5 Retrieved Gender Information and Tweet Sample.	65
Table 2.6 Roget's Domains of the frequent semantic fields in CECA.	68
Table 2.7 Five-year incidence of CECA headwords	69
Table 2.8 Descriptive analysis of CECA headwords	71
Table 2.9 Descriptive information of English informal words	73
Table 2.10 American Equivalents of CECA words.	75
Table 2.11 British Equivalents of CECA words.	76
Table 3.1 The identified hyponyms of the most frequently tackled domains in the CECA corpus.	82
Table 3.2 The fifteen most frequent domains in the CECA corpus.	84
Table 3.3 Representation of categorizing CECA words, their hypernyms and their semantic fields according to LIWC.	87
Table 3.4 Gender Differences in using CECA words.	92
Table 3.5 American and British equivalents of CECA headwords	95

List of Figures

Figure	Page
Figure 0.1 The proposed method of translating CECA words	12
Figure 1.1. Antonyms of ‘grand ’in WordNet.	29
Figure 1.2. Hyponym-hypernym application in linguistics.	33
Figure 1.3. WN’s relational semantics aligns with the mental lexicon	38
Figure 1.4. Diagramming tenets of compiling bilingual dictionaries.	46
Figure 2.1. Word-Cloud showing Collocates of “ʕal”	58
Figure 2.2. Hypernym rules excerpted from Ath-thalabi’s Arabic philology	59
Figure 2.3. An age-related female illustration excerpted from Ath-thalabi’s Arabic philology.	60
Figure 2.4. Interface of AWN and linked WN.	61
Figure 2.5. Concordance of the CECA word “ bəs ” representing its senses.	63
Figure 2.6 Concordance of the CECA word “ mɪʃ ” in feminine usage.	66
Figure 2.7 Concordance of the CECA word “ mɪʃ ” in masculine usage.	67
Figure 2.8 Macro and microstructure representation in our lexicon.	77
Figure 3.1 The 10 topper domains in the studied CECA corpus	80
Figure 3.2 The interface of LIWC-2015 showing its main categories.	86
Figure 3.3 Concordance of deception-expressing words	90
Figure 3.4 Charting the output of the conspiratorial ideation concluded from the studied CECA corpus	91

Table of Contents

Abstract.....	viii
Introduction.....	1
0.1. Objectives of the Study.....	2
0.2. Significance of the Study.....	2
0.3 Research Questions	2
0.4. Review of the Literature	3
0.5. Methodology	5
0.5.1. Data Collection.....	6
0.5.2. Data Processing Software.....	8
0.5.3. Procedures of Data Processing	12
0.5.4. Chapterization of the Thesis	14
Chapter 1: Theoretical Preliminaries	16
1.1 The Framework of Equivalence and Lexicography.....	16
1.1.1 Rarity of Full and Zero Equivalents.....	20
1.1.2 Domination of Partial Equivalents.....	21
1.2 Semantic Similarity and Equivalence.....	23
1.3 Semantic Relations as Attributes of Equivalence	24
1.3.1 Lexical Semantic Relations and WordNet.....	25
1.3.1.1 Paradigmatic Relations and WordNet.....	26
1.3.1.2 Syntagmatic Relations.....	32
1.4 WordNet: from Psycholinguistics to Lexicography	35
1.4.1 The models of “Mental Lexicon” Organization.....	35
1.4.2 WordNet’s Psycholinguistic Background and Structure.....	36
1.4.3 The Updated View of the “Mental Lexicon” Matches WN.....	41
1.5 Descriptive Corpus Lexicography.....	42
1.5.1 Descriptive Corpus Studies.....	42
1.5.2 Social Media Streams as Sources for User-Generated Contents	45
1.5.3 Lexicography and Organizing Dictionaries	46

Chapter 2: Building the Lexicon	49
2.1 Scope of the Proposed Bilingual Lexicon	49
2.2 Processing Arabic and English Corpora	51
2.2.1 Compiling a CECA Wordlist.....	51
2.2.2 Descriptive Analysis of CECA Headwords	58
2.2.3 Frequency and Normalized Ratios.....	69
2.2.4 Annotating Online English Corpora	74
2.3 Finding Possible Equivalents	76
2.4 Final Organization of the Lexicon.....	79
2.4.1 Macrostructure	79
2.4.2 Microstructure.....	79
Chapter 3: Results, Findings and Outlook	81
3.1 Sociolinguistic Perspectives of the Constructed Lexicon	81
3.1.1 Frequent Semantic Fields and the Egyptian Society	82
3.1.2 Gender Differences in Using CECA Words.....	94
3.2 Translation Findings of the Constructed Lexicon	97
3.2.1 Cultural and Linguistic Differences between CECA and Colloquial English.....	98
3.2.2 The Scope of Equivalence in Colloquial Languages.....	99
3.2.3 Online Translatability of CECA Words	100
Conclusion	103
References.....	106
English Summary.....	115
Appendices.....	117
Arabic Summary	i

Abstract

Essentially adopting a descriptive approach to studying corpora, this thesis constructs an Arabic-English lexicon of Contemporary Egyptian Colloquial Arabic (CECA) words using social media streams. It extracts the most frequent CECA words, from a five-million-word corpus of Egyptian Arabic tweets and posts, published in the span from 2012 to 2016. Within the linguistic approach to translation studies, the constructed lexicon portrays the linguistic context within which a word is authentically used in order to identify the possible equivalents in informal American and British English corpora. On translating CECA words into English, the informal American English is proved to harmonize cross-culturally more to the Egyptian culture than the British English does. Moreover, some sociolinguistic information is revealed through the analysis of the CECA corpus. There seems a striking predilection of the Egyptian societal concerns to sentiment and intellectuals. However, deception, malediction and describing outer shapes come atop. Gender preferences of using CECA words demonstrate a significant stratification.

Keywords: *Translation Equivalence, Colloquial Arabic, Bilingual Lexica, Corpus, Arabic Ontology*