Ain Shams University
Faculty of Al-Alsun
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-Alsun,
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 Biritsh 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

List of Figures

Table of Contents

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*