



Ain Shams University
Faculty of Al-Asun
English Department



A Computational Prototype Theory-Based Study of Hyponym Typicality in Arabic and English Using the Web as Corpus

*A Thesis submitted to the Faculty of Al-Asun in partial fulfillment of the requirements
for the Master Degree in Linguistics*

By

Fadia Ahmed Abdullah Mohamed

Teaching Assistant at the English Department,
Faculty of Asun, Ain Shams University.

Supervisors

Dr. Khaled Elghamry

Associate Professor of Linguistics, English Department,
Faculty of Asun, Ain Shams University.

Dr. Fayrouz Fouad

Lecturer in Linguistics, English Department,
Faculty of Asun, Ain Shams University.



كلية الألسن
قسم اللغة الإنجليزية

أنماط الكلمات ذات الحقل الواحد

دراسة حاسوبية تقابلية بين الإنجليزية والعربية فى ضوء نظرية النموذج الأمثل

من خلال شبكة المعلومات

رسالة مقدمة ضمن متطلبات الحصول على درجة الماجستير فى تخصص اللغويات

مقدمة من

فادية أحمد عبد الله محمد

المشرفان

د.خالد عبد الحميد الغمري

أستاذ مساعد اللغويات، قسم اللغة الإنجليزية، كلية الألسن جامعة عين شمس

د.فيروز فؤاد

مدرس اللغويات، قسم اللغة الإنجليزية، كلية الألسن جامعة عين شمس

2017

ACKNOWLEDGMENTS

After praising ALLAH, The All-Knowing, for enabling me to conduct this study, I would like to acknowledge **Prof. Khaled Elghamry** for his supervision and for his guidance. My acknowledgment extends also to **Dr. Fayrouz Fouad** for her valuable comments and fruitful instructions.

My acknowledgements extend equally to my examiners: **Prof. Khaled Tawfik** (Faculty of Arts, Cairo University) and **Prof. Dr. Nagwa I. Younis** (Faculty of Education, Ain Shams University) for their careful review of my thesis. I also appreciate their patience and leniency they showed towards me. Last but not least, I'd also like to thank Prof. **Fadwa Kamal** (Head of English Department, Faculty of Al-Asun) for her effort in meeting our greatest expectations.

List of Abbreviations

GCM	:	Generalized context model
FM s	:	Family Members
FR	:	Family Resemblance
LDLTAL	:	Longman Dictionary of Language Teaching and Applied Linguistics
PT	:	Prototype Theory
SC	:	Semantic Category

List of Figures

Figure 1. The methodology of the present study	7
Figure 2. Web-based toppler bootstrapped prototypes of emotions in the four studied domains.	40
Figure 3. Web-based toppler prototypes of emotions latticed according to semantic relations ...	42
Figure 4. Web-based toppler bootstrapped prototypes of drinks in the four studied domains	47
Figure 5. Web-based toppler prototypes of drinks latticed according to semantic relations	49
Figure 6. Web-based toppler prototypes of diseases latticed according to semantic relations	55
Figure 7. Web-based toppler prototypes of crime latticed according to semantic relations	57
Figure 8. Web-based toppler prototypes of crime latticed according to semantic relations	64
Figure 9. Web-based toppler bootstrapped prototypes of crime in the four studied domains	62
Figure 10. Web-based toppler bootstrapped prototypes of emotions in the four studied domains	72
Figure 11. Web-based toppler prototypes of emotions latticed according to semantic relation ...	74
Figure 12. Web-based toppler bootstrapped prototypes of drinks in the four studied domains ...	79
Figure 13. FMs of the category drinks	81
Figure 14. Web-based toppler prototypes of diseases latticed according to semantic relations ...	86
Figure 15. Web-based toppler prototypes of crime latticed according to semantic relations	88
Figure 16. Web-based toppler bootstrapped prototypes of crime in the four studied domains	93
Figure 17. Web-based toppler prototypes of crime latticed according to semantic relations	96

List of Tables

Table	Page
Table 1: Example of three-level categorization in Rosch's prototype theory	 21
Table 2. Mean values of the topper prototypes of 'emotions' in the four web domains	 38
Table 3. Mean values of the topper prototypes of drinks in the four web domains	 46
Table 4. Mean values of the topper prototypes of disease in four web domains.	 53
Table 5. Mean values of the topper prototypes of crime in four web domains.	 61
Table 1. Mean values of the topper prototypes of "مشاعر" in the four web domains	 70
Table 7. Mean values of the topper prototypes of drinks in the four web domains	 78
Table 8. Prototypes of أمراض in the four domains in (2010-2016)	 85
Table 9. Mean values of the topper bootstrapped prototypes of crime in four web domains	 92
Table 10. Arabic and English FMs of Emotions	 101
Table 11. Arabic and English FMs of drinks	 104
Table 12. Arabic and English FMs of crimes	 106
Table 13. Arabic and English FM of diseases.	 109

Table of Contents

0. Introduction	1
0.1 Context of Study.....	1
0.2 Study objectives	3
0.3 Hypotheses of the Study.....	4
0.4 Significance of the Study	12
0.5 Sources of data	4
0.6 Description of Data	5
0.7 Methodology	5
0.7.1 Tools of the analysis	6
0.7.2 Procedure of the analysis	8
0.8 Review of literature	8
Chapter One: Theoretical Preliminaries	13
1. Prototype Theory	13
1.1 Cognitive Categorization and Principles of Prototype Theory	15
1.1.1 Levels of Categorization in the Prototype theory	17
1.1.2 Category Members and Family Resemblance	21
1.1.3 The Prototype and the Goodness of Example (Typicality)	23
1.2 Linguistic Instantiation of Prototype Relations.....	24
1.2.1 Lexical Semantics and levels of Categorization in PT	24
1.3 Relevant Theories.....	28
1.3.1 PT as a Response to the Fuzziness Set of the Classical Theory	29
1.3.2 PT and the Exemplar Theory	31
1.4 The Cross-cultural Approach to PT	31

1.4.1 Culture and Language.....	32
1.4.2 Culture and the Prototype Theory	33
Chapter Two: Analysis of the English Data	35
2. Analysis of the English Data.....	35
2.1 Emotions.....	36
2.1.1 Emotions in the Four Domains	36
2.1.2 Prototype and FMs of Emotions	42
2.2 Drinks (beverages)	43
2.2.1 Drinks in the four domains	43
2.2.2 Prototype and FMs of Emotions	49
2.3 Diseases.....	50
2.3.1 Diseases in the four domains	50
2.3.2 Prototype and family members of diseases.....	56
2.4 Crimes	58
2.4.1 Crimes in the four domains	58
2.4.2 Prototype and family members in the category	63
Chapter Three: Analysis of the Arabic Data.....	67
3. Analysis of Arabic Data.....	67
3.1 The category ‘مشاعر’.....	67
3.1.1 The Category ‘مشاعر’ in the Four Domains	68
3.1.2 Prototype and family members.....	74
3.2 The category ‘مشروبات’	76
3.2.1 The category ‘مشروبات’ in the four domains	76
3.2.2 Prototype and Family Members	81
3.3 The category ‘امراض’	82

3.3.1 The category ‘امراض’ in the four domains.....	82
3.3.2 Prototype and family members.....	87
3.4The category ‘جرائم’.....	89
3.4.1 The category ‘جرائم’ in the four domains	90
3.4.2 Prototype and family members.....	95
Chapter Four.....	99
4. Contrastive Analysis	99
4.1 Prototypicality of emotions in Arabic and English	99
4.1.1Conceptual categorization of emotions in Arabic and English	100
4.1.2 Arabic and English FMs of Emotions	101
4.2 Prototypicality of Drinks in Arabic and English.....	102
4.2.1 Conceptual categorization of drinks in Arabic and English	103
4.2.2 Arabic and English FM in Drinks.....	103
4.3 Prototypicality of Crimes in English and Arabic	105
4.3.1 Conceptual categorization of crimes in Arabic and English	105
4.3.2 Arabic and English FMs in Crimes	105
4.4 Prototypicality of Diseases in Arabic and English.....	107
4.4.1 Conceptual categorization of diseases in Arabic and English	108
4.4.2Arabic and English FMs in Diseases	108
Conclusion	111
References.....	116
English Summary.....	120
Arabic Summary	123
Arabic Abstract.....	126

Abstract

Using web-as-corpus, this study explores the prototypicality of four semantic categories: *crimes*, *emotions*, *diseases* and *drinks*, in two corpora: Arabic and English. Theoretically implicating on Rosch's prototype theory (1975), the study aims at quantitatively categorizing the different members of four semantic categories in four web domains: *.com*, *.org*, *.info* and *.edu* over a 7-year span (2010-2016). The study answers several questions including 'how effective is the prototype theory in specifying a category prototypes?', 'are there any differences or similarities in the ranking of the family members of the same concept?', 'Can changing domains affect the conceptualization and categorization?', as well as 'how far does the cultural, or societal, level influence the cognition of a certain concept?'. Results reveal that the prototype of emotions in Arabic is 'خوف' but the English prototype is "anger". 'خوف' in the concordance of the Arabic corpus correlates to theological, political and economic predisposing factors –stipulating custody of thoughts and acts. However, 'anger' is traced in people who are less controlled by 'fear'. However, the cultural difference is almost effaced as regards the best example of 'drinks' in the oriental and occidental conceptualizations: *coffee*. Still, variance exists when it comes to the herbal preferences versus alcoholic predilection of beverages in the centroid area near coffee. Hierarchical semantic clusters have proven rich as regards the cognitive lexicalization of enumerating slew of members under each category. There, synonymous and hyponymic relations represent the most frequent bonds among FMs of the same category. The study concludes that Rosch's prototype theory is applicable in retrieving the synchronic organismic perception of the contemporary categorization at the individual and expert levels. Therefore, the study recommends ushering our proposed methodology towards cognitive semantics and mental lexicon.

Keywords: Prototype Theory, Web-as-Corpus, Categorization, Categorization and Semantic Hierarchical Structures.

Introduction

0.1 Context of Study

Prototypicality and categorization are interlinked concepts imported from cognitive psychology to cognitive linguistics. Categorization involves recognizing a new entity as part of something abstract conceived with other real instances. However, Linguistic categorization is a core conceptual aspect of attributing meanings and features to certain concepts. Emanating from environmental knowledge or personal perception, categorizations shape the taxonomical orientation of the individual semantic memory, at the smaller unit, and the societal conceptualization, at a larger measure (Croft, & Cruse, 2004). Concepts and categories are two elements that cannot be seen separated from each other. As Aitchinson (1987) points out, concepts have a categorization function used for classifying new entities and extracting inferences about them.

Biria and Bahadoran-Baghbaderani (2016) refer to the graded differences among the elements categorized under the same concept or category. Some elements are better examples than other co-elements. Many theories are embraced within prototypicality such as the exemplar theory and the prototype theory. Such elements differ in their representativeness of the named category.

The purpose of the present corpus-based study is to investigate how prototype theory works in defining categories in real life. The theory is introduced by Rosch (1975) in order to explain how semantic categories are represented in our mind. Several experiments prove the functioning of

prototype theory, but in everyday life we often categorize instances based on our culturally bound definitions rather than on similarity to a typical instance.

A category is a set of attributes that share characteristics of groups of people or objects, or “a number of objects that are considered equivalent” (Rosch, & Lloyd, 1978:27). The category plays an important role in word recognition because categorization can enable identification of an object, as ‘people often define a concept by reference to typical instances’ (Longman Dictionary of Language Teaching and Applied Linguistics, 2003, p. 432).

Aitchinson (1987) distinguishes between ‘identification criteria’ and ‘stored knowledge’, i.e. the attributes that are crucial to the identification of a concept, and the attributes that people attribute to the objects through their learnt knowledge of the world through the lens of their culture. In this regard, the impact of one's culture is of high importance, as there is evidence that prototypes vary from language to language, and from culture to culture. As a case in point, biology students influenced by the basic academic knowledge tend to categorize bat as a mammal, but based on its appearance, a bat can be categorized as a bird.

According to Rosch (1978), prototype theory (PT) ‘suggests that many mental concepts we have are really prototypes. It has been useful in investigations into how concepts are formed, and to what extent certain concepts can be considered universal or specific to certain cultures/languages’ (p.57). A prototype is ‘a person or object which is considered (by many people) to be typical of its class or group’ (LDLTAL, 2003, p.432). This thesis is divided into four chapters. Chapter one tackles the theoretical preliminaries of prototypicality and concepts perception.

Chapter two deals with the application of the prototype theory to four English concepts: ‘crimes, emotions, diseases and drinks’ and with how to extract the prototype and the most representative words of a concept. The Arabic analysis of the Arabic counterparts of the same concepts is discussed in chapter three. Eventually, chapter four summarizes and contrasts the English data with the Arabic data to show how cultural differences impact the cognitive process.

0.2 Research Questions

The study answers several questions including

1. How effective is the prototype theory in specifying category prototypes?
2. Are there any differences or similarities in the ranking of the family members of the same concept?
3. Can changing domains affect the conceptualization and categorization? and
4. How far does the cultural, or societal, level influence the cognition of a certain concept?

0.3 Study Objectives

This study aims at investigating the validity of the prototype theory to cognitive linguistics and exploring the following points:

- (a) The effectiveness of the prototype theory in determining the most representative hyponyms of some concepts in Arabic and English data

- (b) The impact of the different domains on arranging the hyponyms in the concepts
- (c) Contrasting English with Arabic when it comes to ranking the prototypes of the same concept
- (d) The cultural factors behind such differences

0.4 Hypotheses of the Study

This study hypothesizes that the pattern ‘X such as’ in English and ‘ص ‘مثل’ in Arabic can be the linguistic reference to the abstract prototypicality relation. Moreover, the prototypical members of a semantic category are hypothesized to linguistically exist in the concordance of the target word representing the semantic category. The study also hypothesizes that the four internet domains: .com, .org, .info and .edu linguistically express different categorizations of the same concept. The final hypothesis, in this study, is the similarities and differences between the Arabic and English worlds regarding the linguistic categorization and prototypes of the studied semantic categories.

0.5 Sources of Data

The data is a web-based collection of snippets from the web search engine GOOGLE. Some programs are used like EditPad Pro and AntConc to analyze the data. To collect the data, the pattern ‘X such as’ is used to collect snippets from GOOGLE and inserting them into EditPad Pro to compartmentalize the data. The concepts that are used in English are (*crimes, emotions, diseases and drinks.*). In line with the English data, the Arabic concepts are (مشروبات-أمراض مشاعر- جرائم). Thus, the study covers the topics

of health, culture, society, and economy. The domains that are analyzed are (.com, .org, .info, .edu). Plus, the study depends on analyzing 100 snippets per domain in each concept in each month over the period of seven years from 2010 to 2016.

0.6 Description of Data

The analyzed data consists of eight Arabic and English comparable corpora representing four semantic categories: crimes, emotions, diseases and drinks. There are also sub corpora that include the analysis of four domains in each one of the concepts. The domains that are analyzed are (.com, .org, .info, .edu). The Arabic corpora are 486,618 tokens, divided as follows: drinks 92,433, emotions 48,698 crimes 83,742 and diseases 26,1745 tokens. The comparable English corpora are 3,380,406, divided as follows: drinks 634,656, emotions 739,600, crimes 812,286 and diseases 1,193,864. Arabic and English data saved in UTF-8 files to facilitate digital processing.

0.7 Methodology

This study uses AntConc content analysis program to quantitatively analyze eight corpora representing four semantic categories in Arabic and English. It further divides each corpus chronologically into seven sub-corpora covering the semantic domain in the time span (2010-2016). AntConc is adjusted to extract a 5-word concordance to the left and right of the target pattern, lexicalizing the semantic category, in the Arabic and English corpora, respectively. That is to say, only collocates after the pattern ‘X such as’ are extracted in English.