Semantic Similarity Framework for Thai Conversational Agents

Khukrit Osathanunkul

This thesis is presented for the degree of Doctor of Philosophy

School of Computing, Mathematics and Digital Technology

Manchester Metropolitan University

February 2014

Abstract

Conversational Agents integrate computational linguistics techniques and natural language to support human-like communication with complex computer systems. There are a number of applications in business, education and entertainment, including unmanned call centres, or as personal shopping or navigation assistants. Initial research has been performed on Conversational Agents in languages other than English. There has been no significant publication on Thai Conversational Agents. Moreover, no research has been conducted on supporting algorithms for Thai word similarity measures and Thai sentence similarity measures. Consequently, this thesis details the development of a novel Thai sentence semantic similarity measure that can be used to create a Thai Conversational Agent. This measure, Thai Sentence Semantic Similarity measure (TSTS) is inspired by the seminal English measure, Sentence Similarity based on Semantic Nets and Corpus Statistics (STASIS). A Thai sentence benchmark dataset, called 65 Thai Sentence pairs benchmark dataset (TSS-65), is also presented in this thesis for the evaluation of TSTS. The research starts with the development a simple Thai word similarity measure called TWSS. Additionally, a novel word measure called a Semantic Similarity Measure, based on a Lexical Chain Created from a Search Engine (LCSS), is also proposed using a search engine to create the knowledge base instead of WordNet. LCSS overcomes the problem that a prototype version of Thai Word semantic similarity measure (TWSS) has with the word pairs that are related to Thai culture. Thai word benchmark datasets are also presented for the evaluation of TWSS and LCSS called the 30 Thai Word Pair benchmark dataset (TWS-30) and 65 Thai Word Pair benchmark dataset (TWS-65), respectively. The result of TSTS is considered a starting point for a Thai sentence measure which can be illustrated to create semantic-based Conversational Agents in future. This is illustrated using a small sample of real English Conversational Agent human dialogue utterances translated into Thai.

Copyright

Declaration

No part of this thesis has been submitted in support of an application for any other degree or qualification at this or any other institute of learning. Apart from those parts of the project containing citations to the work of others and apart from the assistance mentioned in the acknowledgements, this thesis is my own work.

Signed ______________________________

Khukrit Osathanunkul

Acknowledgements

I would like to start by thanking my exceptional supervisors, Dr James D. O'Shea and Dr Keeley Crockett, for their guidance, patience, continuous support and encouragement during my PhD study. I appreciate it more than I can say. I am also very grateful to my advisor Dr Zuhair Bandar for his foresight, kindness and always having the right advice at the right time.

I would like to express my gratitude to my colleagues within the Intelligent System group. Special thanks must go to Ms Suphitcha Ratnavanija who has provided me with inspiration, encouragement and good ideas and also to Ms Praow Weeraprechamate who is extremely helpful. I am especially grateful to Dr Olawa Titiloye for being a wonderful friend and confidant, and for sharing their thoughts and feelings, and to many others whom I do not have space to name.

I also would like to thank my family members, dad and mum (**นาย และ นาง ไมตรี โอสถานันต์กุล**), sisters and brothers for being very supportive and encouraging throughout my life. I am grateful to my grandparents, **นาย บุญธรรม และ นาง ฮวย ใจอินทร์** without whose encouragement I would have never made this.

Contents

Chapter 1: Introduction **20**

1.1 Overview 21

1.2 Research Question 23

1.3 Thesis Objective 24

1.4 Thesis Structure 24

Chapter 2: Related Works **26**

2.1 Introduction 27

2.2 Word Semantic Similarity Measure 27

2.2.1 Word Semantic Similarity Measure in English 27

2.2.2 Non-English Word Similarity Measure 29

2.3 Sentence Semantic Similarity Measure in English 30

2.3.1 Latent Semantic Analysis (LSA) 31

2.3.2 Sentence Similarity based on Semantic Nets and Corpus Statistics (STATIS) 31

2.3.3 Sentence Semantic Similarity Measure in Other Languages 34

2.4 Benchmark Dataset 35

2.4.1 Word Benchmark Dataset 35

2.4.2 Sentence Benchmark Dataset 36

2.5 Challenges of Thai Language 37

2.5.1 Basis of Thai Language 37

2.5.1.1 Thai Alphabet 37

2.5.1.2 Thai Consonants 38

2.5.1.3 Thai Vowel 39

2.5.1.4 Tone Make 39

2.5.1.4.1 Tone Rule with Tone Marks 41

2.5.1.4.2 Tone Rule without Tone Mark 42

2.5.1.5 Grammar 43

2.5.1.5.1 Verbs 43

2.5.1.5.2 Adjectives and Adverbs 45

2.5.1.5.3 Nouns and Pronouns 46

2.5.1.5.4 Expressions of Time, Place, Quantity 49

2.5.2 Thai Word Order 50

2.5.3 Thai WordNet 51

2.5.4 Alternative Knowledge Other Than WordNet 51

2.5.5 Previous Research on Thai Semantic Similarity Measure 52

2.6 Conclusion 52

Chapter 3: A Prototype Version of Thai Word Semantic Similarity Measures 54

3.1 Introduction 55

3.2 A Prototype Version of Thai Word Semantic Similarity Measures (TWSS) 55

3.3 Methodology for Creating a Thai Word Benchmark Dataset (TWS-30) 57

3.3.1 Participants 57

3.3.2 Materials 58

3.3.3 Procedure 59

3.3.4 TWS-30 59

3.3.5 Discussion of TWS-30 60

3.4 Evaluation of the Thai Word Semantic Similarity Measure 62

3.4.1 Methodology 62

3.4.2 Semantic Similarity Ratings 62

3.4.3 Discussion 63

3.5 Conclusion 67

Chapter 4: 65 Word Pair Thai Benchmark Dataset (TWS-65) 68

4.1 Introduction 69

4.2 Theme Words 70

4.3 Methodology of Selecting Word Pairs from Theme Words 71

4.3.1 High Semantic Similarity Word Pair 72

4.3.1.1 Participants 72

4.3.1.2 Materials 72

4.3.1.3 Procedure 72

4.3.1.4 Results 73

4.3.2 Medium Semantic Similarity Word Pair 73

4.3.2.1 Participants 73

4.3.2.2 Materials 74

4.3.2.3 Procedure 74

4.3.2.4 Results 74

4.3.3 Low Semantic Similarity Word Pair 75

4.3.4 TWS-65 Word Pairs Dataset 75

4.4 Application of the Methodology to Rating for TWS-65 77

4.4.1 Participants 77

4.4.2 Materials 77

4.4.3 Procedure 78

4.4.4 TWS-65 78

4.4.5 Evaluation of TWS-65 with TWS-30 80

4.5 Evaluation of the Thai Word Semantic Similarity Measure 81

4.5.1 Methodology 81

4.5.2 Semantic Similarity Rating Results 82

4.5.3 Discussion 83

4.6 Conclusion 85

Chapter 5: Word Similarity based on Lexical Chain Created from Search Engine (LCSS) 86

5.1 Introduction 87

5.2 Semantic Similarity Measure based on Lexical Chain Created from Search Engine (LCSS) 88

5.2.1 Overview of LCSS Algorithm 89

5.2.2 STEP 1: Send a Request to Search Engine with “W1 (Space Bar) W2” as Input 91

5.2.3 STEP 2: Word Extraction from HTML Pages 92

5.2.4 STEP 3: Insert Lexical Chains into the Lexical Database 94

5.2.5 STEP 4: Searching for All Lexical Chain in the Lexical Database 95

5.2.5.1 STEP 4.1: Searching All Possible Lexical Chain between the Two Target Words 96

5.2.5.2 STEP 4.2: Maximum Range of Lexical Chain 97

5.2.6 STEP 5: Selecting the Best Available Lexical Chain 97

5.2.7 STEP 6: Calculate the Semantic Similarity Rating between Two Target Words 100

5.2.8 An Example of LCSS Usage with Thai Word Pairs 101

5.3 Training Dataset and Testing Dataset to Evaluate LCSS 105

5.3.1 Training Dataset 106

5.3.2 Testing Dataset 106

5.4 Experiment to Find the Most Suitable Alpha Parameter in LCSS in Thai Language 107

5.4.1 Methodology 108

5.4.2 LCSS Rating with a Different *Alpha* Parameter 108

5.4.3 Results 109

5.5 Evaluation of Testing Dataset (TWS-51) with LCSS 110

5.5.1 Methodology 110

5.5.2 LCSS Semantic Similarity Rating 110

5.5.3 Discussion 111

5.6 Conclusion 116

Chapter 6: New Thai Word Semantic Similarity Measure (nTWSS) 117

6.1 Introduction 118

6.2 New Thai Word Semantic Similarity Measure (nTWSS) 118

6.3 Experiment to Find the Most Suitable δ Parameter 120

6.3.1 Methodology 120

6.3.2 nTWSS Rating with Different *δ* Parameter 121

6.3.3 Results 122

6.4 Evaluation of TWS-51 with nTWSS 122

6.4.1 Methodology 122

6.4.2 Results 123

6.4.3 Discussion 124

6.5 Conclusion 127

Chapter 7: 65 Sentence Thai Benchmark Dataset (TSS-65) 128

7.1 Introduction 129

7.2 Creation of the TSS-65 Dataset 129

7.2.1 Experiment to Find the Most Suitable Definition for Each Noun in TWS-65 130

7.2.1.1 Participants 130

7.2.1.2 Materials 130

7.2.1.3 Procedure 131

7.2.1.4 Results 131

7.2.1.5 TSS-65 Sentence Pairs 147

7.3 Methodology for Rating TWS-65 153

7.3.1 Participants 153

7.3.2 Materials 153

7.3.3 Procedure 153

7.3.4 TSS-65 154

7.4 Discussion of the TSS-65 155

7.5 Conclusion 157

Chapter 8: Thai Short Text Semantic Similarity Measures (TSTS) 158

8.1 Introduction 159

8.2 A Thai Short Text Semantic Similarity Measure (TSTS) 159

8.2.1 Construction of the Joint Word Set 160

8.2.2 Formation of the Lexical Semantic Vectors 160

8.2.3 Calculation of the Sentence Similarity 162

8.3 Evaluation of the Thai Shot Text Semantic Similarity Measure 162

8.3.1 Methodology 162

8.3.2 Results 162

8.3.3 Discussion 163

8.4 Evaluation of TSTS usage with Conversational Agent Log Files 166

8.4.1 Methodology 167

8.4.2 Materials 167

8.4.3 Results 168

8.4.4 Discussion 169

8.5 Conclusion 169

Chapter 9: Conclusion and Future Work 170

9.1 Introduction 171

9.2 Summary of the Work 171

9.3 Summary of Contribution 173

9.4 Future Work 174

References 176

Appendixes **183**

Appendix 1.1 Ethics Statement 184

Appendix 1.2 Instructions 185

Appendix 1.3 Sample Card 186

Appendix 1.4 Sample Rating Sheet (for TWS-30) 187

Appendix 1.5 Personal Data Sheet 188

Appendix 1.6 Semantic Anchors 189

Appendix 1.7 Sample Rating Sheet (for TWS-65) 190

Appendix 2.1 Instructions for High Similarity Word Pairs 191

Appendix 2.2 List of Theme Words 192

Appendix 2.3 High Similarity Word Pairs Recording Sheet 193

Appendix 2.4 Instructions for Medium Similarity Word Pairs 194

Appendix 2.5 Medium Similarity Word Pairs Recording Sheet 195

Appendix 3.1 Instructions (for TSS-65) 196

Appendix 3.2 Sample Question Sheet 197

Appendix 4.1 Ethics Statement (for TSS-65) 198

Appendix 4.2 Instructions (for TSS-65) 199

Appendix 4.3 Sample Card (for TSS-65) 200

Appendix 4.4 Sample Rating Sheet (for TSS-65) 201

List of Figures

Figure 1.1 The Turing Test 21

Figure 2.1 Extract of WordNet 29

Figure 2.2 An Overview of STATIS 31

Figure 2.3 The Vowel in Thai Language 40

Figure 2.4 The Tone Makes in Thai Language 41

Figure 2.5 Thai WordNet Progess 51

Figure 3.1 Overview of TWSS 56

Figure 3.2 Extract of WordNet 57

Figure 3.3 Scatter between TWS-30 and R&G 61

Figure 3.4 The Correlation between Thai Human Rating and Thai Machine Rating 64

Figure 4.1 The Correlation between TWS-65 Rating and TWSS 83

Figure 5.1 An Example of the Chain of Words 90

Figure 5.2 An Overview of LCSS Algorithm 90

Figure 5.3 An Example of Results from Google 92

Figure 5.4 An Example of Two Sentences 93

Figure 5.5 An Example of Two Sentences with and without the Presence of Function Words 93

Figure 5.6 An Example of Lexical Chain before and after Remove Function Word 94

Figure 5.7 The Database 95

Figure 5.8 A Complex Chain of Words 95

Figure 5.9 Chain of Word between Word Monk and Word Priest 96

Figure 5.10 An Example of Three Sentences 98

Figure 5.11 A Chain of Word between Word Monk and Priest with Word Frequency and Link Frequency 98

Figure 5.12 An Example of Result from Google for Words *พืช (Plant)* and *ป่าไม้ (Forest)* 102

Figure 5.13 A Chain of Word between Word *พืช (Plant)* and *ป่าไม้ (Woods)* with Word Frequency and Link Frequency 103

Figure 5.14 General Form of Correlation Triangle 112

Figure 5.15 Specific Correlation Triangle for TWSS vs LCSS 113

Figure 5.16 Specific Correlation Triangle for TWSS vs LCSS with the *Problem Word Pairs* 115

Figure 6.1 An Overview of nTWSS 119

Figure 6.2 The Correlation between TWS-51 and nTWSS 124

Figure 6.3 Specific Correlation Triangles for TWSS vs nTWSS 126

Figure 7.1 The Correlation between TWS-65 and TSS-65 156

Figure 8.1 An Overview of TSTS 160

Figure 8.2 The Correlation between TSS-65 and TSTS 164

Figure 8.3 Specific Correlation Triangles for TWSS vs nTWSS 165

List of Tables

Table 2.1 The Forty-Four Consonants in the Thai Language 38

Table 2.2 The Six Final Consonant Sounds 39

Table 2.3 Frequently Used Nouns and Pronouns 47

Table 2.4 The Thai Classifier 48

Table 3.1 The Average of Similarity Rating from 40 Native Thai Speakers 60

Table 3.2 Semantic Similarity of Human Rating and Machine Rating 63

Table 3.3 The Pearson Product-Moment Correlation Coefficients. 65

Table 3.4 The Exception of Translate Word 66

Table 4.1 List of Theme Words 71

Table 4.2 High Similarity Word Pairs 73

Table 4.3 Medium Similarity Word Pairs 75

Table 4.4 TWS-65 Word Pairs 75

Table 4.5 The Average of Similarity Rating from 40 Native Thai Speakers 78

Table 4.6 TWS-65 Correlation Coefficients with Mean Human Judgment 80

Table 4.7 The Average of Similarity Rating for the 14 Word Pairs in Both Datasets 80

Table 4.8 Semantic Similarity between Human Rating and TWSS 82

Table 4.9 Correlation Coefficients 83

Table 4.10 Problem Word Pairs 84

Table 5.1 TWS-51 106

Table 5.2 The Average of Similarity Rating for the 14 Word Pairs in Both Datasets 108

Table 5.3 The Pearson Product-Moment Correlations in Each Value of Alpha 109

Table 5.4 Semantic Similarity Rating for Human Participants and LCSS 110

Table 5.5 The Pearson Product-Moment Correlation Coefficients (r) 111

Table 5.6 Problem Word Pairs 114

Table 6.1 Semantic Similarity Rating of nTWSS 121

Table 6.2 Correlation Coefficients 122

Table 6.3 nTWSS Semantic Similarity Rating 123

Table 6.4 The Pearson Product-Moment Correlation Coefficients (*r*) 125

Table 7.1 Noun *ลายมือชื่อ*(*Autograph*) 131

Table 7.2 Noun *ลายเซ็น (Signature)* 131

Table 7.3 Noun *เด็กผู้ชาย (Boy)* 132

Table 7.4 Noun *เด็กหนุ่ม (Lad)* 132

Table 7.5 Noun *ฝั่งทะเล (Shore)* 132

Table 7.6 Noun *ชายฝั่ง (Coast)* 133

Table 7.7 Noun *ป่าช้า (Cemetery)* 133

Table 7.8 Noun *สุสาน (Graveyard)* 133

Table 7.9 Noun *การเดินทาง (Journey)* 134

Table 7.10 Noun *การท่องเที่ยว (Travelling)* 134

Table 7.11 Noun *ทาส (Slave)* 134

Table 7.12 Noun *ข้ารับใช้ (Serf)* 135

Table 7.13 Noun *อุปกรณ์ (Equipment)* 135

Table 7.14 Noun *เครื่องมือ (Tool)* 135

Table 7.15 Noun *เที่ยงวัน (Midday)* 136

Table 7.16 Noun *กลางวัน (Noon)* 136

Table 7.17 Noun *อัญมณี (Gem)* 136

Table 7.18 Noun *เพชรพลอย (Jewel)* 137

Table 7.19 Noun *เนินเขา (Hill)* 137

Table 7.20 Noun *ภูเขา (Mountain)* 137

Table 7.21 Noun *ป่าไม้ (Forest)* 138

Table 7.22 Noun *พงไพร (Woods)* 138

Table 7.23 Noun *ยานพาหนะ (Vehicle)* 138

Table 7.24 Noun *รถยนต์ (Car)* 139

Table 7.25 Noun *อาหาร (Food)* 139

Table 7.26 Noun *ผลไม้ (Fruit)* 139

Table 7.27 Noun *แก้ว (Glass)* 140

Table 7.28 Noun *ถ้วย (Bowl)* 140

Table 7.29 Noun *นักบวช (Priest)* 140

Table 7.30 Noun *พระ (Monk)* 141

Table 7.31 Noun *นักมายากล (Magician)* 141

Table 7.32 Noun *พ่อมด (Wizard)* 141

Table 7.33 Noun *ผ้าฝ้าย (Muslin)* 142

Table 7.34 Noun *ผ้าไหม (Silk)* 142

Table 7.35 Noun *ครู (Teacher)* 142

Table 7.36 Noun *อาจารย์ (Lecturer)* 143

Table 7.37 Noun *นิตยสาร (Magazine)* 143

Table 7.38 Noun *หนังสือ (Book)* 143

Table 7.39 Noun *วัด (Temple)* 144

Table 7.40 Noun *โบสถ์ (Church)* 144

Table 7.41 Noun *ลุง (Uncle)* 144

Table 7.42 Noun *ป้า (Aunt)* 145

Table 7.43 Noun *สุนัข (Dog)* 145

Table 7.44 Noun *หมา (Dog)* 145

Table 7.45 Noun *โรงภาพยนต์ (Cinema)* 146

Table 7.46 Noun *โรงละคร (Theatre)* 146

Table 7.47 Noun *พืช (Plant)* 146

Table 7.48 Noun *ต้นไม้ (Tree)* 147

Table 7.49 TSS-65 Sentence Pairs 148

Table 7.50 TSS-65 Sentence Pairs with Human Rating 154

Table 7.51 Correlation Coefficients with Mean Human Judgment 155

Table 7.52 The Odd Pair 156

Table 8.1 Semantic Similarity of Human Rating, STASIS, and TSTS 163

Table 8.2 Correlation Coefficients 164

Table 8.3 The Chosen 15 Sentence Pairs 167

Table 8.4 The Results of the Chosen 15 Sentence Pairs 168

List of Abbreviations

ANOVA	Analysis of Variance
B&M	Battig and Montague
CA	Conversational Agent
HTML	Hyper Text Markup Language
LCSS	Word Similarity based on Lexical Chain Created from Search Engine
LD	Link Density
LF	Link Frequency
LSA	Latent Semantic Analysis
NTS	Native Thai speakers
nTWSS	New Version of Thai Word Semantic Similarity Measure
PMI	Point-Wise Mutual Information
R&G	Rubenstein and Goodenough
SP	Sentence Pair
STASIS	Sentence Similarity based on Semantic Nets and Corpus Statistics
STSS	Sentence Semantic Similarity Measure
SVD	Singular Value Decomposition
TSS-65	65 Thai Sentence Pairs Benchmark Dataset
TSTS	Thai Sentence Semantic Similarity Measure
TWSS	Thai Word Semantic Similarity Measure
TWS-30	30 Thai Word Pairs Benchmark Dataset
TWS-51	51 Thai Word Pairs Testing Dataset
TWS-65	65 Thai Word Pairs Benchmark Dataset
UK	United Kingdom
URL	Uniform Resource Locator
WF	Word Frequency
WP	Word Pair

Chapter 1

Introduction

1.1 Overview

Conversational Agents are of increasing importance in the marketplace where the consumer is becoming ever increasingly mistrusting of forceful salespersons and looking for new ways to conduct their everyday activities. Conversational Agents save firms a lot of money these days, whether through unmanned call centres or as personal shopping or navigation assistants, as the need to employ staff has been greatly reduced. The term ‘Conversational Agent’ within this thesis refers to the text dialogue variant, although Conversational Agents take many other forms such as the inclusion of an animated ‘avatar’ that is able to follow the mouse pointer so that it maintains realistic eye contact and body language whilst in communication (Embodied Conversational Agent). The idea of the original Conversational Agent started as long ago as the 1950s when a test of a computer’s intelligence was benchmarked against a human being by Alan Turing (Turing, 1948; Turing, 1950; Turing, 1952) as part of his research. His famous ‘Alan Turing Scrapbook’ describes the quest in detail of discovering ‘to what extent (the machine) could think for itself’ (Hodges, 1997) and assuming it could think for itself, the drive to find out how well it would perform with natural language queries. Certainly, the so called ‘Imitation Game’ was a good test of this (Turing, 1950). The game involved a human participant at a computer screen engaged in dialogue with the Conversational Agent via a keyboard and computer screen and also a human being responding with their keyboard. If a human participant could not tell computer and human apart when situated behind a screen when receiving the response to his or her questions, the Conversational Agent is said to have passed the ‘Turing Test’.

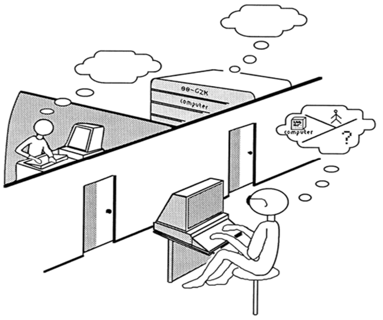


Figure 1.1: The Turing Test (Copeland, 2000)

Conversational Agents are a good measurement of a computer’s simulated human intelligence, although other qualities such as humour and wit need to be included to give the agent a social human-like presence. The agent should be able to select the right response rather than respond from the prior experience of education from everyday living as a human being would otherwise. It is mainly due to some of these aforementioned characteristics that Turing’s prediction of computers being able to fool human beings into believing they are real people has still not been fully achieved. Conversational Agents are used in many areas of commerce, including as a ‘Help Desk’ (Göker, 1998), ‘Customer Self-service’, ‘Fault Diagnosis’ (McSherry, 2001), and ‘Product Recommendation’ (Ricci, 2002) and many more. The Conversational Agent in most of these areas can be used as a ‘pattern matcher that has canned responses to anticipated inputs’ (Sammut, 2007). Natural Language Processing supports syntactic and semantic analysis and is therefore essential for a convincing Conversational Agent.

Conversational Agents can be separated in two groups:

- Pattern matching Conversational Agent
- Semantic-based Conversational Agent.

Writing a pattern matching Conversational Agent is a time consuming process. Scripting using the structural patterns of sentences requires the script writer to consider every permutation that the user may send as input (Sammut, 2001). Conversely, a semantic similarity measure can compute the similarity of meaning of many diverse human utterances against a single prototype sentence, representing the general intention of the user (O’Shea et al., 2009). Therefore, whilst a rule capturing the intention of human users in a pattern matching Conversational Agent will contain many carefully-crafted patterns, each rule in a semantic similarity-based Conversational Agent will contain a few prototype sentences (ideally one). Thus, semantic similarity as a scripting technique reduces the volume of work and skill required from a scripter. For instance, multiple sentences with similar semantics ought not to be incorporated in the same rule. Accordingly, solely one permutation could be adequate in most cases. In addition, natural language scripting is regarded as significantly easier to maintain and more intuitive to write. This semantic-based Conversational Agent approach has been conducted in English and a few other languages.

Unfortunately, research is still lacking for a Thai Conversational Agent. Moreover, research on natural language resources in Thai is also limited (Aroonmanakun, 2007; Thoongsup, 2009). Therefore, the aim of this thesis is to propose a Thai sentence semantic similarity measure, called TSTS, which can be used to create a Thai semantic-based Conversational Agent. To create a Thai sentence semantic similarity measure, a Thai word similarity measure is needed. Again, there is no research on Thai word similarity measures at the time of writing. However, related works on English word similarity measures can be used as a starting point. Also, Thai benchmark datasets are needed to evaluate those measures.

The contributions in this thesis are as follows:

- A review and discussion of Thai natural language resources
- Creation of the first Thai word semantic similarity measure (TWSS)
- Methodology for creating the first Thai word semantic similarity benchmark dataset (TWS-30)
- Application of the methodology to rating TWS-30
- Evaluation of TWSS with TWS-30
- Creation of a 65 Thai word pairs benchmark dataset (TWS-65)
- Application of the methodology to rating TWS-65
- Evaluation of TWS-30 with TWS-65
- Evaluation of TWSS with TWS-65
- Creation of a word similarity measure based on a lexical chain created from a search engine (LCSS)
- Creation of a testing dataset (TWS-51)
- Evaluation of LCSS with TWS-51
- Creation of a new word measure specifically for the Thai language (nTWSS)
- Evaluation of nTWSS with TWS-51
- Creation of a 65 Thai sentence pairs benchmark dataset (TSS-65)
- The application of the methodology to rate TSS-65
- Evaluation of TWS-65 with TSS-65
- Creation of the first Thai sentence similarity measure (TSTS)
- Evaluation of TSTS with TSS-65
- An illustration of the use of TSTS with representative dialogue utterances for a future Thai Conversational Agent.

The degree to which the contributions answer the research questions is discussed in Chapter 9, Section 9.2.

1.2 Research Questions

This thesis investigates the following research questions:

- Can a semantic-based Conversational Agent be developed in Thai?

- Can an English word similarity measure be developed for the Thai language by translating Thai words into English?
- Can a WordNet based English word similarity measure produce a similarity rating between words that are based on Thai culture?
- Can a search engine provide an alternative natural language resource for a Thai word similarity measure?
- Can a combination of TWSS and LCSS provide a better model of human perception of Thai word semantic similarity than either separately?
- Can a Thai word measure be used to develop a Thai sentence similarity measure?
- Is the developed Thai sentence similarity measure feasible to use to develop Thai Conversational Agents?

1.3 Thesis Objective

The objectives of this thesis are to address the research issues:

1. Adapting an English word similarity measure for Thai words by using translation from Thai to English.
2. Creating Thai word benchmark datasets to evaluate the Thai word similarity measure.
3. Developing a Thai word similarity measure to address weakness in (1) and include Thai culture by using a search engine to create the measure's knowledge.
4. Developing a sentence similarity measure for the Thai language based on the Thai word similarity measure.
5. Creating a Thai sentence benchmark dataset to evaluate the Thai sentence similarity measure.
6. Illustrating the use of the Thai sentence similarity measure for future Thai Conversational Agents.

1.4 Thesis Structure

This chapter presents an overview of the research. This research proposes three word similarity measures, two Thai word benchmark datasets, one Thai sentence similarity measure, and one sentence benchmark dataset. The remaining chapters are summarised as follows:

- Chapter 2 presents a background to this thesis that introduces related previous research on word and sentence similarity measures, and the fundamentals of the

Thai language and current natural language resources research on the Thai language.

- Chapter 3 presents a Thai word similarity measure based on the English WordNet (TWSS), the first Thai word benchmark dataset (TWS-30), and evaluates the experimental results.
- Chapter 4 presents a Thai word benchmark dataset (TWS-65) based on Thai culture which covers the weakness of TWS-30, and evaluates the experimental results between TWS-30 and TWS-65. An evaluation of TWSS and TWS-65 is also discussed in this chapter.
- Chapter 5 proposes a word measure that uses a search engine (LCSS) to produce its knowledge and evaluates the experimental results between TWS-65 and LCSS.
- Chapter 6 proposes a Thai word similarity measure (nTWSS) that results from combining TWSS and LCSS judgments and evaluates the experimental results between TWS-65 and LCSS.
- Chapter 7 presents the first Thai word benchmark dataset (TSS-65) that is based on Thai culture and evaluates the experimental results between TWS-65 and TSS-65.
- Chapter 8 proposes a Thai sentence similarity measure (TSTS) that results from nTWSS and evaluates the experimental results between TSS-65 and TSTS. This chapter also illustrates the potential of TSTS to create a Thai Conversational Agent.
- Chapter 9 provides a conclusion of the thesis and suggests direction for future work.

Chapter 2

Related Works

2.1 Introduction

Conversational Agents are applied in a broad range of areas including business (Lemon, 2006), education (Kopp, 2005) and entertainment (Ibrahim, 2002). Conversational Agents may be used in unmanned call centres or as personal shopping or navigation assistants to reduce operating costs and provide 24/7 access for users. Most Conversational Agents use English. However some work has been done in Chinese (Huang, 2000) and Japanese (Ehsani, 2000). Little or no work has been done in Thai.

The chief barrier to deploying Conversational Agents effectively in the real world is the labour cost of scripting and maintenance. Consequently, a new generation of sentence semantic similarity-based agents is being introduced to overcome the problem.

This chapter presents a background to this thesis that introduces related previous research including English word similarity measures, non-English word similarity measures, English sentence similarity measures, non-English sentence similarity measures, and an overview of the Thai language and current state of research on Thai WordNet. In addition, a review of Thai similarity measures is discussed in this chapter. The aim of this chapter is to investigate the research question: Can a semantic-based Conversational Agent be developed in Thai?

The remainder of this chapter is organized as follows: Section 2.2 reviews English and non-English word semantic similarity measures; Section 2.3 reviews English and non-English sentence semantic similarity measures; Section 2.4 discusses the benchmark datasets; Section 2.5 discusses the fundamentals of the Thai language and Thai linguistics resources that are available for this research; and Section 2.6 presents the conclusions.

2.2 Word Semantic Similarity Measure

This section will give a summary of English and non-English word measures. One of the most significant word measures, Li's measure (Li et al., 2003), is also reviewed in this chapter. Li's measure is also used as a prototype to create a non-English measure.

2.2.1 Word Semantic Similarity Measure in English

There are numerous approaches to word semantic similarity, including thesaurus based (Morris, 1991; Jarmasz, 2004), dictionary based (Kozima, 1993), and WordNet based (Rada, 1989; Wu, 1994). All use a lexical resource such as a directed graph or network and their semantic similarity is measured from the particular graph or network. The highlighted word measures in English are the following:

Rada et al. (1989) use the minimum number of edges separating two concepts that contain the compared word to calculate the similarity between two target words. This measure is a starting point for the edge counting based methods.

Information theory based measures were first proposed by Resnik (1995). Resnik’s measure (Resnik, 1995) used an ontology and a corpus together as the measure’s knowledge. There are a number of measures developed from the original work of Resnik. Jiang and Conrath (1997) proposed an improvement over Resnik’s original measure by taking the edge counting based methodology and the information content was added as a decision factor. Another measure that developed from Resnik’s measure is one proposed by Lin (1998). Lin’s approach (1998) calculated the similarity of two target words by the combination of the information content of the compared concepts assuming their independence (Li et al., 2003).

Bollegala et al. (2007) calculated a number of popular relatedness metrics based on page counts for a Web search engine such as point-wise mutual information (PMI); three coefficients are combined, which are the Dice coefficient, the Jaccard coefficient, and the Simpson coefficient with lexico-syntactic patterns as model features. To rank word pairs, the model parameters were trained by using Support Vector Machines (SVM). Another word measure based on a Web search engine was implemented by Sahami and Heilman (2006). A vector is represented in each snippet from the result of a search engine, and weighted with the TF*IDF score. The semantic similarity between two queries is calculated as the inner product between the centroids of the respective sets of vectors.

One of the most significant word measures is Li’s measure (Li et al., 2003). Li’s word measure is a measure based on WordNet (Miller, 1995). WordNet was developed by Princeton University and is a machine-readable lexical database which is organized by word senses. Words in WordNet can be broken down into: 'nouns', 'verbs', 'adjectives' and 'adverbs', which are grouped into sets of synonyms. These synonyms are called 'synsets' and are connected by means of 'conceptual-semantic' and 'lexical' relations. Figure 2.1 is part of the hierarchy of WordNet.

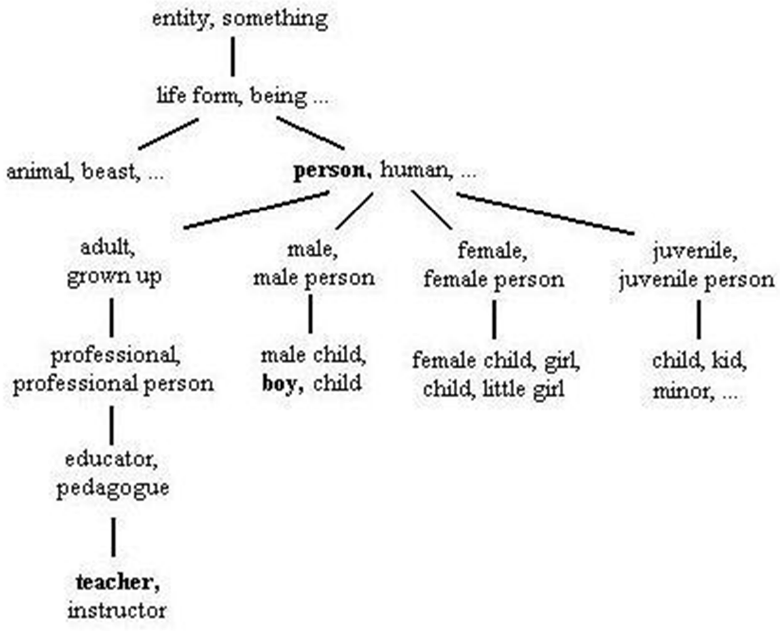


Figure 2.1: Extract of WordNet

Li’s measure approximates the semantic similarity between two words by using WordNet, which does the estimate by looking up their subsumer of two words in WordNet. The Li measure is calculated by the following formulas:

Given two words, w_1 and w_2, the semantic similarity $s(w_1, w_2)$ (Equation 2.1) can be calculated from:

$$s(w_1, w_2) = \tanh(\beta \times h) \times e^{(-\alpha \times d)} \qquad \textbf{Equation 2.1}$$

Where α and β are the length and depth factors respectively, d (Equation 2.2) can be calculated from:

$$d = d_1 + d_2 - (2 \times h) \qquad \textbf{Equation 2.2}$$

where d_1 and d_2 are the depth of $w_1$and w_2 in WordNet, and h is the depth of their least common subsumer in WordNet.

2.2.2 Non-English Word Semantic Similarity Measure

The aim of this section is to give an overview of a non-English word measure that will be implemented. There are a number of word measures in a number of languages include Chinese, Malay, and Arabic.

Guan (2002) proposes a measure of semantic similarity for Chinese words by using HowNet (Dong, 2006) as the measure's knowledge. HowNet is a bilingual common sense ontology (Chinese-English) online. There are three steps for this measure. Firstly, a concept feature file from HowNet is used to create a sememe network. Secondly, the semantic similarity degrees between sememes are given by quantifying their semantic paths in the sememe network and using a sememe weighting method. Lastly, the word measure for Chinese words is presented by combining these components.

Noah (2007) proposes a Malay word similarity measure by using a dictionary as knowledge. Lesk's method (Lesk, 1986) is adopted to use with the word measure. The similarity of words is calculated by referring to the ratio between the counts of meanings containing any of the words in the set of uniquely overlapped words found in the meanings with all the meaning associated with the word.

For Arabic, the Arabic word measure was implemented in 2013. Almarsoomi (2013) proposed a method to measure the semantic similarity between two Arabic words in the Arabic knowledge base. The measure is modified from the English word measure WordNet-based (Li et al., 2003) but uses Arabic WordNet (Elkateb, 2006) as knowledge instead. The measure has achieved the Pearson Product-Moment correlation coefficient (Blalock, 1979) between the measure and Arabic dataset (Almarsoomi, 2012) at a value of 0.894.

2.3 Sentence Semantic Similarity Measure in English (STSS)

According to O'Shea et al. (2013), there are at least 50 measures created between 2004 and 2012, including improvements of existing measures. A number of these measures are proposed based on STASIS (Ferri et al., 2007) and on LSA (Jin and Chen, 2008). Example of these are WordNet based measure (Kennedy and Szpakowicz, 2008; Quarteroni and Manandhar, 2008), or thesaurus-based measures (Inkpen 2007; Kennedy and Szpakowicz, 2008). Other technique measures include: TF*IDF variants (Kimura et al., 2007); a measure based on string similarity (Islam, 2008); other cosine measures (Yeh et al., 2008); Jaccard coefficient and other word overlap measures (Fattah and Ren, 2009); grammar based measures (Achananuparp et al., 2008); graph or tree based measures (Barzilay and McKeown, 2005); concept expansion (Sahami and Heilman, 2006); and a directional relation between text fragments measure (Corley et al., 2007). However, this section is focused on two significant sentence similarity measures: Latent Semantic Analysis (LSA) and Sentence Similarity, based on Semantic Nets and Corpus Statistics (STASIS). Also presented is a review of a non-English sentence similarity measure.

2.3.1 Latent Semantic Analysis (LSA)

Latent Semantic Analysis is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text (Landauer et al., 1998). There are two stages of LSA. First, a matrix of words is created based on the number of times a word appeared in a specific context; the word order in a sentence is not taken into consideration (Landauer et al., 1998). Second, the application of Singular Value Decomposition (SVD) is used to decompose the word matrix to reduce its size. SVD is a mathematical matrix decomposition technique that reduces the dimensional representation of the word matrix by trying to keep the entries that have the strongest relationship between the words and their occurrences in sentences. As LSA does not take the word order and the word with polysemy (the coexistence of many possible meanings for a word), this might cause an inability to analyze the sentence correctly.

2.3.2 Sentence Similarity based on Semantic Nets and Corpus Statistics (STASIS)

STASIS (Li et al, 2006) uses three elements for the determination of sentence similarity: word similarity; statistical information such as word frequency; and word order similarity. Figure 2.2 shows an overview of STASIS.

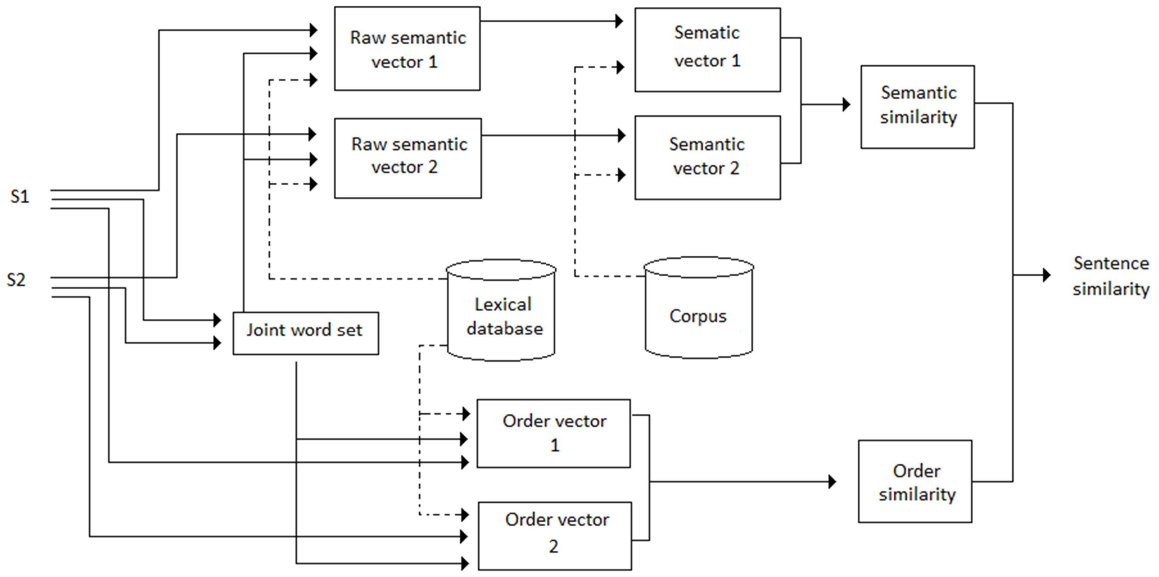


Figure 2.2: An Overview of STASIS

Construction of the Joint Word Set

Equation 2.3 describes a joint word set T derived from all unique words in two sentences: T_1 and T_2.

$$T = T_1 \cup T_2 = \{w_1, w_2, \cdots, w_m\} \qquad \textbf{Equation 2.3}$$

Giving two sentences T_1 and T_2 a joint word set is formed using Equation 2.3:

T_1: The lion is the king of the jungle.

T_2: Lion is a mammal.

A joint word set, T is

T= {The lion is the king of the jungle a mammal}

Formation of the Lexical Semantic Vectors

The vector derived from the joint word set, called the 'lexical semantic vector', denoted by $\check{s}$. $\check{s}$, is derived from the joint word set for each short text, and m equals the number of words in the joint word set. Each entry, $\check{s}_i$(where i=1, 2, ..., m) is determined by the semantic similarity of the corresponding word in the joint word set to a word in the sentence.

For each word in the joint set, there are two possible cases to process when the joint set is scanned.

- Case 1: $\check{s}_i$ is set to 1, if w_i appears in the sentence,
- Case 2: if w_i is not contained in T_1, a semantic similarity score is computed between w_i and each word in the short text T_1, using the word measure described in Li et al. (2003). The most similar word in T_1 to w_i is that with the highest similarity score. If the highest score exceeds a preset threshold, then $\check{s}_i$ is equal the highest score; if not, $\check{s}_i$ is 0.

Equation 2.4 shows how the words are weighted according to their information content (Resnik, 1999), on the assumption that word frequency influences the contribution of the individual words to the overall similarity. Entropy measures are calculated using the Brown Corpus (Francis, 1979):

$$s_i = \check{s} \times I(w_i) \times I(w_j) \qquad \textbf{Equation 2.4}$$

Given that w_i is a word in the joint word set, and w_j is its associated word in the sentence, $I(w_i)$ is the information content of w_i in the corpus. The value of $I(w_i)$ can be [0,1] and it is defined as:

$$I(w_i) = \frac{-\log(p(w_i))}{\log(N+1)} \qquad \textbf{Equation 2.5}$$

Where $p(w_i)$ is the probability of a word w_i, and N is the total number of words in the corpus, $p(w_i)$ can be calculated as:

$$p(w_i) = \frac{n+1}{N+1} \qquad \textbf{Equation 2.6}$$

where n is the word frequency of the word w in the corpus.

Calculation of the Semantic Similarity component

Lastly, the semantic similarity between T_1 and T_2 (S_s) is calculated using the cosine similarity measure between two vectors, as shown in Equation 2.7:

$$\mathrm{S_s} = \frac{\mathrm{s_1 \times s_2}}{\|\mathrm{s_1}\| \times \|\mathrm{s_2}\|} \qquad \textbf{Equation 2.7}$$

Formation of the Word Order Vectors

Given two sentences T_1 and T_2, the following sentence pair illustrates the importance of word order:

T_1: The male lion kills the poor tiger.

T_2: The male tiger kills the poor lion.

Then, using Equation 2.1 to create joint word set:

T = {The male lion kills the poor tiger}

A unique index number is assigned for each word in T_1 and T_2 by the order of the word that appears in the sentence. For instance, in T_1 for the word *lion,* the index number is 3 and 6 for *poor*. A word vector r_1 and r_2 is creating based on the joint word set. From the T_1 and T_2, the vectors $r_1$and r_2 are produced as:

r_1: {1 2 3 4 5 6 7}

r_2: {1 2 7 4 5 6 3}

Calculation of the Word Order Similarity Component

Word order similarity (S_r) is calculated as shown in Equation 2.7:

$$\mathrm{S_r} = 1 - \frac{\|r_1 - \mathrm{r}_2\|}{\|r_1 + \mathrm{r}_2\|} \qquad \textbf{Equation 2.7}$$

Calculation of Overall Sentence Similarity

Finally, the overall similarity between two sentences $S(T_1, T_2)$ is calculated using two components S_s and S_r as in Equation 2.8:

$$\mathrm{S}(T_1, T_2) = \delta S_s + (1 - \delta) S_r \qquad \textbf{Equation 2.8}$$

where δ is a constant in the range 0.5 $<\delta <$ 1 which adjusts the relative contributions of semantic and word order.

2.3.3 Sentence Semantic Similarity Measure in Other Languages

There are a number of non-English sentence measures including in Chinese, Malay, and Arabic. The recent Chinese sentence similarity measure is proposed by Ru Li (2009); the measure calculates the similarity between two Chinese sentences based on Chinese FrameNet (You, 2005) and Chinese Dependency Graph as its knowledge. The measure calculates the similarity by distance between two frames in Chinese FrameNet; also another component of similarity is obtained by looking from a Chinese Dependency Graph. Then, the similarity between two Chinese sentences is combined from those two components.

A sentence similarity measure for two Malay sentences is proposed by Noah (2007). The measure uses the method of STASIS (Li et al., 2006) to produce the sentence similarity between two Malay sentences. Although there was no short text benchmark dataset available at that time, Noah (2007) claims that the experiment has shown consistent and encouraging results which indicate the potential use of this modified approach.

An Arabic sentence similarity measure (Almarsoomi, 2013) is under development, and the measure is also derived from STASIS.

2.4 Benchmark Datasets

This section summarises benchmark datasets that are available in English. Using a benchmark dataset of word pairs or sentence pairs with similarity values that are obtained from human judgment is the only way to authenticate a semantic similarity measure (Resnik, 1999; Gurevych and Niederlich, 2005; O'Shea et al., 2013). A semantic similarity measure is evaluated by using its correlation (normally Pearson's Product-Moment correlation coefficient) with the human ratings.

This section can be separated into two parts: Word benchmark dataset and Sentence benchmark dataset.

2.4.1 Word Benchmark Datasets

Two word benchmark datasets commonly used for an evaluation of word similarity measure in English are:

- Rubenstein and Goodenough word pairs dataset (Rubenstein and Goodenough, 1965)
- Miller and Charles word pairs dataset (Miller and Charles, 1991).

Rubenstein and Goodenough (1965) built the most influential English word benchmark dataset. There are two steps to create this word dataset. The first is creating 65 word pairs ranging from maximum to minimum similarity of meaning. A list of 48 English nouns were separated into two groups (A and B); the 65 word pairs dataset is produced by choosing one word from group A and one from group B. The second step is to collect the human similarity ratings of the 65 word pairs. The participants were asked to rate the similarity of word pairs. The words pairs were scaled using a rating scale that ran from 0 (minimum similarity) to 4 (maximum similarity). The Rubenstein and Goodenough dataset, nevertheless, was published without grounds for the specific choices of 48 nouns and the method of choosing the word pairs.

After 25 years from the creation of the Rubenstein and Goodenough dataset, Miller and Charles (1991) recreated the Rubenstein and Goodenough experiment, but examined only 30 word pairs from the 65 word pairs of the Rubenstein and Goodenough dataset to avoid an inherent bias towards low similarity. The participants, 38 undergraduate students, were asked to rank the 30 word pairs using the same rating scale as Rubenstein and Goodenough from 0 to 4; the participants were all native English speakers. Comparing human ratings from the two datasets, a correlation coefficient of 0.97 was obtained, which is a high value. Moreover, in 1995, the Miller and Charles experiment was reproduced by Resnik (1995). The 10 participants, comprising computer science graduate students and post-doctoral students were asked to rank the subset of 30 word pairs from Miller and Charles. A correlation coefficient of 0.96 was obtained from this experiment. From the results of two reproduced experiments, it indicates the Rubenstein and Goodenough dataset has shown stability over the years. Plus, this stability indicates that the use of human ratings could be a reliable reference for the purpose of comparison with similarity measures.

Over 5 decades, the Rubenstein and Goodenough dataset is still valuable (Pirro, 2009). Therefore, the Rubenstein and Goodenough dataset methodology was chosen for use in producing a Thai word benchmark dataset. However, following the Miller and Charles experiment which used only 30 word pairs should be a good starting point. The creation of 30 Thai word pairs (TWS-30) is explained in detail in Chapter 3. O'Shea (2013) established that a combination of Rubenstein and Goodenough sorting, Charles' semantic anchors and other instruction produced ratings can be treated as ratio scale.

2.4.2 Sentence Benchmark Dataset

There are four notable datasets that demonstrate the ongoing state of sentence benchmark datasets which are:

- LEE50 (Lee et al., 2005)
- STSS-65 (Li et al., 2006),
- Mitchell400 (Mitchell and Lapata, 2008)
- S2012-T6 (Agirreet et al., 2012).
- STSS-131 (O'Shea et al., 2013).

LEE50 was made in 2005 using all distinct combinations of 50 email reviews of headline news stories (in the range of 51-126 words); that is, 1,225 text pairs with human ratings. Published in 2006, STSS-65 was created by replacing the words with naturalistic sentences (in the range of 5–33 words) from their dictionary definitions in the Collins Cobuild Dictionary (Sinclair, 2001), from the 65 Rubenstein and Goodenough word pairs (Rubenstein and Goodenough, 1965). STSS-131 used the best practice established from STSS-65 to rate a more representative set of English sentences. Mitchell400, presented in 2008 (Guo and Diab, 2012; Mitchell and Lapata, 2008), contains 400 pairs of simple sentences (in the range of 3 words), built by using intransitive verbs and going with subject nouns extracted from CELEX (Baayen, 1993) and the British National Corpus (Burnard, 1995). The S2012-T6 dataset contains approximately 5,200 sentence pairs (in the range of 4-61 words), separated into training set, testing set and evaluation set for Machine Learning.

The Thai sentence benchmark dataset, which will be produced in this thesis, will be based on the methodology of creating STSS-65 as the STSS-65 is built specifically for sentence similarity measure evaluation (O’Shea et al., 2013). The creation of the Thai sentence benchmark dataset will be explained in more detail in Chapter 7.

2.5 Challenges of Thai Language

The aim of this section is to describe the nature of the Thai language, including the current stage of the research in the Thai similarity measure. Also, a review of Thai word order, Thai WordNet, and the current stage of Thai similarity research is discussed in this section.

2.5.1 Basis of Thai language

The Thai language was formalised in 1283 by King Ramakamhaeng of Sukhothai. The Thai alphabet is derived from the Khmer alphabet which, in turn, was derived from Brahmic script from the Indic family. In the Thai language, there are a number of components that are different from European languages, which include consonants, vowel characters, tenses, levels of politeness, verb-to-noun conversion. The purpose of this section is to provide the information needed to understand the challenges of the Thai language.

2.5.1.1 The Thai Alphabet

The Thai alphabet uses forty-four consonants and fifteen basic vowel characters. These are horizontally placed, left to right, with no intervening space, to form syllables, words and sentences. Vowels are written above, below, before, or after the consonant they modify, although the consonant always sounds first when the syllable is spoken. The vowel characters (and a few consonants) can be combined in various ways to produce numerous compound vowels.

2.5.1.2 The Thai Consonants

Table 2.1 shows the forty-four consonants in the Thai alphabet which produce twenty-one initial consonant sounds when used at the beginning of a syllable. The forty-four consonants in the Thai alphabet are divided into three classes, which include: low class with twenty-four consonants (shown in blue); middle class with nine consonants (shown in green); and high class with eleven consonants (shown in red). The classes are important for determining the tone with which a syllable should be spoken. Since many of the consonants produce the same sound, each consonant has an acrophonic word (a system in which an alphabetic letter is represented by a word that starts with the sound of the initial letter) that is conventionally used to identify it uniquely.

Table 2.1: The Forty-Four Consonants in Thai Language (Simon, 1998)

Letter	Name/Meaning	Transliteration	Letter	Name/Meaning	Transliteration
ก ไก่	**kokai/chicken**	k	ท ทหาร	thothahan/soldier	th/t
ข ไข่	**khokai/egg**	k	ธ ธง	tho thong/flag	th/t
ฃ ฃวด	**khokhuat/bottle**	kh/k	น หนู	no nu/mouse	n
ค ควาย	**khokhwai/water buffalo**	kh/k	บ ใบไม้	bobaimai/leaf	b/p
ฅ ฅน	**khoknon/person**	kh/k	ป ปลา	popla/fish	p
ฆ ระฆัง	**khora-khang/bell**	kh/k	ผ ผึ้ง	pho phueng/bee	ph
ง งู	**ngongu/snack**	ng	ฝ ฝา	fofa/lid	f
จ จาน	**chochan/plate**	ch/j	พ พาน	pho phan/tray	ph/p
ฉ ฉิ่ง	**chochang/cymbals**	ch	ฟ ฟัน	fo fan/teeth	f
ช ช้าง	**chochang/elephant**	ch/t	ภ สำเภา	pho samphao/sailing boat	ph/p
ซ โซ่	**so so/chain**	s/t	ม ม้า	mo ma/horse	m
ฌ เฌอ	**chochoe/bush**	ch	ย ยักษ์	yo yak/ogre	y
ญ หญิง	**yoying/women**	y/n	ร เรือ	roruea/boat	r/n
ฎ ชฎา	**do cha-da/hairdress**	d/t	ล ลิง	lo ling/monkey	l/n
ฏ ปฏัก	**to pa-tak/goad**	t	ว แหวน	woweng/ring	w

ฐ ฐาน	**thosa-than/base**	th/t	**ศ** ศาลา	so sala/pavilion	s/t
ฑ มณโฑ	**thonangmon-tho/dancer**	th/t	**ษ** ฤาษี	so rue-si/Hermit	s/t
ฒ ผู้เฒ่า	**thophuthao/old person**	th/t	**ส** เสือ	so suea/tiger	s/t
ณ เณร	**no nen/novice monk**	N	**ห** หีบ	ho hip/chest	h
ด เด็ก	**do dek/child**	d/t	**ฬ** จุฬา	lo chula/kite	l/n
ต เต่า	**to tao/turtle**	t	**อ** อ่าง	o ang/basin	o
ถ ถุง	**thothung/sack**	th	**ฮ** นกฮูก	honok-huk/owl	k

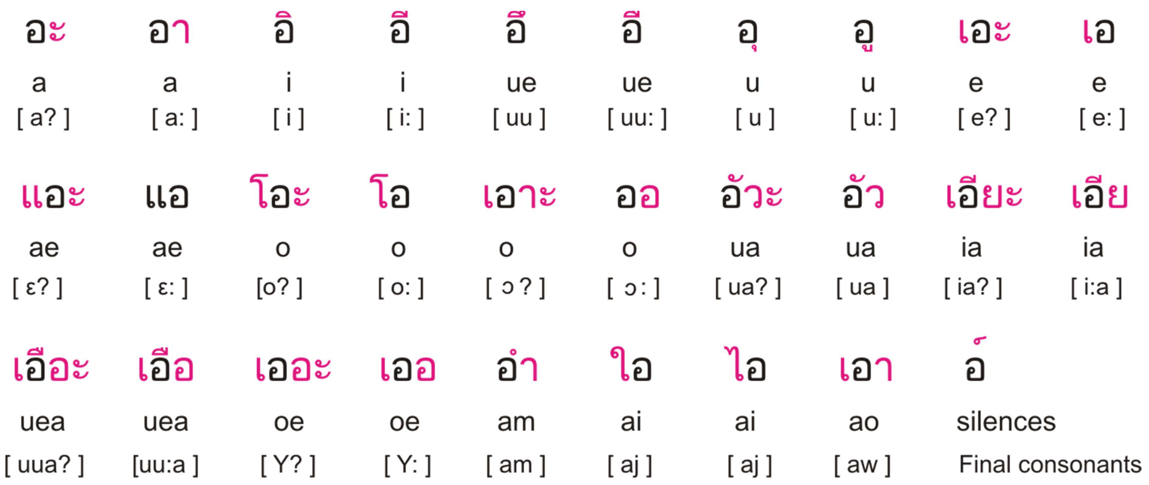
In the Thai consonants, there are some that never appear at the end of a syllable: ฉ, ซ, ผ, ฝ, ห, ฮ. In addition, in the Thai language when consonants appear at the end of a syllable, they can be separated into two groups: live consonant endings (k, p and t) and dead consonant endings (m, n and ng). Each group produces three final consonant sounds. This distinction is important for the tone rules. Table 2.2 below shows six final consonant sounds.

Table 2.2: The Six Final Consonant Sounds (Thai-language.com, 1999)

Six Final Consonant Sounds				
sound		**low**	**mid**	**high**
dead	-k	ค, ฅ, ฆ	ก	ข, ฃ
	-p	พ, ฟ, ภ	บ, ป	
	-t	ช, ฌ, ฑ, ฒ, ท, ธ	จ, ฎ, ฏ, ด, ต	ฐ, ถ, ศ, ษ, ส
live	-m	ม		
	-n	ญ, ณ, น, ร, ล, ฬ		
	-ng	ง		

2.5.1.3 Thai Vowel

The basic Thai vowel is shown in Figure 2.3. The letter *o ang* (อ) acts as a silent vowel carrier at the beginning of words that start with a vowel.



อะ	อา	อิ	อี	อึ	อื	อุ	อู	เอะ	เอ
a	a	i	i	ue	ue	u	u	e	e
[a?]	[a:]	[i]	[i:]	[uu]	[uu:]	[u]	[u:]	[e?]	[e:]
แอะ	แอ	โอะ	โอ	เอาะ	ออ	อัวะ	อัว	เอียะ	เอีย
ae	ae	o	o	o	o	ua	ua	ia	ia
[ɛ?]	[ɛ:]	[o?]	[o:]	[ɔ ?]	[ɔ :]	[ua?]	[ua]	[ia?]	[i:a]
เอือะ	เอือ	เออะ	เออ	อำ	ใอ	ไอ	เอา	อ์	
uea	uea	oe	oe	am	ai	ai	ao	silences	
[uua?]	[uu:a]	[Y?]	[Y:]	[am]	[aj]	[aj]	[aw]	Final consonants	

Figure 2.3: The Vowel in Thai Language (Simon, 1998)

The vowel karan (์) silences final consonants usually used with foreign words written in Thai such as *computer* (คอมพิวเตอร์) and *cartoon* (การ์ตูน).

Thai vowels are more complicated to use than English vowels because in English, those terms with short and long duration (when using the vowels in spoken language) do not impart meaning. This is unlike Thai where each vowel is pronounced using either a short or a long duration and do impart meaning. For example, if in Thai, the word "ka (กะ)" is spoken with a short duration it means "to estimate something" but if the word is said with long duration it is spoken as "kaa (กา)" which means "a crow".

2.5.1.4 Tone Make

There are four tones in the Thai language, maiehk, maitoh, maidtree, maijaidtawa and they are shown in Figure 2.4.

Tones are very important as there are so many short words which are spelled differently but can sound the same to a Westerner's ear. By having a different tone for each word Thai people can then understand what is being said. Basically, there are five tones: middle, low, high, rising, and falling. The middle tone is usually produced without any tone mark. However, there are some tone rules that can be separated into two parts: tone rules when there are tone marks, and tone rule when there are no tone marks.

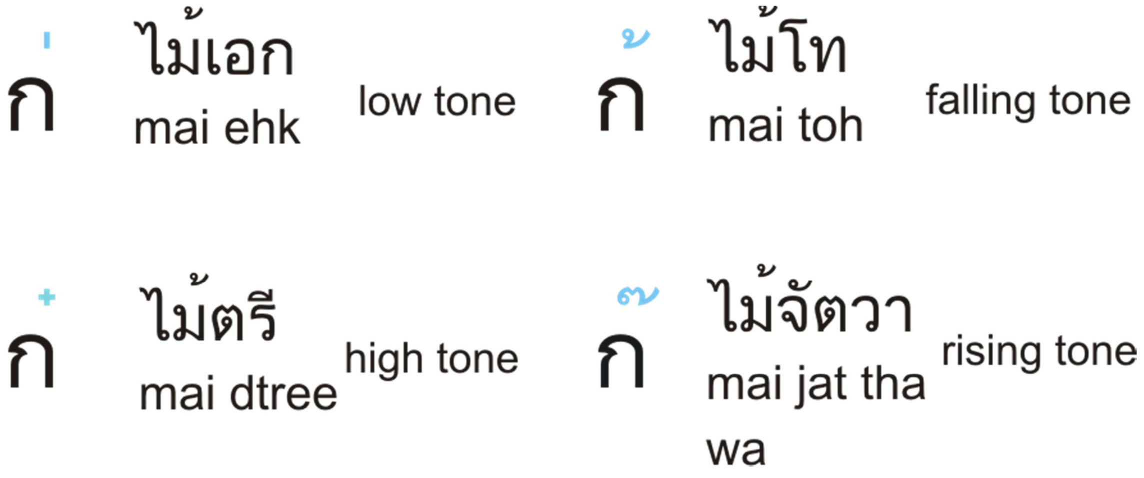


Figure 2.4: The Tone Makes in Thai Language (Simon, 1998)

2.5.1.4.1 Tone Rule with Tone Marks

In the case where there are tone marks, these can be separated into three groups of rules by consonants classes.

Tone rule with low class consonants

There are twenty-four consonants that are low class consonants. Three tones are possible (middle, falling, and high) for this class and two of the tone marks can be used, which are mai eak, and mai toh. There are two rules for this class of consonant.

- A low class consonant produced with a mai eak tone mark will create a falling tone.
- A low class consonant produced with a mai toh tone mark will create a high tone.

The middle tone can be created without any tone mark. The twenty-four consonants in this class are shown below.

ค ฆ ง ช ซ ฌ ญ ณ ฑ ฒ ท ธ น พ ฟ ม ย ร ล ว ภ ฬ ฮ

Tone rule with middle class consonants

There are nine consonants that are middle class consonants. All five tones are possible for this class and all four tone marks can be used, which are mai eak, mai toh, mai dtree, and mai juttawa. There are four rules for this class of consonant.

- A middle class consonant produced with a mai eak tone mark will create a low tone.
- A middle class consonant produced with a mai toh tone mark will create a falling tone.
- A middle class consonant produced with a mai dtree tone mark will create a high tone.

- A middle class consonant produced with a mai juttawa tone mark will create a rising tone.

The middle tone can be created without any tone mark. The nine consonants in this class are shown below.

ก จ ด ฎ ฏ ด ต บ ป อ

Tone rule with high class consonants

There are eleven consonants that are high class consonants. Three tones are possible (low, falling and rising) for this class and two of the tone marks can be used, which are mai eak, and mai toh. There are two rules for this class of consonant.

- A high class consonant produced with a mai eak tone mark will create a low tone.
- A high class consonant produced with a mai toh tone mark will create a falling tone.

The rising tone can be created without any tone mark. The nine consonants in this class are shown below.

ข ฉ ฐ ถ ผ ฝ ศ ษ ส ห

2.5.1.4.2 Tone Rule without Tone Mark

In a case when there are no tone marks, it can be separated into two groups of rules by a live or a dead syllable.

Tone rule with live syllable

A live syllable is either:

- an open syllable with a long vowel
- a closed syllable with a live consonant ending.

Two tones are possible (middle and rising) for live consonant endings. All rules for live syllables are shown below.

- A low class consonant produced with a live syllable will create a middle tone.
- A middle class consonant produced with a live syllable will create a middle tone.
- A high class consonant produced with a live syllable will create a rising tone.

Tone rule with dead syllable

A dead syllable is either:

- an open syllable with a short vowel
- a closed syllable with a dead consonant ending.

Three tones are possible (high, falling and rising) for a live consonant ending. All rules for a dead syllable are shown below.

- A low class consonant produced with a short vowel and dead syllable will create a high tone.
- A low class consonant produced with a long vowel and dead syllable will create a falling tone.
- A middle class consonant produced with a dead syllable will create a low tone.
- A high class consonant produced with a dead syllable will create a low tone.

2.5.1.5 Grammar

In comparison with English and other European languages, there is very little in the way of fixed rules in Thai grammar. There are no definite or indefinite articles, no verb conjugations, noun declensions or object pronouns. Moreover, past and future tenses are often indicated only by context, or with the words "already (laaeo: แล้ว)" or "will (ja: จะ)" tacked on. This may make it seem quite simple, but the lack of structure can end up making understanding sentences more difficult than other languages with stricter grammatical rules.

2.5.1.5.1 Verbs

In the Thai language, verbs do not change with the person, tense, voice, or number as English does. However, tenses are often indicated only by context or tense markers before or after the verb.

Typically, the past tense can be indicated by laaeo (แล้ว) after the verb. In addition, dai (ได้) is also used to indicate the past tense by being placed before the verb. It is also possible to have those two words in one sentence.

For instance:

- ***dai (ได้)***

  Sentence: เขาได้กิน

  Transliteration: khao **dai** kin

Translation: S/he ate

- ***laaeo (แล้ว)***

  Sentence: เขากินแล้ว

  Transliteration: khao kin **laaeo**

  Translation: S/he ate or He has already eaten

- ***dai (ได้) and laaeo (แล้ว)***

  Sentence: เขาได้กินแล้ว

  Transliteration: khao **dai** kin **laaeo**

  Translation: S/he ate or He has already eaten

Moreover, the word muea wan (yesterday: เมื่อวาน) can be an indicated action which took place in the past. This word can be added either at the beginning of a sentence or the end of a sentence.

Sentence: เขากินแล้วเมือวานนี้

Transliteration: khao kin laaeo **muea wan nee**

Translation: S/he already ate yesterday

The present tense can be often indicated by kamlang (currently: กำลัง) before the verb for ongoing action (as in the English -ing form). Also, it can be indicated by yu (อยู่) after the verb, or by both.

For example:

- ***kamlang (currently: กำลัง)***

  Sentence: เขากำลังวิ่ง

  Transliteration: khao **kamlang** wing

  Translation: S/he is running

- ***yu (อยู่)***

  Sentence: เขาวิ่งอยู่

  Transliteration: khao wing **yu**

  Translation: S/he is running

- ***kamlang(currently: กำลัง) and yu (อยู่)***

  Sentence: เขากำลังวิ่งอยู่

  Transliteration: khao **kamlang** wing **yu**

  Translation: S/he is running

The future tense can be indicated by ja (will: จะ) before the verb or by a time expression indicating the future.

For example:

- ***ja (will: จะ)***

  Sentence: เขาจะวิ่ง

  Transliteration: khao **ja** wing

  Translation: S/he will run or He/She is going to run

The passive voice is indicated by the insertion of thuk (ถูก) before the verb. This describes an action that was experienced by rather than controlled by the person.

For example:

- ***thuk(ถูก)***

  Sentence: เขาถูกตี

  Transliteration: khao **thuk** ti

  Translation: S/he is hit

Negation is indicated by placing mai (not: ไม่) before the verb.

For example:

- ***mai (not: ไม่)***

  Sentence: เขาไม่กิน

  Transliteration: khao**mai** kin

  Translation: S/he does not eat.

2.5.1.5.2 Adjectives and Adverbs

In the Thai language, there is no specific rule about where adverbs or adjectives should be. There are a number of words that can be used in either function. They follow the word they modify, which may be a noun, verb, or another adjective or adverb.

For instance:

Sentence: คนอ้วน

Transliteration: khon **uan**

Translation: a fat person

Sentence:	คนอ้วนๆ[1]
Transliteration:	khon **uan uan**
Translation:	a very/rather fat person

Sentence:	คนที่อ้วนเร็วมาก
Transliteration:	khon thi **uan** reo mak
Translation:	a person who becomes/became fat very quickly

Sentence:	คนที่อ้วนเร็วมากๆ[1]
Transliteration:	khon thi **uan** reo mak mak
Translation:	a person who becomes/became fat very very quickly

For the comparative in Thai, this is often expressed as "A X kwa (กว่า) B" which means A is more X than B.

For example:

Sentence:	ฉันอ้วนกว่าเขา
Transliteration:	chan uan **kwa** khao
Translation:	I am fatter than her/him

In the case of the superlative in Thai, it takes the form "A X thi sut (ที่สุด)" which means A is the most X.

For example:

Sentence:	เขาอ้วนที่สุด
Transliteration:	khao uan **thi sut**
Translation:	S/he is the fattest

2.5.1.5.3 Nouns and Pronouns

In Thai, nouns are neither singular nor plural. There are some specific words that can point out which nouns are plural. The word called phuak (พวก) can be used as a prefix to a noun or pronoun to indicate which noun is plural.

[1] Mai ya mohk (ๆ) - the repetition character in written Thai.

For example:

Sentence:	เด็ก
Transliteration:	dek
Translation:	Child

Sentence:	พวกเด็ก
Transliteration:	**phuak** dek
Translation:	A group of children

In addition, there are some nouns and pronouns that can be used as plural by adding Mai ya mohk (ๆ) at the end of the word.

For example:

Sentence:	เด็กๆ
Transliteration:	dek dek
Translation:	A group of children

Subject pronouns are often omitted, while nicknames are often used where English would use a pronoun. There are specialised pronouns in the royal (Royal family) and sacred (Monk) Thai languages. The nouns and pronouns that often appear in common conversation are shown in Table 2.3.

Table 2.3: Frequently Used Nouns and Pronouns

Word	Transliteration	Translation
ผม	phom	**I (masculine; formal)**
ดิฉัน	dichan	**I (feminine; formal)**
ฉัน	chan	**I (masculine or feminine; informal)**
คุณ	khun	**you (polite)**
ท่าน	than	**you (polite to a person of high status)**
เธอ	thoe	**you (informal, usually use with girl/woman)**
เรา	rao	**we/us, I/me/you (casual)**
เขา	khao	**he/him, she/her**
มัน	man	**It**
พวกเขา	phuak khao	**they/them**

As shown in Table 2.3, the word rao (เรา: we) can represent the first person (I), second person (you), or both (we), depending on the context. Another thing that makes Thai more complicated than European languages is that the word ‘I’ in English can only mean “I” unlike in Thai. There are a number of words that can mean ‘I’ such as phom (ผม), chan (ฉัน), dichan (ดิฉัน), nuu (หนู), and gra maawm (กระหม่อม). Each word expresses a different gender, age, level of politeness, status, and relationship between the speaker and listener.

Moreover, there are classifiers (used as a measure word) that are used with plurals. A classifier is almost always used in the Thai language unlike in English or European languages. There are a number of words that can be classifiers. Examples of those words are shown in Table 2.4.

Table 2.4: The Thai Classifier

Word	Transliteration	Used with
อัน	un	for small objects, things (in general)
ฉบับ	cha bub	for letters, newspapers
ช่อ	chaw	for bunches of flowers
บาน	baan	for windows, doors, picture frames, mirrors
ใบ	bai	for round hollow objects , leaves
ดอก	dork	for flowers
ดวง	duang	for stars, postage stamps
ฟอง	fong	for poultry eggs
ห่อ	hor	for bundles, parcels
แก้ว	gaew	for drinking glasses, tumblers
คำ	cum	for words, mouthful of food
คัน	cun	for vehicles, umbrellas, cars
คน	kon	for a person, a child, human beings
คู่	koo	for pairs of articles, forks and spoons
แก้ว	gaew	for drinking glasses, tumblers
กล่อง	gluk	for matchboxes
ก้อน	gon	for lumps of sugar, stones
กระบอก	gra bawk	for guns, cannon
กอง	gong	for piles or heaps of stones, sand
ลำ	lum	for boats, ships, aeroplanes
หลัง	lung	for houses, mosquito nets
เล่ม	lem	for books, candles, scissors
เม็ด	met	for smaller things, fruit pits, pills
มวน	muan	for cigarettes
องค์	ong	for holy personages, kings, also for monks
แผ่น	phaen	for sheets of paper, pieces of plank

The Thai classifiers are used with plurals under the term of “noun-number-classifier”. As shown in Table 2.4, those words can be only used with specific nouns.

For example:

- ***Cha bub (ฉบับ)***

Sentence:	จดหมายหนึ่งฉบับ
Transliteration:	joht maay neung **cha bap**
Translation:	one letter

- ***Chaw (ช่อ)***

Sentence:	ดอกไม้หลายช่อ
Transliteration:	daawk maai laay **chaw**
Translation:	a bunch of flowers

- ***Duang (ดวง)***

Sentence:	คืนนี้มีดาวหลายดวง
Transliteration:	kheuun nee mee daao laay **duang**
Translation:	This evening there are many stars.

2.5.1.5.4 Expressions of Time, Place, Quantiy

In English, the words “what, when, where, why, and who” are mainly used at the beginning of a sentence unlike in Thai, where these words are always at the end of a sentence. In addition, “khrai (ใคร: who)” and “Tum mai (ทำไหม: why)” can often be used at the beginning of the sentence. Examples of question sentences are shown below.

- ***A rai (อะไร: what)***

Sentence:	คุณชื่ออะไร
Transliteration:	kun cheuu **a rai**
Translation:	What is your name?

- ***Meuua rai (เมื่อไร: when)***

Sentence:	คุณกลับบ้านเมื่อไร
Transliteration:	kun glab barn **meuua rai**
Translation:	When do you go home?

- ***Tee nhai (ที่ไหน: where)***

  Sentence: คุณมาจากที่ไหน

  Transliteration: kun mar jarg **tee nhai**

  Translation: Where do you come from?

- ***khrai (ใคร: who)***

  Sentence: คุณคือใคร

  Transliteration: kun kheuu **khrai**

  Translation: Who are you?

  Sentence: ใครไปกรุงเทพ

  Transliteration: **khrai** bpai groong thaehp

  Translation: Who goes to Bangkok?

- ***Tum mai (ทำไม: why)***

  Sentence: ทำไมถึงทำแบบนี้

  Transliteration: **tum mai** theung tham baaep nee

  Translation: Why do you do it like that?

  Sentence: ทำแบบนี้ทำไม

  Transliteration: tham baaep nee **tum mai**

  Translation: Why do you do it like that?

2.5.2 Thai Word Order

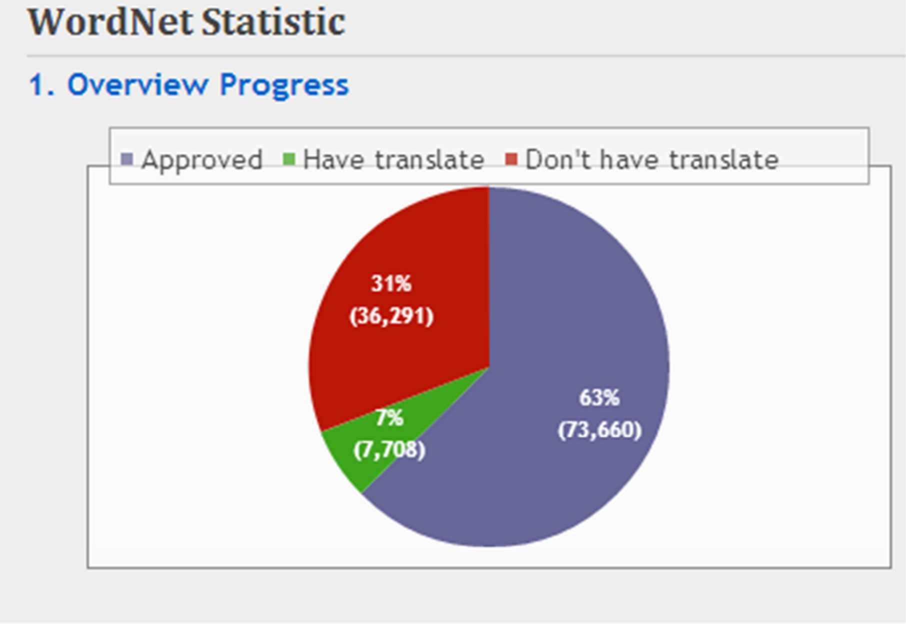
A number of similarity measures take word order as an important part to produce the semantic similarity between two sentences including STASIS. However, in the Thai language, word order is not an important part of determining the meaning of the sentence, for instance, in the 5 given sentences:

S1: เมื่อวานฝนตกที่กรุงเทพฯ (yesterday raining at Bangkok)
S2: ฝนตกเมื่อวานที่กรุงเทพฯ (raining yesterday at Bangkok)
S3: ที่กรุงเทพฯฝนตกเมื่อวาน (at Bangkok raining yesterday)
S4: ที่กรุงเทพฯเมื่อวานฝนตก (at Bangkok yesterday raining)
S5: เมื่อวานกรุงเทพฯฝนตก (yesterday Bangkok raining)

All five sentences mean exactly the same: ‘It was raining yesterday in Bangkok’. This example is the example of one of a number of cases showing that the word order in Thai does not play an important role to determine meaning.

2.5.3 Thai WordNet

There is a Thai WordNet (Thoongsup, 2009) which is still under implementation. Thai WordNet is created from English WordNet (Miller, 1995) by using a translation of a Thai-English dictionary (Sornlertlamvanich, 2008) to create senses in Thai. Figure 2.5 shows the current progress on Thai WordNet.



Source: http://th.asianwordnet.org/statistic, Date: 10/01/14

Figure 2.5: Thai WordNet Progress

From Figure 2.5, only 63% of Thai words have been approved into Thai WordNet. Plus, Thai WordNet pays no attention to the Thai words that have more than one meaning in Thai culture, as Thai WordNet is constructed by using English WordNet (Thoongsup, 2009). This makes the use of Thai WordNet unreliable to create any similarity measure that is based on the Thai language.

2.5.4 Alternative Knowledge Other Than WordNet

As discussed in Section 2.5.3, Thai WordNet is not reliable. However, there are a number of English similarity measures that use alternative knowledge other than WordNet, including corpus and search engines. The Thai corpus was first introduced in 2007 by

Wirote Aroonmanakun (2007), called “Thai National Corpus”. In 2009, there was a report of the current stage of the corpus (Aroonmanakun, 2009). Since 2006, the Thai National Corpus could only collect fourteen million words out of its aim of eighty million words (Aroonmanakun, 2009). This means the Thai National Corpus had completed only 17.5% over three years since the project first started. Therefore, the Thai National Corpus is not chosen to develop the Thai similarity measure.

Further Thai lexical knowledge can be obtained from a search engine. Consequently, a search engine will be chosen as lexical knowledge for the Thai measure. This subject will be discussed in Chapter 5.

2.5.5 Previous Research on Thai Semantic Similarity Measure

Unfortunately, no research has been conducted regarding both Thai word similarity measures and Thai sentence similarity measures. Therefore, this was an encouraging factor for the novelty of this research. The research in English similarity measures provides a good starting point. There are two significant sentence measures in English, which are STASIS and LSA. LSA architecture superficially seems to be a good choice to use in developing Thai sentence similarity as both LSA and the Thai language pay no attention to word order. LSA requires substantial corpora to produce the similarity rating. However, due to the lack of natural language resources in Thai, there is only one Thai corpus, which is still in the development process as mentioned in Section 2.5.4. Therefore, LSA architecture is not suitable at this time. According to Pirro (2009) and Hliaoutakis (2006), the STASIS is more effective for larger scale applications as it is simple and fast to calculate. Accordingly, the STASIS architecture was chosen to develop a Thai sentence similarity measure. Moreover, STASIS architecture has also influenced development in a number of languages including Malay and Arabic. As the aim of this thesis is to propose a Thai sentence similarity measure that can be applied to create a Thai Conversational Agent, a Thai word similarity measure is needed as a first step. Furthermore, Thai word and sentence benchmark datasets are also needed for the Thai measure evaluation. The creation of the first Thai word measure will be described in more detail in Chapter 3.

2.6 Conclusion

This chapter presented an overview of word and sentence similarity measures. Also reviewed were some of the non-English similarity measures and English benchmark datasets; fundamentals and difficulties of Thai language have also been discussed. As mentioned, the aim of this research is to propose a Thai sentence similarity measure. Unfortunately, to date, there is no reported work on Thai similarity. Therefore, the research question ‘Can a semantic-based Conversational Agent be developed in Thai?’ cannot be given an immediate answer ‘YES’. The Thai language simply does not yet have the resources to support this. Therefore, the main focus of this work is to create a suitable framework to support future work. The STASIS architecture is chosen to develop a Thai sentence similarity measure. However, to propose a first Thai sentence similarity measure, a Thai word similarity measure is needed. The creation of the first Thai word measure can be found in Chapter 3.

Chapter 3

A Prototype Version of Thai Word Semantic Similarity Measures

3.1 Introduction

The aim of this thesis is to develop the first Thai sentence semantic similarity measure. To do this, the first step is to develop a Thai word semantic similarity measure. As mentioned in Chapter 2, there has been no research on Thai word semantic similarity. This chapter presents an experiment with new benchmark datasets to investigate the application of a WordNet-based (Miller, 1995) machine measure to Thai similarity as a starting point for a Thai word measure. Because there is no functioning Thai WordNet (Sornlertlamvanich, 2009; Thoongsup, 2009), as mentioned in Section 2.5.3, the aim of this chapter is to investigate the research question: Can a WordNet-based word similarity measure be developed for the Thai language by translating Thai words into English and using the English Language WordNet in a word similarity algorithm?

The contributions in this chapter are:

- Creation of a first Thai word semantic similarity measure (TWSS)
- Methodology for creating the first Thai word semantic similarity benchmark dataset (TWS-30)
- Application of the methodology to rating TWS-30
- Evaluation of TWSS with TWS-30.

The rest of this chapter is organized as follows: Section 3.2 describes the development of the Thai word semantic similarity measure works; Section 3.3 describes the collection of a Thai word similarity benchmark dataset from participants using a method based on Miller and Charles (1991) and O'Shea et al. (2008); Section 3.4 discusses human and machine similarity ratings; and Section 3.5 is the conclusion.

3.2 A prototype version of Thai Word semantic similarity measure (TWSS)

In this research, Li's word similarity measure (Li et al., 2003) provides a starting point for a prototype of a Thai word semantic similarity measure as STASIS architecture was chosen to develop a Thai sentence similarity measure, as mentioned in Section 2.5.5. However, it is not possible to use Li's measure (2003) to calculate the Thai words without any modification as the Thai and English languages are different, as discussed in Section 2.5. Also, Thai WordNet is very immature and not suitable for a Thai word similarity measure. Therefore, Li's measure (2003) is modified by using a Thai-English translation as a starting point.

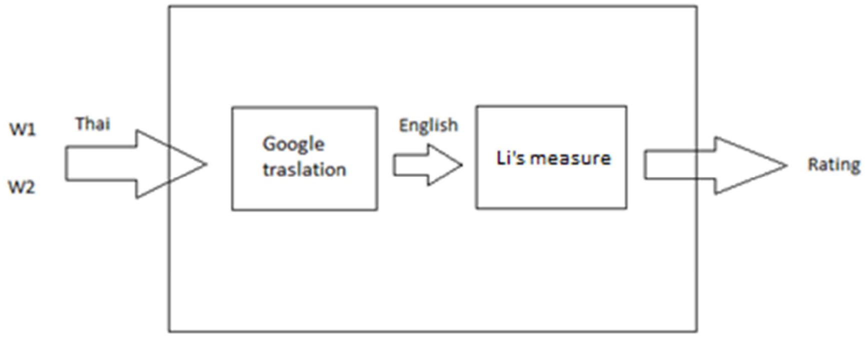


Figure 3.1: Overview of TWSS

Figure 3.1 shows an overview of the prototype Thai Word Semantic Similarity Measure (TWSS). The Li algorithm (Li et al., 2003) was adapted by using a machine translation of Thai words to English before submitting them to the algorithm. This was done by choosing the first sense (the most frequently used sense) returned by the Google translation utility. There are two reasons that Google Translate was chosen for use in this research. First, the Google translation engine uses the United Nations' parallel corpus to train their translation engine (Och, 2005). The United Nations' parallel corpus consists of around 300 million words per language (Eisele, 2010). Second, apart from English to Thai, it can translate over 53 languages for which the methodology could be adapted for further research.

TWSS calculates the similarity between words by looking up their subsumer in WordNet. TWSS is processed by the following steps:

- Given w_1 and w_2 are two Thai words
- The two Thai words are translated into English by Google translation
- Calculate the rating of two Thai words $s(w_1,w_2)$ by using Li's measure (Li et al., 2003).

The TWSS rating can be calculated as follows:

Given two words, w_1 and w_2, the semantic similarity $s(w_1, w_2)$ (Equation 3.1) can be calculated from:

$$s(w_1, w_2) = \tanh(\beta \times h) \times e^{(-\alpha \times d)}$$ **Equation 3.1**

where d can be calculated from Equation 3.2, α = 0.2 and β = 0.6, which is the optimal value reported by (Li el at., 2003).

$$d = d_1 + d_2 - (2 \times h)$$ **Equation 3.2**

where d_1 and d_2 are the depth of w_1 and w_2, and h is the depth of their least common subsumer in WordNet.

For example, if *w1* is *teacher* and *w2* is *boy* in Figure3.2, the depth of *w1* and *w2* are 7 and 5, respectively, and the synset of *person* is called the subsumer for the words of *teacher* and *boy*. Therefore, *h* for *teacher* and *boy* is 3, and the *d* of *teacher* and *boy* is 6.

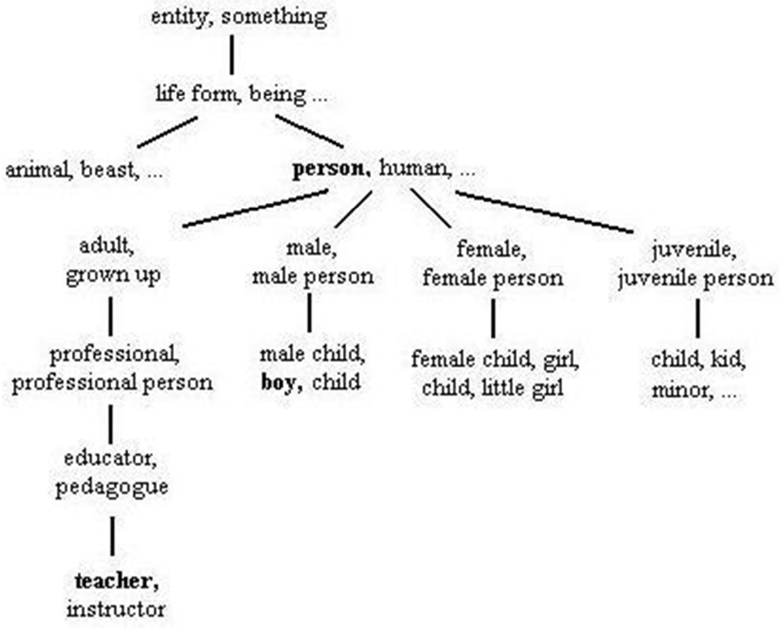


Figure 3.2: Extract of WordNet (Li et al., 2003)

3.3 Methodology for Creating a Thai Word Benchmark Dataset (TWS-30)

The aim of this section is to describe the methodology for creating the first Thai word benchmark dataset, which will be called “TWS-30”. Also, it will present TWS-30 so that it can be used to evaluate the prototype TWSS by calculating and comparing the similarity rating between Human and TWSS in Section 3.4.

3.3.1 Participants

The experiment used 40 participants to provide a safe margin above the group size of 32 which has been sufficient to obtain statistically significant results in prior work (O’Shea et al., 2013). In addition, prior work has shown that a diverse group of students can represent the general population (O’Shea et al., 2013).

Similarity ratings were collected from 40 native Thai speakers to create a benchmark dataset. The participants had an equal number of Art/Humanities and Science/Engineering backgrounds. They consisted of 12 undergraduates and 28 postgraduates studying in 4 different UK universities. The average age of the participants was 25 and standard deviation was 2.8, with 23 males and 17 females. The overall breakdown of qualifications was: 45% Bachelor’s degrees; 8% PhDs; 42.5% Master’s. This is comparable with participant groups used for English word similarity by both Rubenstein and Goodenough (1965) and Miller and Charles (1991).

3.3.2 Materials

Following the previous practice of Miller and Charles (1991) and O’Shea et al. (2008), a representative subset of 30 word pairs evenly spread across the similarity range was chosen from the Rubenstein and Goodenough dataset (1965). The original Rubenstein and Goodenough 65 word pair dataset is biased towards low similarity, and so Miller and Charles (1991) selected a subset of 30 word pairs to avoid an inherent bias towards low similarity. The important issue has been raised (O’Shea et al., 2008) that these words are not a representative sample, consisting of largely concrete nouns. However, the set has been widely used in prior word studies (Miller and Charles, 1991; Resnik, 1995; O’Shea et al., 2008). The semantic properties of these words are well understood by researchers in English and this advantage is considered important in creating a Thai dataset at this early stage in the field. Those 30 word pairs were translated into Thai by a native Thai speaker using the first meaning from an established Thai-English dictionary (Trakultaweekoon, 2007). Each word pair was printed on a separate card using a standard Thai font. A questionnaire was produced containing instructions for recording similarity ratings and a small amount of personal data (Name, Confirmation of being a native Thai speaker, Age, Gender, and Academic background) was collected. Semantic anchors were also provided to guide the participants. Appendix 1 contains the following examples of experimental materials:

- Appendix 1.1 The Ethics Statement
- Appendix 1.2 The Instruction Sheet
- Appendix 1.3 A Sample Card

- Appendix 1.4 Sample Rating Recording Sheet
- Appendix 1.5 The Person Data Collection Sheet
- Appendix 1.6 Semantic Anchors.

3.3.3 Procedure

The participants were asked to perform the following established procedure (Rubenstein and Goodenough, 1965; Charles, 2000; O'Shea et al., 2008):

1. Please sort the cards into four groups in a rough order of the similarity of meaning of the word pair.
2. After sorting the cards into groups, order the cards in each group according to similarity of meaning (i.e. the card that contains the lowest similarity of meaning is at the top of the group).
3. Please recheck the cards in every group. You may change a word pair to other groups at this stage.
4. Please rate the semantic similarity rating of each pair of words by writing a number between 0.0 (minimum similarity) and 0.9 for the first group, 1.0 and 1.9 for the second group, 2.0 to 2.9 for the third group, 3.0 and 4.0 (maximum similarity) for the fourth group on the recording sheet. You can use the first decimal place (e.g. 2.5) to show finer degrees of similarity. You may also assign the same value to more than one pair.

The cards were shuffled into a random order before being given to the participants. The participants were supervised by the experimenter during the experiment. Previous work (O'Shea et al., 2008) has found no evidence to support the idea of the order of presentation of the sentences in the pair biasing similarity judgment.

3.3.4 TWS-30

The benchmark dataset is shown in Table 3.1. R&G words are the original words from Rubenstein and Goodenough (1965). Translated words are the Thai words translated by Google translation as described in Section 3.2. Column *WP* shows the number of the word pair of TWS-30. Column *Human* presents the human rating for the Thai word pairs.

Table 3.1: The Average of Similarity Rating from 40 Native Thai Speakers

WP	**Word Pair**				***Human***
	W₁		*W₂*		
	R&G Word	Translated Word	R&G Word	Translated Word	
1	Cord	สายไฟ	Smile	รอยยิ้ม	0.078
5	Autograph	ลายมือชื่อ	Shore	ชายฝั่ง	0.022
9	Asylum	ที่หลบภัย	Fruit	ผลไม้	0.068
13	Boy	เด็กผู้ชาย	Rooster	นกตัวผู้	0.682
17	Coast	ฝั่งทะเล	Forest	ป่าไม้	0.632
21	Boy	เด็กผู้ชาย	Sage	นักปราชญ์	0.598
25	Forest	ป่าไม้	Graveyard	สุสาน	0.548
29	Bird	นก	Woodland	ป่าเขา	0.595
33	Hill	เนินเขา	Woodland	ป่าเขา	2.162
37	Magician	นักมายากล	Oracle	คำทำนาย	1.260
41	Oracle	คำทำนาย	Sage	นักปราชญ์	1.298
47	Furnace	เตาหลอม	Stove	เตาไฟ	1.613
48	Magician	นักมายากล	Wizard	พ่อมด	1.570
49	Hill	เนินเขา	Mound	ภูเขา	2.420
50	Cord	สายไฟ	String	เชือก	0.882
51	Glass	แก้ว	Tumbler	ถ้วยแก้ว	3.125
52	Grin	ยิ้มกว้าง	Smile	รอยยิ้ม	2.330
53	Serf	ทาส	Slave	ข้ารับใช้	3.345
54	Journey	การเดินทาง	Voyage	การท่องเที่ยว	2.788
55	Autograph	ลายมือชื่อ	Signature	ลายเซ็น	3.223
56	Coast	ฝั่งทะเล	Shore	ชายฝั่ง	3.218
57	Forest	ป่าไม้	Woodland	ป่าเขา	2.830
58	Implement	อุปกรณ์	Tool	เครื่องมือ	3.335
59	Cock	ไก่ตัวผู้	Rooster	นกตัวผู้	1.515
60	Boy	เด็กผู้ชาย	Lad	เด็กหนุ่ม	2.425
61	Cushion	เบาะ	Pillow	หมอน	2.035
62	Cemetery	ป่าช้า	Graveyard	สุสาน	3.400
63	Automobile	รถยนต์	Car	รถเก๋ง	3.080
64	Midday	เที่ยงวัน	Noon	กลางวัน	3.008
65	Gem	อัญมณี	Jewel	เพรชพลอย	3.075

3.3.5 Discussion of TWS-30

An appropriate measure of consistency is the correlation coefficient. Similarity measurements have usually been treated as being on a ratio scale in previous word similarity works (Rubenstein and Goodenough, 1965; Miller and Charles, 1991; Resnik, 1999; Charles, 2000). Previous word similarity work has also made the untested assumption that data are normally distributed. However, a recent thorough investigation has established that the English STSS methodology used as the model for this work does produce data suitable for Pearson Product-Moment correlation coefficient (O’Shea et al., 2013). Consequently, the Pearson Product-Moment correlation coefficient (Blalock, 1979) is appropriate.

Calculating Pearson Product-Moment correlation coefficient between TWS-30 and Rubenstein and Goodenough the result is:

- Pearson's *r* = 0.857 (P-Value < 0.01)

For both *r*, a value of +1 indicates perfect correlation, 0 indicates no relationship and -1 indicates a perfect negative correlation. P-values indicate the likelihood of obtaining the result by chance.

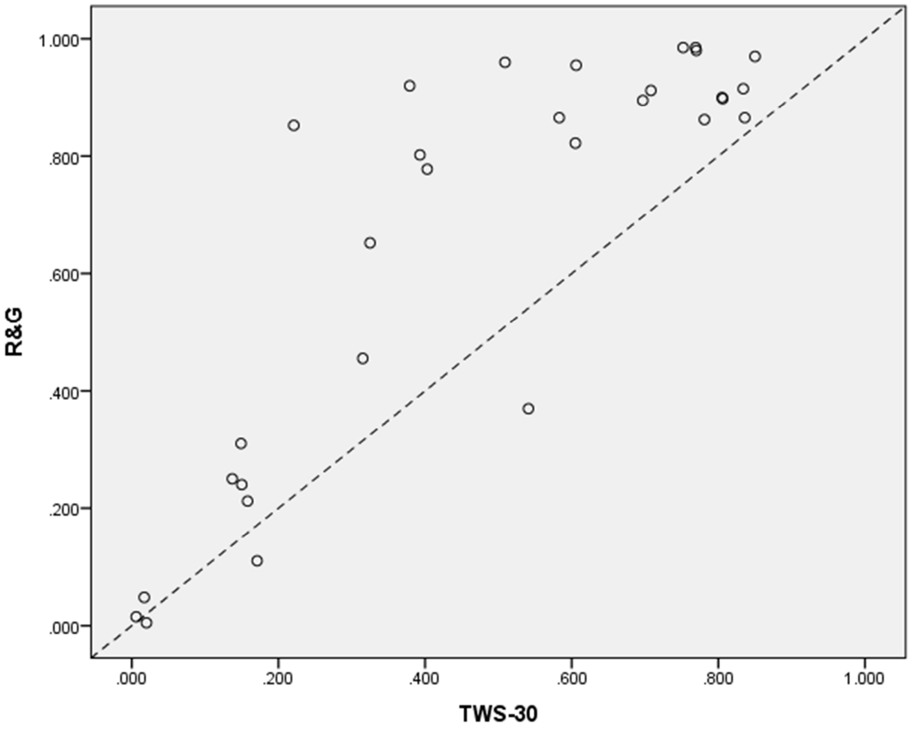


Figure 3.3: Scatter between TWS-30 and R&G

Figure 3.3 shows data points between TWS-30 and R&G. Most of the data points are near the linear line (dotted line). It is quite reasonable not to have a prefect correlation. This is because some of the words are polysemy (the coexistence of many possible meanings for a word) e.g. in English, word 'glass' and 'crane'. The same as Thai, the word 'แก้ว' (glass) means either 'glass' or 'crystal'. It has been conjectured that participants adopt the most similar pair of senses for polysemous words (Rubenstein and Goodenough, 1965). However, there are some data points that are far away from the linear line. The worst data point comes from word pair 50 (Cord-String) top-left in the figure. The word pair 50 (Cord-String) from the R&G English human rating was given as 3.41. Yet, after translation into Thai as 'สายไฟ-เชือก', the Thai human rating was given as 0.882. This is because in Thai there is the word 'สายไฟ' which has a meaning similar to 'Cable' in English. Thus, it can be said that Thai humans rate the word pair 50 (Cord-String) according to its meaning in Thai.

3.4 Evaluation of the Thai Word Semantic Similarity Measure

The aim of this section is to describe a series of experiments that were conducted using TWS-30 to evaluate the prototype TWSS measure described in Section 3.2.

3.4.1 Methodology

To evaluate the prototype TWSS measure, a benchmark dataset is required. The word benchmark dataset described in Section 3.3 can now be used to evaluate the TWSS measure described in Section 3.2. The methodology is follows:

- Translate all word pairs in TWS-30 into English using Google translation
- Calculate the TWSS rating for each word pair in TWS-30.

The Pearson Product-Moment correlation coefficients (*r*) between Thai human rating and TWSS will be calculated and shown in Section 3.4.3.

3.4.2 Semantic Similarity Ratings

Table 3.2 shows the semantic similarity ratings for the translated word pairs. Column *WP* is the number of the word pair as shown in Table 3.1. Column *Thai Human Rating* is the human rating for the Thai word pairs. Column *Thai Machine Rating* is the machine rating for the Thai word pairs using TWSS described in Section 3.2. Column *English Human Rating* is the human rating obtained from Rubenstein and Goodenough (1965) for the purposes of comparison. Column *English Machine Rating* is the machine rating for the English word pairs using Li’s word measure (Li et al., 2003). Human ratings are calculated as the mean of the ratings provided by the set of participants for each word pair. All of the measures have been scaled in the range 0 to 1 to aid comparison.

Table 3.2: Semantic Similarity of Human Rating and Machine Rating

WP	Thai		English	
	Thai Human Rating	Thai Machine Rating	English Human Rating	English Machine Rating
1	0.020	0.097	0.005	0.070
5	0.006	0.070	0.015	0.050
9	0.017	0.016	0.048	0.156
13	0.171	0.110	0.110	0.107
17	0.158	0.322	0.212	0.320
21	0.150	0.365	0.240	0.366
25	0.137	0.176	0.250	0.175
29	0.149	0.145	0.310	0.200
33	0.541	0.322	0.370	0.320
37	0.315	0.298	0.455	0.245
41	0.325	0.365	0.652	0.366
47	0.403	0.448	0.778	0.548
48	0.393	0.991	0.802	0.366
49	0.605	1.000	0.822	0.817
50	0.221	0.214	0.852	0.814
51	0.781	0.818	0.862	0.817
52	0.583	0.996	0.865	0.667
53	0.836	0.544	0.865	0.818
54	0.697	0.819	0.895	0.547
55	0.806	0.816	0.898	0.818
56	0.805	0.801	0.900	0.817
57	0.708	0.978	0.912	1.000
58	0.834	0.816	0.915	0.817
59	0.379	1.000	0.920	1.000
60	0.606	0.811	0.955	0.670
61	0.509	0.816	0.960	0.817
62	0.850	0.999	0.970	1.000
63	0.770	1.000	0.980	1.000
64	0.752	1.000	0.985	1.000
65	0.769	0.999	0.985	1.000

3.4.3 Discussion

According to O’Shea et al. (2013), the Pearson correlation coefficient has been suitable for measuring the assumption between human and machine rating of semantic similarity since the 1960s (Rubenstein and Goodenough, 1965).

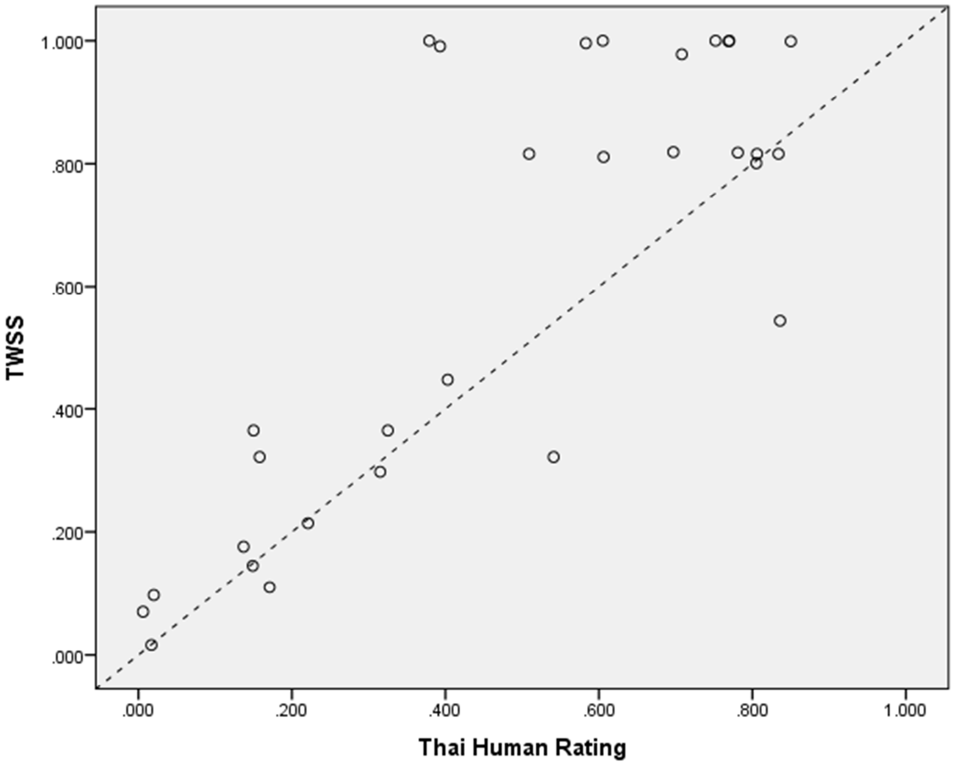


Figure 3.4: The Correlation between Thai Human Rating and Thai Machine Rating

The experimental results in Table 3.2 suggest that the TWSS measure and semantic similarity of human rating provides good results. As can be seen in Figure 3.4, most of the data points are near the linear line (dotted line). The data points indicate how well the measure performs. The closer the data point to the linear line, the better the measure performs. The Pearson Product-Moment correlation coefficients obtained from these results were:

- Pearson’s r = 0.823 (P-Value < 0.01)

Table 3.3 illustrates the agreement of both of the machine measures with human ratings by calculating the Pearson Product-Moment correlation coefficients between the human ratings and the machine ratings over the dataset. It is important to investigate how effective the semantic similarity measure is. This can be achieved by comparing its performance with the ‘average’ human. Also, the upper and lower of the expected performance can be set using the correlation for the best and worst humans. Leave-one-out resampling technique (Resnik, 1995) is used to find the correlation coefficient of each participant with rest of the group and calculating the average.

Table 3.3: The Pearson Product-Moment Correlation Coefficients.

	Correlation *r*	P-value
Thai human similarity rating and TWSS	0.823	0.000
English human similarity rating and Li's measure	0.911	0.000
Thai human similarity rating and English human similarity rating	0.857	0.000
Average of the correlation of all participant	0.842	-
Worst Thai native speaker participant and the rest of the group	0.606	-
Best Thai native speaker participant and the rest of the group	0.933	-

Table 3.3 shows the Pearson Product-Moment correlation coefficients. The Thai machine measure performs close to the English machine measure, with a difference of 0.088 between the two correlation coefficients. The Thai machine measure also performs better than the correlation between the worst performing human and the rest of the group (r = 0.606), which supports the view that it could form the basis of an effective algorithm. Furthermore, because the best performing human achieved the correlation of 0.933, it shows this benchmark dataset is capable of measuring considerable improvement over the current algorithm and should be useful to researchers on Thai semantic similarity.

Word pairs 37 (Magician-Oracle) and 41 (Oracle-Sage) in Table 3.2 illustrate an interesting problem. Both pairs of nouns contain the word *Oracle*. In general, *Oracle* means either 'a message given by an oracle' or 'someone who gave advice to people or told them what would happen'; the definition can be found in the Longman Dictionary (Mayor, 2009). In this work, the first meaning was taken from the Thai-English dictionary (Trakultaweekoon, 2007), which is *คำทำนาย* is likely to mean 'prediction'. After we translated the word back to English via Google Translate, the first meaning from the Google translation was chosen, and is *prophecy*. Consequently, the TWSS rating that was obtained was low because their subsumer is *entity*. The human rating was significantly higher than the machine rating, as shown in Table 3.2. This shows that the way that the TWSS calculates the rating for pairs of nouns is based on only this first meaning that comes up in the dictionary. In a debriefing session after the experiment, the participants reported selecting a word sense based on all of their personal knowledge of a word. The TWSS cannot predict which sense a human will use. Table 3.4 illustrates the words found to raise problems of ambiguity during translation.

Table 3.4: The Exception of Translate Word

R&G word	Thai word	Google word	R&G word	Thai word	Google word
Cord	สายไฟ	Wire	Voyage	การท่องเที่ยว	Travel
String	เชือก	Rope	Shore	ชายฝั่ง	Coast
Sage	นักปราชญ์	Savant	Autograph	ลายมือชื่อ	Signature
Oracle	คำทำนาย	Prophecy	Jewel	เพชรพลอย	Gem
Cushion	เบาะ	Pad	Stove	เตาไฟ	Fireplace
Rooster	นกตัวผู้	Bird	Wizard	พ่อมด	Necromancer
Woodland	ป่าเขา	Forest	Implement	อุปกรณ์	Equipment
Serf	ทาส	Thrall	Asylum	ที่หลบภัย	Shadow
Automobile	รถยนต์	Car	Mound	ภูเขา	Mountain
Journey	การเดินทาง	Travel			

Moreover, word pair 64 (Midday-Noon) in Table 3.2 also illustrates another problem. Both words translate to the same word, which is 'เที่ยงวัน' in Thai, by using the Google translation. Nevertheless, in Thai culture, the word *Midday* (เที่ยงวัน) means 12.00pm but the word *Noon* (กลางวัน) can mean around 11.00am – 13.00pm. This is one of the reasons why the Thai Human rating for word pair 64 (0.752) is significantly different from the English Human rating (0.985). This means the measure also cannot fully predict the ratings for those words that have different meanings in the Thai culture.

The paired sample t-test was used to find whether or not Thai Human rating and TWSS rating over the dataset were statistically significantly different (α=0.05) from the hypotheses:

- H_0: There is no statistically significant difference between Human rating and TWSS rating.
- H_1: There is a statistically significant difference between Human rating and TWSS rating.

The result is:

- t = -3.439, df = 29 (P-Value < 0.01)

As a result, the null hypothesis is rejected. This supports the view that the ratings by Human are statistically significantly different from the rating of TWSS in that procedure. This means there is room for improvement of TWSS, and TWS-30 is capable of measuring future improvement.

Another paired sample t-test was conducted find whether or not Thai Human ratings and English Human ratings from Rubenstein and Goodenough (1965) over the dataset were statistically significantly different (α=0.05) from the hypotheses:

- H_0: There is no statistically significant difference between Thai Human ratings and English Human ratings.
- H_1: There is a statistically significant difference between Thai Human ratings and English Human ratings.

The result is:

- $t = -3.439$, $df = 29$ (P-Value < 0.01)

From the result, the null hypothesis is rejected and that means the ratings by Thai Human are statistically significantly different from the ratings by English Human in that procedure. This can be explained as the TWS-30 is a dataset whereby the word pairs are based on English and translated into Thai. The participants of TWS-30 were native Thai speakers and were asked to rate English based word pairs and that makes these two datasets statistically significantly different. However, the correlation coefficient between the two datasets is still high ($r = 0.857$).

The benchmark dataset from Section 3.3.4 is a dataset that represents a subset of 30 word pairs chosen from the R&G dataset. Because this dataset was created based on English words, the TWSS could not perform very well with the word pairs 54 and 64. Moreover, the Human rating and TWSS rating are statistically significantly different from the dataset. To clarify this particular problem, a benchmark data set with Thai culture needed to be created. This will be explained in more detail in Chapter 4.

Although the dataset is small, it illustrates that the semantic meaning of words when translated from English to Thai is lost. The result of this research is encouraging, however, and indicates the potential for the creation of a TWSS measure.

3.5 Conclusion

This chapter described how the prototype TWSS measure work was developed and described the methodology for the creation of a Thai benchmark dataset (TWS-30) from human participants, as well as discussing the experimental results. This work was published in Osathanunkul (2011). As mentioned in Section 3.4.3, this measure cannot fully predict those word pairs that relate to Thai culture as TWS-30 was built based on an English dataset (Rubenstein and Goodenough, 1965). Thus, to experiment on words related to the Thai culture, a more effective evaluation is needed before it is possible to accept or reject a particular algorithm as a component of a Thai STSS measure. Therefore, a new benchmark dataset based on Thai culture is needed. This will be discussed in Chapter 4.

Chapter 4

65 Word Pair Thai Benchmark Dataset

(TWS-65)

4.1 Introduction

Chapter 3 established the potential of a Thai word semantic similarity measure (TWSS). However, TWSS establishes a baseline against which the performance of a specifically Thai-oriented measure can be compared. Chapter 3 also identified shortcomings in the English-oriented evaluation benchmark dataset. Further development requires an expanded Thai word similarity benchmark dataset. Therefore, the aim of this chapter is to create a Thai benchmark dataset (TWS-65) based on Thai culture and should provide a more effective evaluation. To date, no prior work has been reported on Thai word benchmark datasets. The question is what the right way to create one is; the answer is that there is no right way to do so. However, following procedures previously practised in other languages prior to the Thai language should prove effective. Hence, this research will create a Thai benchmark dataset following the Rubenstein and Goodenough (1965) procedure, and yet Thai culture will be taken into account in terms of creation. This TWS-65 will contain 65 word pairs, the same amount as the original R&G benchmark dataset. This chapter will describe the methodology for the creation of TWS-65 and will discuss:

- The selection of theme words
- The selection of word pairs which separate into three distinct categories:
    - High similarity word pairs
    - Medium similarity word pairs
    - Low similarity word pairs
- Collecting ratings from Thai native speakers for all the word pairs.

The aim of this chapter is to investigate the research question: Can a WordNet based English word similarity measure produce a similarity rating between words based on Thai culture?

The contributions in this chapter are:

- A methodology for creating TWS-65
- Application of the methodology to rating TWS-65
- Evaluation of TWS-30 with TWS-65
- Evaluation of TWSS with TWS-65.

The rest of this chapter is organized as follows: Section 4.2 sets out the method of selecting theme words for TWS-65; Section 4.3 describes the method of forming high, medium, and low word pairs.; Section 4.4 describes the collection of rating for the Thai word similarity benchmark dataset from the participants using a method based on Miller and Charles

(1991) and O’Shea et al. (2008) with a discussion of TWS-65 versus. TWS-30; Section 4.5 compares human ratings with TWSS ratings over the TWS-65 dataset and Section 4.6 is the conclusion.

4.2 Theme Words

Prior to the creation of TWS-65, theme words related to the Thai culture need to be established. Also, since theme words are required to represent Thai culture, they cannot simply be a wholesale replication of Rubenstein and Goodenough (1965). The R&G dataset methodology (Rubenstein and Goodenough, 1965) is chosen to produce a TWS-65, as discussed in Section 2.4.3. However, The R&G dataset was published without grounds for the specific choices of 48 nouns and the method of choosing the word pairs.

As Rubenstein and Goodenough word pairs are a good starting point for creating a TWS-65, the first sixteen pairs of theme words which are related to the Thai culture were adopted from Rubenstein and Goodenough (1965) whose work was first produced in 1965 and has been extensively referenced up to the present day. There is no evidence of categories for each word pair in the Rubenstein and Goodenough dataset. However, Battig and Montague (1969) provide a good source of categories, some of which map to Rubenstein and Goodenough categories. The six pairs of theme words are referred from Battig and Montague (1969), which separate nouns into 56 categories. These pairs cover certain noun categories missing from Rubenstein and Goodenough and these six pairs of theme words are related to Thai culture. Now 22 pairs of theme words are selected to create TWS-65. However, the Rubenstein and Goodenough (1965) dataset used 24 pairs of theme words to create 65 word pairs. The additional two pairs are listed by native Thai speakers (NTS). These two pairs are regarded as semantically similar solely in the Thai language and have been widely agreed by over 20 native Thai speakers.

Table 4.1 shows the list of theme words. Column *WP* shows the number of word pairs. Column *List of theme words* shows 24 pairs of theme words in Group A and Group B. Column *Source* indicates the category of theme word pairs where R&G is Rubenstein and Goodenough, B&M is Battig and Montague, and NTS is native Thai speakers.

Table 4.2: List of Theme Words

WP	**List of theme words**				***Source***
	Group A		**Group B**		
1	Autograph	ลายมือชื่อ	Signature	ลายเซ็น	R&G
2	Boy	เด็กผู้ชาย	Lad	เด็กหนุ่ม	R&G
3	Coast	ฝั่งทะเล	Shore	ชายฝั่ง	R&G
4	Cemetery	ป่าช้า	Graveyard	สุสาน	R&G
5	Journey	การเดินทาง	Voyage	การท่องเที่ยว	R&G
6	Slave	ทาส	Serf	ข้ารับใช้	R&G
7	Implement	อุปกรณ์	Tool	เครื่องมือ	R&G
8	Midday	เที่ยงวัน	Noon	กลางวัน	R&G
9	Gem	อัญมณี	Jewel	เพชรพลอย	R&G
10	Hill	เนินเขา	Mound	ภูเขา	R&G
11	Forest	ป่าไม้	Woodland	พงไพร	R&G
12	Automobile	ยานพาหนะ	Car	รถยนต์	R&G
13	Food	อาหาร	Fruit	ผลไม้	R&G
14	Glass	แก้ว	Tumbler	ถ้วย	R&G
15	Priest	นักบวช	Monk	พระ	R&G
16	Magician	นักมายากล	Wizard	พ่อมด	R&G
17	Cotton	ผ้าฝ้าย	Silk	ผ้าไหม	B&M
18	Teacher	ครู	Lecturer	อาจารย์	B&M
19	Magazine	นิตยสาร	Book	หนังสือ	B&M
20	Temple	วัด	Church	โบสถ์	B&M
21	Uncle	ลุง	Aunt	ป้า	B&M
22	Dog	สุนัข	Dog	หมา	B&M
23	Cinema	โรงภาพยนต์	Theatre	โรงละคร	NTS
24	Plant	พืช	Tree	ต้นไม้	NTS

4.3 Methodology of Selecting Word Pairs from Theme Words

The aim of this section is to describe the methodology for selecting Thai word pairs for TWS-65 which follows the same way as the Rubenstein and Goodenough (1965) dataset. Rubenstein and Goodenough separated 65 word pairs into three classes: 20 high similarity word pairs; 21 medium similarity word pairs; and 24 low similarity word pairs. TWS-65 will have the same number of word pairs in any range of similarity as Rubenstein and Goodenough. According to Rubenstein and Goodenough (1965), Miller and Charles (1991) and O’Shea et al. (2008), the most difficult word pairs to select are in the medium similarity range. To achieve this, an experiment needs to be conducted. Also, an experiment was also conducted to find high similarity pairs rather than relying on the author’s subjective opinion. However, low similarity word pairs are easy to construct as most of the word pairs that are constructed are likely to be low similarity pairs. TWS-65 word pairs will be presented in Section 4.3.4. This experiment is separated into two phases as follows:

- Phase 1: Selecting high semantic similarity word pairs
- Phase 2: Selecting medium semantic similarity word pairs.

Those two phases of the experiment were done by the same participants on the same day.

4.3.1 High Semantic Similarity Word Pair

The aim of this section is to describe the methodology for finding high similarity word pairs.

4.3.1.1 Participants

To select high similarity word pairs, 20 native Thai speakers were selected. The participants had 9 Arts or Humanities backgrounds and 11 Science or Engineering backgrounds. They consisted of 8 undergraduates and 12 postgraduates studying in 6 different UK universities. The average age of the participants was 24; standard deviation was 4.8, with 11 males and 9 females. These participants were not the same participants who rated the pairs in TWS-30.

4.3.1.2 Materials

Twenty-four candidate theme word pairs were chosen from the Rubenstein and Goodenough, Battig and Montague, and Thai native speakers, as shown in Table 4.2. Each theme word pair was printed in separate groups (A and B), as shown in Table 4.1, in a random order using a standard Thai font. A questionnaire was produced containing instructions for choosing high similarity word pairs and specifying a small amount of personal data (Name, Confirmation of being a native Thai speaker, Age, Gender, and Academic background). The examples of experimental materials are:

- Appendix 1.5 The Person Data Collection Sheet
- Appendix 2.1 The Instruction Sheet
- Appendix 2.2 List of Theme Words
- Appendix 2.3 High Similarity Word Pairs Recording Sheet.

4.3.1.3 Procedure

The participants were asked to perform the following procedure:

1. Please read through all words in Group A and Group B.
2. Please enter the best 20 word pairs that you think are strongly related in meaning. Each pair of words chosen should have one from group A and one from group B.

4.3.1.4 Results

The high semantic similarity word pairs produced from this experiment are shown in Table 4.2. Twenty word pairs were chosen to be high similarity word pairs. This number of high similarity word pairs is the same number as Rubenstein and Goodenough (1965) high similarity word pairs in their dataset. Columns W_1 and W_2 are the word pairs. Participants in the column *Number* indicate the number of participants choosing these word pairs as high semantic similarity. As there were 20 participants, the maximum number is 20.

Table 4.2: High Similarity Word Pairs

W_1		W_2		*Number*
เด็กผู้ชาย	Boy	เด็กหนุ่ม	Lad	20
เที่ยงวัน	Midday	กลางวัน	Noon	20
โรงภาพยนต์	Cinema	โรงละคร	Theatre	20
การเดินทาง	Journey	การท่องเที่ยว	Voyage	20
ครู	Teacher	อาจารย์	Lecturer	20
นักบวช	Priest	พระ	Monk	20
ฝั่งทะเล	Coast	ชายฝั่ง	Shore	20
สุนัข	Dog	หมา	Dog	20
อัญมณี	Gem	เพชรพลอย	Jewel	19
เนินเขา	Hill	ภูเขา	Mountain	19
ยานพาหนะ	Automobile	รถยนต์	Car	18
ลายมือชื่อ	Autograph	ลายเซ็น	Signature	18
อุปกรณ์	Implement	เครื่องมือ	Tool	18
นิตยสาร	Magazine	หนังสือ	Book	17
ป่าไม้	Forest	พงไพร	Woods	17
ป่าช้า	Cemetery	สุสาน	Graveyard	16
ทาส	Slave	ข้ารับใช้	Serf	15
ผ้าฝ้าย	Cotton	ผ้าไหม	Silk	15
พืช	Plant	ต้นไม้	Tree	14
แก้ว	Glass	ถ้วย	Cup	12

4.3.2 Medium Semantic Similarity Word Pair

The aim of this section is to describe the methodology for finding medium similarity word pairs.

4.3.2.1 Participants

The medium similarity word pairs were collected from 20 native Thai speakers. The 20 native Thai speakers were the same participants who selected the high similarity word pairs.

4.3.2.2 Materials

Twenty-four pairs of theme words were selected from the Rubenstein and Goodenough, Battig and Montague and Thai native speakers, as shown in Table 4.2. Each theme word pair was printed in separate groups (A and B), as shown in Table 4.1, in random order using a standard Thai font. A questionnaire was produced containing instructions for choosing medium similarity word pairs and a small amount of personal data (Name, Confirmation of being a native Thai speaker, Age, Gender, and Academic background). The examples of experimental materials are:

- Appendix 1.5 The Personal Data Collection Sheet
- Appendix 2.2 List of Theme Words
- Appendix 2.4 The Instruction Sheet
- Appendix 2.5 Medium Similarity Word Pairs Recording Sheet.

4.3.2.3 Procedure

The participants were asked to perform the following procedure:

1. Please read through all words in Group A and Group B.
2. Please enter the best 21 pairs of words that you think are related in meaning, which have not been selected in the High semantic similarity word pairs. Each pair of words chosen should have one from group A and one from group B.

4.3.2.4 Results

The medium semantic similarity word pairs obtained from the experiment are shown in Table 4.3. Twenty-one word pairs were chosen to be medium similarity word pairs. This number of medium similarity word pairs is the same number as Rubenstein and Goodenough (1965) medium similarity word pairs in their dataset. Columns W_1 and W_2 are the word pairs. Participants in the column *Number* indicate that the number of participants choosing these word pairs. As there were 20 participants, the maximum number is 20. This number also includes the number of participants that were selecting these word pairs as high similarity word pairs; i.e. the column *Number* is the number of participants choosing these word pairs as high and medium similarity word pairs.

Table 4.3: Medium Similarity Word Pairs

W_1		*W_2*		*Number*
วัด	Temple	โบสถ์	Church	20
อาหาร	Food	ผลไม้	Fruit	18
นักมายกล	Magician	พ่อมด	Wizard	18
ลุง	Uncle	ป้า	Aunt	18
วัด	Temple	สุสาน	Graveyard	15
นักบวช	Priest	พ่อมด	Wizard	14
พืช	Plant	พงไพร	Woods	14
ป่าไม้	Forest	ต้นไม้	Tree	14
ผ้าฝ้าย	Cotton	ต้นไม้	Tree	12
วัด	Temple	พระ	Monk	12
เด็กผู้ชาย	Boy	อาจารย์	Lecturer	11
อุปกรณ์	Implement	รถยนต์	Car	11
เนินเขา	Hill	ชายฝั่ง	Shore	10
ลุง	Uncle	อาจารย์	Lecturer	10
ป่าไม้	Forest	ภูเขา	Mountain	10
นักมายกล	Magician	เครื่องมือ	Tool	9
วัด	Temple	พงไพร	Woods	9
ครู	Teacher	ป้า	Aunt	9
ป่าไม้	Forest	ผลไม้	Fruit	8
สุนัข	Dog	เด็กหนุ่ม	Lad	8
ฝั่งทะเล	Coast	พงไพร	Woods	8

4.3.3 Low Semantic Similarity Word Pair

Twenty-four low semantic similarity word pairs were chosen at random from the theme word pairs, and these low similarity word pairs are not the same as the high similarity word pairs or medium similarity word pairs. Moreover, these low similarity word pairs were screened to avoid any higher similarity word pairs by chance.

4.3.4 TWS-65 Word Pairs Dataset

Table 4.4 exhibits the TWS-65 candidate word pairs. Column *Source* indicates the similarity category of the word pairs. Columns *W_1* and *W_2* are the word pairs. The next step is to obtain the actual Human rating for the word pairs.

Table 4.4: TWS-65 Word Pairs

Source	*W_1*		*W_2*	
High	เด็กผู้ชาย	Boy	เด็กหนุ่ม	Lad
High	เที่ยงวัน	Midday	กลางวัน	Noon
High	โรงภาพยนต์	Cinema	โรงละคร	Theatre
High	การเดินทาง	Journey	การท่องเที่ยว	Voyage
High	ครู	Teacher	อาจารย์	Lecturer
High	นักบวช	Priest	พระ	Monk

High	ฝั่งทะเล	Coast	ชายฝั่ง	Shore
High	สุนัข	Dog	หมา	Dog
High	อัญมณี	Gem	เพชรพลอย	Jewel
High	เนินเขา	Hill	ภูเขา	Mountain
High	ยานพาหนะ	Automobile	รถยนต์	Car
High	ลายมือชื่อ	Autograph	ลายเซ็น	Signature
High	อุปกรณ์	Implement	เครื่องมือ	Tool
High	นิตยสาร	Magazine	หนังสือ	Book
High	ป่าไม้	Forest	พงไพร	Woods
High	ป่าช้า	Cemetery	สุสาน	Graveyard
High	ทาส	Slave	ข้ารับใช้	Serf
High	ผ้าฝ้าย	Cotton	ผ้าไหม	Silk
High	พืช	Plant	ต้นไม้	Tree
High	แก้ว	Glass	ถ้วย	Cup
Medium	วัด	Temple	โบสถ์	Church
Medium	อาหาร	Food	ผลไม้	Fruit
Medium	นักมายากล	Magician	พ่อมด	Wizard
Medium	ลุง	Uncle	ป้า	Aunt
Medium	วัด	Temple	สุสาน	Graveyard
Medium	นักบวช	Priest	พ่อมด	Wizard
Medium	พืช	Plant	พงไพร	Woods
Medium	ป่าไม้	Forest	ต้นไม้	Tree
Medium	ผ้าฝ้าย	Cotton	ต้นไม้	Tree
Medium	วัด	Temple	พระ	Monk
Medium	เด็กผู้ชาย	Boy	อาจารย์	Lecturer
Medium	อุปกรณ์	Implement	รถยนต์	Car
Medium	เนินเขา	Hill	ชายฝั่ง	Shore
Medium	ลุง	Uncle	อาจารย์	Lecturer
Medium	ป่าไม้	Forest	ภูเขา	Mountain
Medium	นักมายากล	Magician	เครื่องมือ	Tool
Medium	วัด	Temple	พงไพร	Woods
Medium	ครู	Teacher	ป้า	Aunt
Medium	ป่าไม้	Forest	ผลไม้	Fruit
Medium	สุนัข	Dog	เด็กหนุ่ม	Lad
Medium	ฝั่งทะเล	Coast	พงไพร	Woods
Low	ฝั่งทะเล	Coast	รถยนต์	Car
Low	อาหาร	Food	ถ้วย	Cup
Low	เที่ยงวัน	Midday	โรงละคร	Theatre
Low	เที่ยงวัน	Midday	การท่องเที่ยว	Voyage
Low	ยานพาหนะ	Automobile	เพรชพลอย	Jewel
Low	ทาส	Slave	หมา	Dog
Low	สุนัข	Dog	เครื่องมือ	Tool
Low	การเดินทาง	Journey	สุสาน	Graveyard
Low	อัญมณี	Gem	ลายเซ็น	Signature
Low	การเดินทาง	Journey	กลางวัน	Noon
Low	แก้ว	Glass	ข้ารับใช้	Serf
Low	อาหาร	Food	ลายเซ็น	Signature
Low	เด็กผู้ชาย	Boy	หมา	Dog

Low	นิตยสาร	Magazine	ป้า	Aunt
Low	นักบวช	Priest	หนังสือ	Book
Low	สุนัข	Dog	เด็กหนุ่ม	Lad
Low	วัด	Temple	พงไพร	Woods
Low	ป่าช้า	Cemetery	หมา	Dog
Low	ฝั่งทะเล	Coast	พงไพร	Woods
Low	นักมายากล	Magician	เครื่องมือ	Tool
Low	ครู	Teacher	หนังสือ	Book
Low	พืช	Plant	ผ้าไหม	Silk
Low	เนินเขา	Hill	ผลไม้	Fruit
Low	นักมายากล	Magician	ถ้วย	Cup

4.4 Application of the Methodology to Rating for TWS-65

The aim of this section is to describe the methodology for evaluating TWS-65. Also, the 65 Thai word pair benchmark dataset (TWS-65) that can be used to evaluate with TWSS are presented in Section 4.4.4.

4.4.1 Participants

Similarity ratings were collected from 40 native Thai speakers to create a benchmark dataset. Those 40 native Thai speakers were different from the participants who selected the word pairs. The participants had an equal number of Art/Humanities and Science/Engineering backgrounds. They consisted of 15 undergraduates and 25 postgraduates studying in 2 different UK universities and 4 Thai universities. The average age of the participants was 24 and standard deviation was 2.6, with 22 males and 18 females. This is comparable with student participant groups used for English word similarity for both Rubenstein and Goodenough (1965) and Miller and Charles (1991).

4.4.2 Materials

Following previous practice, Miller and Charles (1991) and O’Shea et al. (2008), the representative subset of 65 word pairs in Table 4.4 were used. Each word pair was printed on a separate card using a standard Thai font. A questionnaire was produced containing instructions for recording similarity ratings and a small amount of personal data (Name, Confirmation of being a native Thai speaker, Age, Gender, and Academic background). Semantic anchors were also provided to guide the participants. The examples of experimental materials are:

- Appendix 1.1 The Ethics Statement
- Appendix 1.2 The Instruction Sheet

- Appendix 1.3 A Sample Card
- Appendix 1.5 The Person Data Collection Sheet
- Appendix 1.6 Semantic Anchors
- Appendix 1.7 Sample Rating Recording Sheet.

4.4.3 Procedure

The participants were asked to perform the following procedure:

1. Please sort the cards into four groups in a rough order of the similarity of meaning of the word pair.
2. After sorting the cards into groups, order the cards in each group according to similarity of meaning (i.e. the card that contains the lowest similarity of meaning is at the top of the group).
3. Please recheck the cards in every group. You may change a pair of words to other groups at this stage.
4. Please rate the semantic similarity rating of each pair of words by writing a number between 0.0 (minimum similarity) and 0.9 for the first group, 1.0 and 1.9 for the second group, 2.0 to 2.9 for the third group, 3.0 and 4.0 (maximum similarity) for the fourth group on the recording sheet. You can use the first decimal place (e.g. 2.5) to show finer degrees of similarity. You also may assign the same value to more than one pair.

The cards were shuffled into a random order before being given to the participants.

4.4.4 TWS-65

TWS-65 is shown in Table 4.5. These word pairs are the original words pairs from Section 4.3.4 with the average Thai human participant rating. Column *WP* is the number of the word pair. Columns W_1 and W_2 are the word pairs. Column *Human* is the average similarity rating from the 40 native Thai speakers. Column *SD* is the standard deviation.

Table 4.5: The Average of Similarity Rating from 40 Native Thai Speakers

WP	$\boldsymbol{W_1}$		$\boldsymbol{W_2}$		***Human***	***SD***
1	แก้ว	Glass	ข้ารับใช้	Serf	0.058	0.162
2	อาหาร	Food	ลายเซ็น	Signature	0.068	0.206
3	อัญมณี	Gem	ลายเซ็น	Signature	0.098	0.251
4	ฝั่งทะเล	Coast	รถยนต์	Car	0.110	0.257
5	สุนัข	Dog	เครื่องมือ	Tool	0.123	0.335
6	การเดินทาง	Journey	สุสาน	Graveyard	0.135	0.350

7	เที่ยงวัน	Midday	โรงละคร	Theatre	0.175	0.506
8	เที่ยงวัน	Midday	การท่องเที่ยว	Voyage	0.225	0.509
9	ยานพาหนะ	Automobile	เพรชพลอย	Jewel	0.243	0.511
10	เนินเขา	Hill	ผลไม้	Fruit	0.278	0.390
11	นักมายากล	Magician	ถ้วย	Cup	0.278	0.499
12	ป่าช้า	Cemetery	หมา	Dog	0.293	0.506
13	ฝั่งทะเล	Coast	พงไพร	Woods	0.318	0.310
14	นักมายากล	Magician	เครื่องมือ	Tool	0.320	0.541
15	การเดินทาง	Journey	กลางวัน	Noon	0.343	0.581
16	นิตยสาร	Magazine	ป้า	Aunt	0.415	0.560
17	นักบวช	Priest	หนังสือ	Book	0.420	0.768
18	เด็กผู้ชาย	Boy	หมา	Dog	0.440	0.557
19	สุนัข	Dog	เด็กหนุ่ม	Lad	0.530	0.743
20	วัด	Temple	พงไพร	Woods	0.540	0.818
21	ทาส	Slave	หมา	Dog	0.555	0.636
22	อาหาร	Food	ถ้วย	Cup	0.650	0.846
23	ครู	Teacher	หนังสือ	Book	0.983	0.986
24	พืช	Plant	ผ้าไหม	Silk	1.043	0.988
25	เด็กผู้ชาย	Boy	อาจารย์	Lecturer	1.083	1.007
26	โรงภาพยนต์	Cinema	โบสถ์	Church	1.095	1.093
27	ทาส	Slave	เด็กหนุ่ม	Lad	1.160	0.884
28	เนินเขา	Hill	ชายฝั่ง	Shore	1.175	0.905
29	ยานพาหนะ	Automobile	เครื่องมือ	Tool	1.265	0.996
30	ผ้าฝ้าย	Cotton	ต้นไม้	Tree	1.283	1.020
31	อุปกรณ์	Implement	รถยนต์	Car	1.336	0.831
32	ลุง	Uncle	อาจารย์	Lecturer	1.410	1.000
33	ป่าไม้	Forest	ผลไม้	Fruit	1.551	0.805
34	ครู	Teacher	ป้า	Aunt	1.625	0.808
35	นักบวช	Priest	พ่อมด	Wizard	1.720	0.890
36	แก้ว	Glass	เพชรพลอย	Jewel	1.923	1.007
37	นักมายากล	Magician	พ่อมด	Wizard	2.010	0.969
38	วัด	Temple	สุสาน	Graveyard	2.150	0.805
39	พืช	Plant	พงไพร	Woods	2.210	1.005
40	ป่าไม้	Forest	ภูเขา	Mountain	2.255	1.045
41	อาหาร	Food	ผลไม้	Fruit	2.363	0.935
42	แก้ว	Glass	ถ้วย	Cup	2.413	0.922
43	วัด	Temple	พระ	Monk	2.675	0.991
44	ลุง	Uncle	ป้า	Aunt	2.743	0.980
45	ป่าไม้	Forest	ต้นไม้	Tree	2.905	0.848
46	โรงภาพยนต์	Cinema	โรงละคร	Theatre	3.018	0.774
47	เนินเขา	Hill	ภูเขา	Mountain	3.023	0.787
48	เด็กผู้ชาย	Boy	เด็กหนุ่ม	Lad	3.030	0.407
49	ผ้าฝ้าย	Cotton	ผ้าไหม	Silk	3.050	0.758
50	ยานพาหนะ	Automobile	รถยนต์	Car	3.105	0.851
51	ฝั่งทะเล	Coast	ชายฝั่ง	Shore	3.118	0.408
52	อุปกรณ์	Implement	เครื่องมือ	Tool	3.120	0.856
53	ทาส	Slave	ข้ารับใช้	Serf	3.140	0.473
54	การเดินทาง	Journey	การท่องเที่ยว	Voyage	3.188	0.358

55	นิตยสาร	Magazine	หนังสือ	Book	3.198	0.620
56	ลายมือชื่อ	Autograph	ลายเซ็น	Signature	3.210	0.412
57	เที่ยงวัน	Midday	กลางวัน	Noon	3.235	0.445
58	ป่าไม้	Forest	พงไพร	Woods	3.303	0.438
59	อัญมณี	Gem	เพชรพลอย	Jewel	3.318	0.346
60	พืช	Plant	ต้นไม้	Tree	3.410	0.376
61	นักบวช	Priest	พระ	Monk	3.575	0.311
62	ป่าช้า	Cemetery	สุสาน	Graveyard	3.625	0.323
63	วัด	Temple	โบสถ์	Church	3.693	0.230
64	ครู	Teacher	อาจารย์	Lecturer	3.783	0.262
65	สุนัข	Dog	หมา	Dog	3.923	0.129

4.4.5 Evaluation of TWS-65 with TWS-30

TWS-30 and TWS-65 ratings were collected from different groups of participants, but both datasets were rated by using the same procedure. Table 4.6 shows the Pearson product moment correlation coefficients of TWS-65 with 40 participants; the leave-one-out resampling technique was used to find the correlation coefficient of each participant with the rest of the group.

Table 4.6: TWS-65 Correlation Coefficients with Mean Human Judgment

	Correlation *r*
Average of the correlation of all participants	0.883
Worst participant	0.681
Best Participant	0.937

The ANOVA test was used to find whether or not TWS-30 and TWS-65 were statistically significantly different ($\alpha=0.05$) from the hypotheses:

- H_0: There is no statistically significant difference between the two datasets.
- H_1: There is a statistically significant difference between the two datasets.

To do this, the 14 word pairs, which were common to both datasets, were used. Table 4.7 shows the 14 word pairs ratings from both datasets. Columns W_1 and W_2 are the word pairs in English. Column *TWS-30* is the similarity rating from TWS-30 and column *TWS-65* is the similarity rating from TWS-65.

Table 4.7: The Average of Similarity Rating for the 14 Word Pairs in Both Datasets

W_1		***W_2***		***TWS-30***	***TWS-65***
ฝั่งทะเล	Coast	พงไพร	Woods	0.632	0.318
นักมายกล	Magician	พ่อมด	Wizard	1.570	2.010
เนินเขา	Hill	ภูเขา	Mountain	2.420	3.023
เด็กผู้ชาย	Boy	เด็กหนุ่ม	Lad	2.425	3.030

การเดินทาง	Journey	การท่องเที่ยว	Voyage	2.788	3.188
ป่าไม้	Forest	พงไพร	Woods	2.830	3.303
เที่ยงวัน	Midday	กลางวัน	Noon	3.008	3.235
อัญมณี	Gem	เพชรพลอย	Jewel	3.075	3.318
ยานพาหนะ	Automobile	รถยนต์	Car	3.080	3.105
ฝั่งทะเล	Coast	ชายฝั่ง	Shore	3.218	3.118
ลายมือชื่อ	Autograph	ลายเซ็น	Signature	3.223	3.210
อุปกรณ์	Implement	เครื่องมือ	Tool	3.335	3.120
ทาส	Slave	ข้ารับใช้	Serf	3.345	3.140
ป่าช้า	Cemetery	สุสาน	Graveyard	3.400	3.625

The result is:

- Pearson’s r = 0.926 (P-Value < 0.01)
- ANOVA test f = 0.318 , df = 1 (P-Value > 0.05)

As a result, the P-Value for the ANOVA test is greater than 0.05, meaning we fail to reject the null hypothesis, and that it is reliable to assume that the human ratings from TWS-30 and human ratings from TWS-65 are not statistically significantly different.

It could be questioned for what reasons the two sets of ratings are not in perfect agreement. It should be noted, firstly, that it is illogical to anticipate perfect agreement (correlation = 1.0). Even when the Rubenstein and Goodenough word experiments were reproduced (employing the Miller & Charles 30-word subset), correlations of 0.97 (Miller and Charles, 1911) and 0.96 (Resnik, 1999) were acquired.

4.5 Evaluation of the Thai Word Semantic Similarity Measure

The aim of this section is to describe a series of experiments that were conducted using the TWS-65 to evaluate the TWSS measure described in Section 3.2. This established a baseline for improvement by a dedicated Thai word similarity measure.

4.5.1 Methodology

The TWS-65 was used to evaluate the TWSS measure. The TWSS rating was obtained by calculating the similarity of English words translated from Thai word pairs, as mentioned in Section 3.2.

- Translate all word pairs in TWS-65 into English via Google translation.
- Calculate TWSS rating for each word pair in TWS-65.

The Pearson Product-Moment correlation coefficients (*r*) between Thai human ratings and TWSS are calculated and presented in Section 4.5.3.

4.5.2 Semantic Similarity Rating Results

Table 4.8 shows the semantic similarity ratings for the translated word pairs. Column *WP* is the number of the word pairs, as shown in Table 4.5. Column *Human* is the human rating for the Thai word pairs. Column *TWSS* is the machine rating for the Thai word pairs using the algorithm (TWSS) described in Section 3.2. All of the measures have been scaled in the range 0 to 1 to aid comparison.

Table 4.8: Semantic Similarity between Human Rating and TWSS

WP	*Human*	*TWSS*	*WP*	*Human*	*TWSS*
1	0.014	0.144	**34**	0.406	0.244
2	0.017	0.216	**35**	0.430	0.365
3	0.024	0.176	**36**	0.481	0.263
4	0.028	0.144	**37**	0.503	0.991
5	0.031	0.044	**38**	0.538	0.097
6	0.034	0.014	**39**	0.553	0.547
7	0.044	0.097	**40**	0.564	0.322
8	0.056	0.044	**41**	0.591	0.144
9	0.061	0.132	**42**	0.603	0.668
10	0.069	0.144	**43**	0.669	0.128
11	0.069	0.157	**44**	0.686	0.445
12	0.073	0.079	**45**	0.726	0.586
13	0.079	0.144	**46**	0.754	0.818
14	0.080	0.097	**47**	0.756	0.656
15	0.086	0.053	**48**	0.758	0.811
16	0.104	0.128	**49**	0.763	0.664
17	0.105	0.145	**50**	0.776	1.000
18	0.110	0.108	**51**	0.779	0.801
19	0.133	0.108	**52**	0.780	0.816
20	0.135	0.197	**53**	0.785	0.544
21	0.139	0.445	**54**	0.797	0.819
22	0.163	0.360	**55**	0.799	0.670
23	0.246	0.105	**56**	0.803	0.816
24	0.261	0.360	**57**	0.809	1.000
25	0.271	0.365	**58**	0.826	0.991
26	0.274	0.448	**59**	0.829	0.999
27	0.290	0.544	**60**	0.853	0.716
28	0.294	0.520	**61**	0.894	0.298
29	0.316	0.244	**62**	0.906	0.999
30	0.321	0.548	**63**	0.923	0.669
31	0.334	0.445	**64**	0.946	0.668
32	0.353	0.244	**65**	0.981	1.000
33	0.388	0.216			

4.5.3 Discussion

The experimental results in Section 4.5.2 suggest that the TWSS measure and semantic similarity of human rating still provides good results. However, there are a number of data points far from the linear line (dotted line), as can be seen in Figure 4.1. The Pearson Product-Moment correlations obtained from these results are:

- Pearson’s r = 0.807 (P-Value < 0.01)

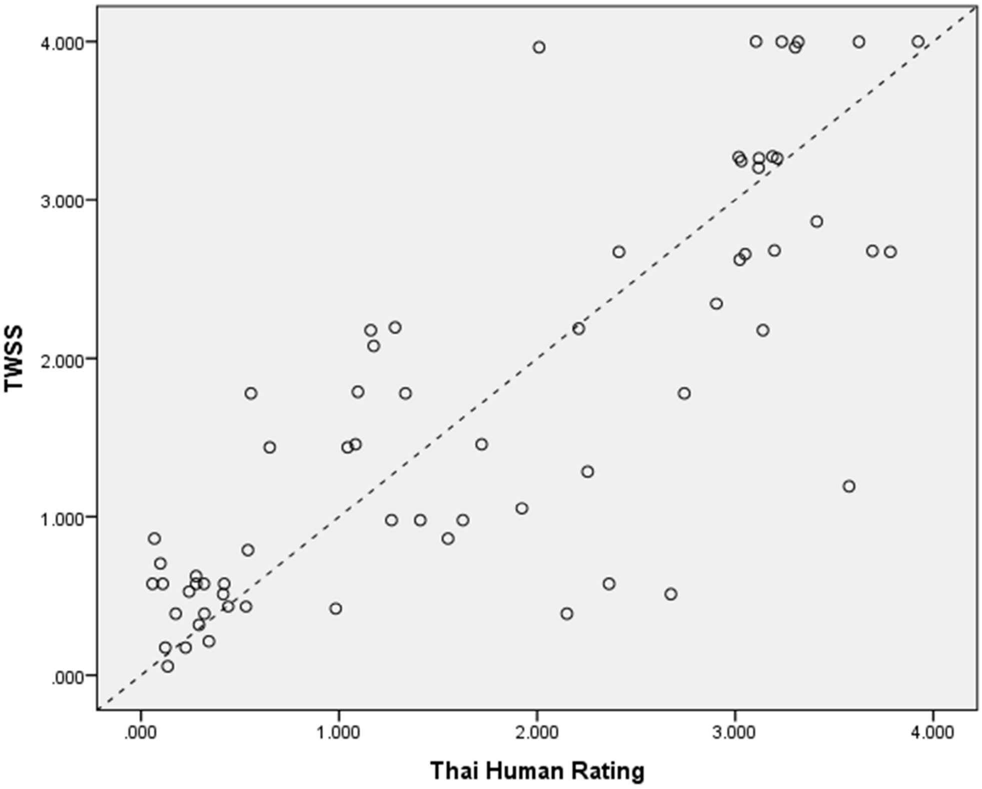


Figure 4.1: The Correlation between TWS-65 Rating and TWSS

Table 4.9 illustrates the agreement of the machine measure with human ratings by calculating the Pearson Product-Moment correlations (r) between the human ratings and the machine ratings over the TWS-65. Also, the correlation coefficients of each participant with the average for the rest of the group over the TWS-65 from Table 4.6 is shown.

Table 4.9: Correlation Coefficients

	Correlation r
Thai human similarity rating and machine similarity measure	0.807
Average of the correlation of all participants	0.883
Worst Thai native speaker participant and the rest of the group	0.708
Best Thai native speaker participant and the rest of the group	0.937

The TWSS performs better than the correlation between the worst performing human and the rest of the group (r = 0.708), which supports the view that it could form the basis of an effective algorithm. Furthermore, because the best performing human achieved the correlation of 0.937, it shows this benchmark dataset is capable of measuring considerable improvement over the current algorithm.

The paired sample t-test was used to find whether or not the Human rating and TWSS rating over the dataset are statistically significantly different (α=0.05) from the hypotheses:

- H_0: There are no statistically significant differences between the Human rating and TWSS rating.
- H_1: There are statistically significant differences between the Human rating and TWSS rating.

The result is:

- t = 0.313, df = 64 (P-Value > 0.05)

From the result, the null hypothesis failed to be rejected; that means the rating procedures by Human are not statistically significantly different from the rating of TWSS. This means that we can accept that the ratings produced by TWSS are representative of human perceptions of similarity over the TWS-65.

Table 4.10: Problem Word Pairs

WP	***W_1***		***W_2***		***Human***	***TWSS***
21	ทาส	Slave	หมา	Dog	0.555	1.781
27	ทาส	Slave	เด็กหนุ่ม	Lad	1.160	2.176
37	นักมายากล	Magician	พ่อมด	Wizard	2.010	3.964
38	วัด	Temple	สุสาน	Graveyard	2.150	0.387
41	อาหาร	Food	ผลไม้	Fruit	2.363	0.578
43	วัด	Temple	พระ	Monk	2.675	0.513
61	นักบวช	Priest	พระ	Monk	3.575	1.194
63	วัด	Temple	โบสถ์	Church	3.693	2.677
64	ครู	Teacher	อาจารย์	Lecturer	3.783	2.671

There are nine word pairs that have different ratings between the Human rating and TWSS rating of more than 1 in TWS-65, as shown in Table 4.10, called the *Problem Word Pairs*. There are six out of nine word pairs and the human rating is higher than the machine rating.

The Pearson Product-Moment correlations between Human rating and TWSS rating obtained from the *Problem Word Pairs* in Table 4.10 are:

- Pearson's r = 0.037 (P-Value > 0.05)

Therefore, it was also worth taking a second opinion in the form of Spearman's ρ (Fenton and Pfleeger, 1998), the Rank correlation coefficient for small sets of data, as well as the Kendall tau rank correlation coefficient (Kendall, 1938).

- Spearman's ρ = 0.083 (P-Value > 0.05)
- Kendall's tau τ = 0.111 (P-Value > 0.05)

For r, ρ, and τ, a value of +1 indicates perfect correlation, 0 indicates no relationship and -1 indicates a perfect negative correlation. The TWSS ratings are statistically significantly different from the Human ratings over the *Problem Word Pairs* (t = 1.232, df = 8, P-value > 0.05). This shows insufficiency in rating performance of the TWSS rating with the *Problem Word Pairs* (r = 0.037). As TWSS using WordNet to perform ratings, is English-based, it results in an inefficient rating performance of these word pairs, which are related mainly to Thai culture. In addition, as previously stated, WordNet is an English-based machine, which is thus incapable of identifying the subsumer of those pairs and results in a lower rating than a human one as seen in those pairs in Table 4.9. Therefore, it could be inferred that TWSS is considered not always efficient.

Accordingly, the flaws of the TWSS rating performance previously mentioned should be taken into account as a pathway to improve the TWSS measure. The new approaches to develop the improved Thai word similarity measure will be discussed in Chapter 5.

4.6 Conclusion

This chapter aimed to describe the creation of TWS-65. To begin with, the chapter covered the essence of theme words and word pairs. Methods of finding theme words primarily related to Thai culture were presented, along with an approach to formulate word pairs from the theme word set. The procedure and explanation of the word pairs were thoroughly reviewed, leading to the presentation of a methodology for creating a Thai Word Benchmark dataset. A rating procedure was adapted from known good practice in English and an experiment performed following the procedure. The captured ratings were presented and the evaluations of TWSS with TWS-65 were reported and discussed. Lastly, although displaying a promising result at this stage, the measure may clearly be improved as a predictor for human similarity perception. An analysis of some difficult cases provides the motivation for the development of a new Thai word semantic similarity measure, which is the subject of the next chapter.

Chapter 5

Word Similarity based on Lexical Chain Created from Search Engine (LCSS)

5.1 Introduction

As observed in Chapter 4, the TWSS is less effective for those word pairs that are related to Thai culture due to the problem of TWSS being derived from the English WordNet and consequently, it has limited accuracy in terms of words that are associated with the Thai culture. There is still no functioning Thai WordNet, which is an essential component in many English word semantic similarity measures (Li et al., 2003). The aim of this chapter is to create a Thai word semantic similarity measure that will overcome the weakness of TWSS, especially in its Thai cultural aspect, to create the new TWSS (nTWSS) without relying on an immature Thai WordNet. This is achieved in a novel measure by creating a lexical chain using knowledge extracted from the Web by a search engine, called 'LCSS'. This unique measure uses completely different components to calculate the similarity rating from other search engine based word measures that are reviewed in Section 2.2.1.

A "lexical chain" is defined as a sequence of related words in the text, short (adjoining words or sentences) or long distances (entire text) (Morris and Hirst, 1991). A chain is independent of the grammatical structure of the text. Consequently, it is viewed as a list of words that captures a portion of the connected structure of the text. A lexical chain can present a context for the resolution of a vague term and enable identification of the notion that the term represents (Morris and Hirst, 1991). This work is based on the conjecture that a lexical chain can substitute for WordNet in a semantic similarity measure.

In Chapter 4, TWSS was evaluated with TWS-65, showing statistical significance. Although TWSS shows reasonably good performance over the evaluation dataset TWS-65, it is clearly capable of improvement. The analysis showed in Chapter 4 that there was an impact on performance cause by a subset of word pairs that relate to Thai culture. This is explicable because TWSS functions with English WordNet.

This chapter proposes an LCSS measure which is expected to perform as well with Thai culture words as with general concepts already encountered in the English culture. The proposed algorithm aims to overcome those problems by using alternative knowledge, which will be provided by a search engine. Following a review of current well-known search engines, Google was selected for use in this research, as the Google search algorithm is a crawler-based engine designed to "crawl" the information on the internet and add it to its database, unlike other search engines which mostly use only PageRank technology and massive listing (Brin, 1998). Moreover, the Google search engine is the most widely used search engine worldwide (Seymour, 2011). There are a number of benefits of using a search engine. Firstly, the search engine can be used in a number of languages which means that LCSS can also be adapted to create knowledge in a wide range of languages and used to calculate the machine similarity rating in that language. Secondly, the data provided from the search engine are up-to–date, meaning this proposed algorithm would cover new words which enter a language over a period of time (for example, slang and fashion words).

LCSS has a trainable parameter and independent data are required for training and testing. Training set (TWS-30) and testing set (TWS-51) are created in this chapter to find the most suitable parameter for the LCSS algorithm.

The aim of this chapter is to investigate the research question: Can a search engine provide an alternative natural language resource for Thai word similarity measure?

The contributions in this chapter are:

- Creation of a word similarity measure based on a lexical chain created from a search engine
- Creation of TWS-51
- Evaluation of LCSS with TWS-51.

This chapter describes LCSS and outlines the methodology for development. Section 5.2 aims to explain LCSS and its idea, which is based on the notion that one word represents one idea or one unit in a sentence and thus, two sentences containing the same words should represent the same ideas in some aspects (Firth, 1957; Simahasan, 2002). Section 5.3 describes the datasets that were used as a training set and a testing set. Section 5.4 discusses Thai human semantic similarity ratings and LCSS ratings with the TWS-51 and the last section is the conclusion.

5.2 A Semantic Similarity Measure based on Lexical Chain Created from Search Engine (LCSS)

This section presents the LCSS algorithm in detail, from its conception, through the steps by which it estimates the semantic similarity between two words and its experimental evaluation. The LCSS algorithm works in the Thai language and all experiments were conducted using Thai words. However, the examples given in this chapter are in the English language to make them easier to understand.

According to Simahasan (2002), “**ประโยค** คือ การนำคำตั้งแต่ ๒ คำขึ้นไป มาเรียงต่อกันแล้วได้ใจความสมบูรณ์ ประโยคประกอบด้วยภาคประธาน และภาคแสดง”, meaning ‘**a sentence** is to connect more than two words together and make the complete idea. A sentence is composed of subject and predicate’ and there is a famous quotation “You shall know a word by the company it keeps” (Firth, 1957); the idea for the LCSS algorithm was inspired by those sentences. While a sentence represents one idea, a word also symbolizes one unit or one idea. Thus, it is fairly easy to suppose that sentence sharing using one or more words will be about similar ideas. It can be assumed that the two sentences might represent the same idea or could be related in certain aspects. However, it also leads to the further question with regards to the two sentences which do not contain any of the same words. How can the algorithm know whether the two sentences are related to each other or not? This question will be discussed later on in this chapter.

5.2.1 Overview of LCSS Algorithm

To find the similarity between two words, LCSS first performs a Google search using the two words as a single search term. From the results, it extracts the first eight WebPages (standard number return by Google) to form a small corpus of text relating the two words. This mini-corpus is used to construct a “Chain of words” between the two terms. To explain *Chain of words*, three sentences are given as an example of how to create a chain of words, as follows:

S1: Monk lives in the temple.

S2: Priest goes to church.

S3: Churches are Christian and temples are Buddhist.

Sharing words between sentences creates the chain of words. For example, S1 and S3 are connected by sharing the word *temple*. Likewise, S2 and S3 are connected by sharing the word *church*. Therefore, S1 and S2 are also linked via S3. An example of these three sentences connected by sharing words with each other is shown in Figure 3.1. The function words also need to be taken out, discussed in detail later in Section 5.2.4.2. After the function words are taken out, the three example sentences are now as follows:

S1: Monk temple

S2: Priest church

S3: Churches Christian temples Buddhist

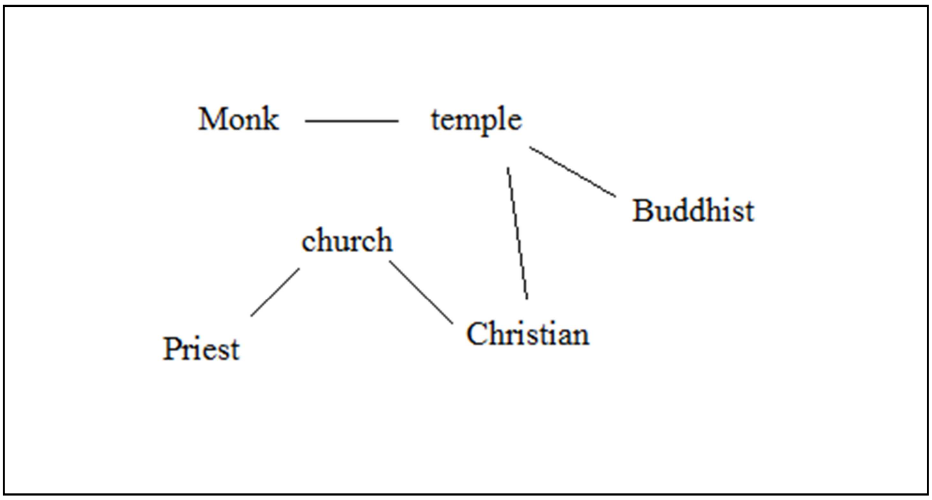


Figure 5.1: An Example of the Chain of Words

As shown in Figure 5.1, the lexical chain between the word *Monk* and the word *Priest* that can be created from this Chain of words is presented as *Monk-temple-Christian-church-priest*. It can be assumed from this lexical chain that the word *Monk* and the word *Priest* might be related to each other.

The following steps make use of a database which contains lexical chains linking pairs of words. This is based on the conjecture that the lexical database will be populated with the most frequent occurrences reducing the need for Google searches. At the start, the database is partially populated using processes, which will be described later. The lexical database can also be automatically extended during the operation of the algorithm. The extensionis saved and becomes a permanent part of a growing linguistic resource.

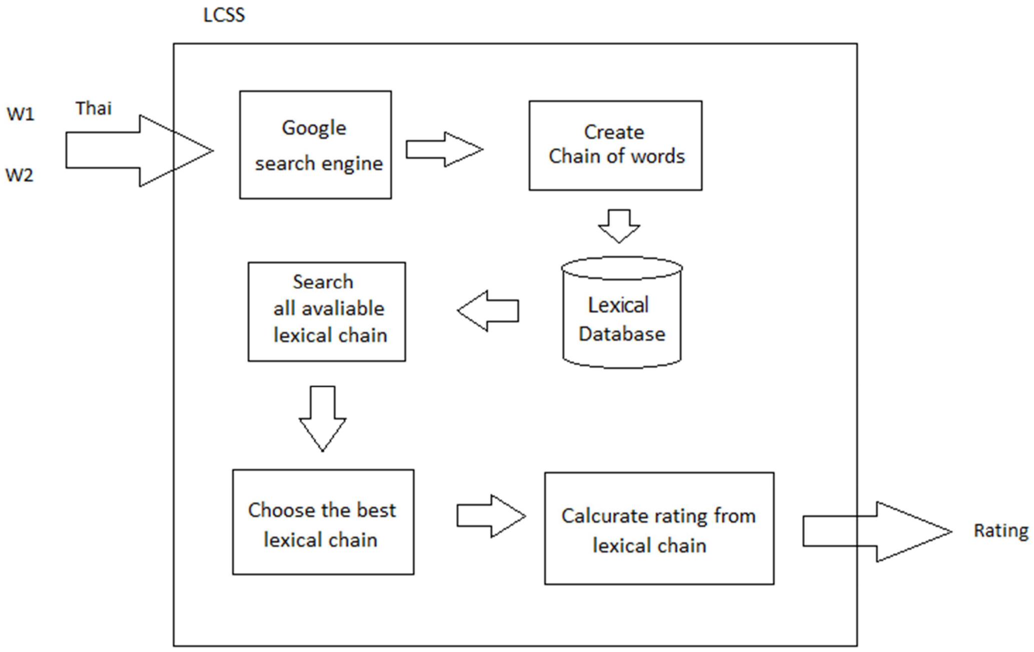

LCSS
W1
W2
Thai
Google search engine
Create Chain of words
Lexical Database
Search all avaliable lexical chain
Choose the best lexical chain
Calcurate rating from lexical chain
Rating

Figure 5.2: An Overview of LCSS Algorithm

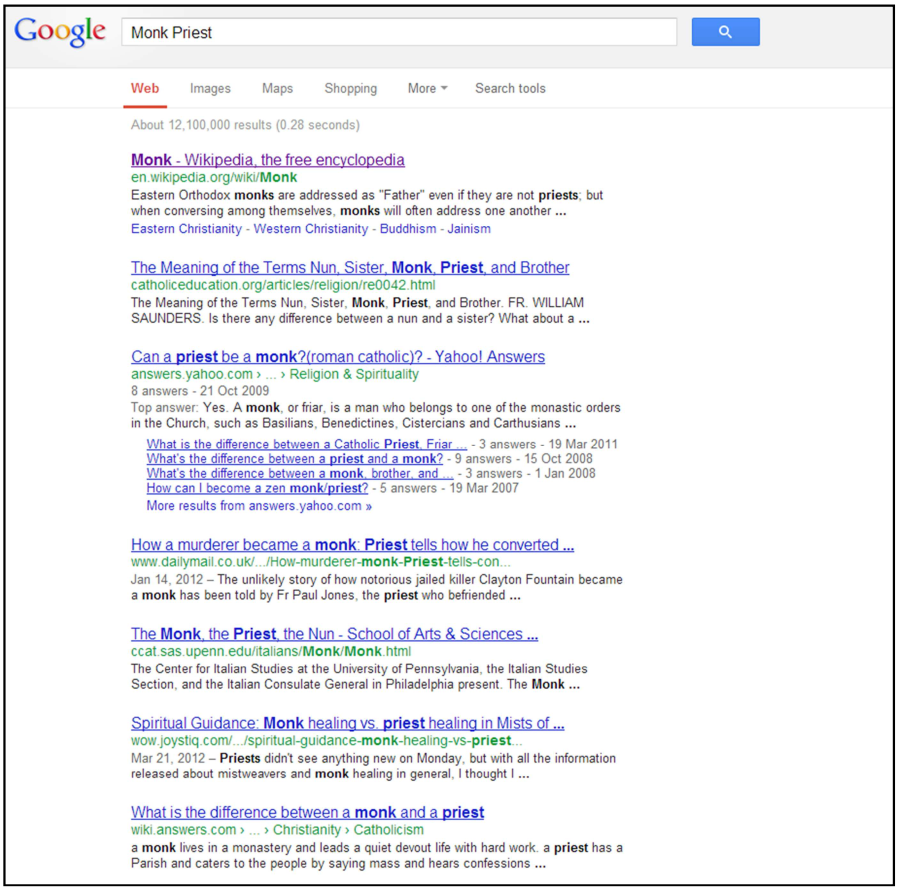
Figure 5.2 shows an overview of the LCSS algorithm, given two words: w1 and w2. The semantic similarity of the LCSS can be calculated by following these steps:

- Step 1: Send a request to the search engine (e.g. Google) with “w1 w2” as input.
- Step 2: Word Extraction from HTML Pages.
- Step 3: Insert Chain of Words into the lexical database.
- Step 4: Search the database for all available lexical chains from w1 to w2 in the database.
- Step 5: Select the best lexical chain available using Equation 5.3.
- Step 6: Calculate the similarity rating *s(w1,w2)* from Equations 5.4 and 5.5.

To simplify the concept, in this chapter, the LCSS will be explained by using the word *Monk* and the word *Priest* as the main two target words. N.B. for the purpose of understanding by non-Thai speakers, English examples have been used in some places, although the algorithm works in the Thai language.

5.2.2 STEP 1: Send a Request to Search Engine with “w1 w2” as Input

Google was selected for use in this research, as explained in Section 5.1. LCSS estimates the semantic similarity between two target words by using data received from Google. Generally, using the searching platform “w1 w2” every time that LCSS estimates the semantic similarity between words, LCSS functions by sending a request to Google by using the Google search API which is an open source code (Technofreak, 2012). The Google ajax api script returns the top eight WebPages from the search back and stores it in the database. For instance, to estimate the semantic similarity rating between the word *Monk* and the word *Priest*, they first submit a request to Google which then returns the top eight WebPage URLs, which will be used to create the chain of words in the next step. However, if the WebPage URL already exists in the database, LCSS has already captured the lexical information so that Webpage is ignored. Figure 5.3 shows an example of response from Google via input with the word *Monk* and the word *Priest*.



Monk Priest

About 12,100,000 results (0.28 seconds)

Monk - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Monk**
Eastern Orthodox **monks** are addressed as "Father" even if they are not **priests**; but when conversing among themselves, **monks** will often address one another ...
Eastern Christianity - Western Christianity - Buddhism - Jainism

The Meaning of the Terms Nun, Sister, **Monk**, **Priest**, and Brother
catholiceducation.org/articles/religion/re0042.html
The Meaning of the Terms Nun, Sister, **Monk**, **Priest**, and Brother. FR. WILLIAM SAUNDERS. Is there any difference between a nun and a sister? What about a ...

Can a **priest** be a **monk**?(roman catholic)? - Yahoo! Answers
answers.yahoo.com › ... › Religion & Spirituality
8 answers - 21 Oct 2009
Top answer: Yes. A **monk**, or friar, is a man who belongs to one of the monastic orders in the Church, such as Basilians, Benedictines, Cistercians and Carthusians ...
What is the difference between a Catholic **Priest**, Friar ... - 3 answers - 19 Mar 2011
What's the difference between a **priest** and a **monk**? - 9 answers - 15 Oct 2008
What's the difference between a **monk**, brother, and ... - 3 answers - 1 Jan 2008
How can I become a zen **monk**/**priest**? - 5 answers - 19 Mar 2007
More results from answers.yahoo.com »

How a murderer became a **monk**: **Priest** tells how he converted ...
www.dailymail.co.uk/.../How-murderer-**monk**-**Priest**-tells-con...
Jan 14, 2012 – The unlikely story of how notorious jailed killer Clayton Fountain became a **monk** has been told by Fr Paul Jones, the **priest** who befriended ...

The **Monk**, the **Priest**, the Nun - School of Arts & Sciences ...
ccat.sas.upenn.edu/italians/**Monk**/**Monk**.html
The Center for Italian Studies at the University of Pennsylvania, the Italian Studies Section, and the Italian Consulate General in Philadelphia present. The **Monk** ...

Spiritual Guidance: **Monk** healing vs. **priest** healing in Mists of ...
wow.joystiq.com/.../spiritual-guidance-**monk**-healing-vs-**priest**...
Mar 21, 2012 – **Priests** didn't see anything new on Monday, but with all the information released about mistweavers and **monk** healing in general, I thought I ...

What is the difference between a **monk** and a **priest**
wiki.answers.com › ... › Christianity › Catholicism
a **monk** lives in a monastery and leads a quiet devout life with hard work. a **priest** has a Parish and caters to the people by saying mass and hears confessions ...

Figure 5.3: An Example of Results from Google

5.2.3 STEP 2: Word Extraction from HTML Pages

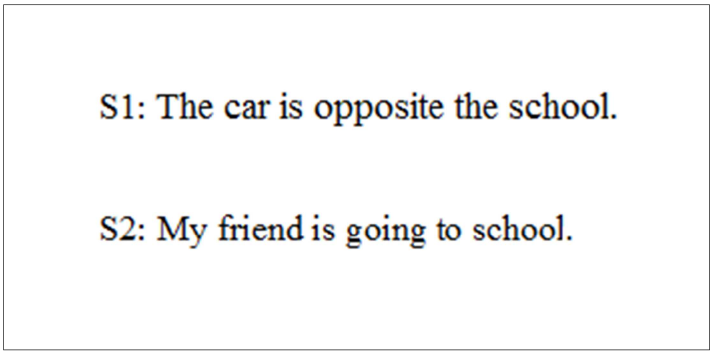
After extracting the WebPage URLs of the top eight WebPages of w1 and w2, LCSS performs three more steps to prepare the data before putting the data into the database, as follows:

- Remove the HTML tags.
- Extract words from sentence. (This step is required in a number of languages where the words in a sentence are connected, including the Thai language.)
- Remove function words.

Firstly, LCSS reads the source code from the entire eight WebPage URLs acquired from STEP1. LCSS deletes all HTML tags by using function strip_tags(); (Php.net, 2001).

Secondly, for the Thai language, there is no space between words in a sentence. A Thai word extraction algorithm (Sornlertlamvanich, 2000) is used to separate words in a sentence so that the LCSS can recognise each word in the sentence.

Lastly, Thai function words need to be taken out. Function words in themselves are, generally speaking, very high frequency. As individual words, they make little or no contribution in the semantic content of a sentence, but patterns of function words contribute structural information which define dialogue acts and so on. The two sentences are given as an example in Figure 3.4 to illustrate the difference of having and not having function words more explicitly.



S1: The car is opposite the school.

S2: My friend is going to school.

Figure 5.4: An Example of Two Sentences

Figure 5.4 shows two sentences. There is a difference in the chain of words that were created by the two sentences with and without the presence of function words, as shown in Figure 5.5.

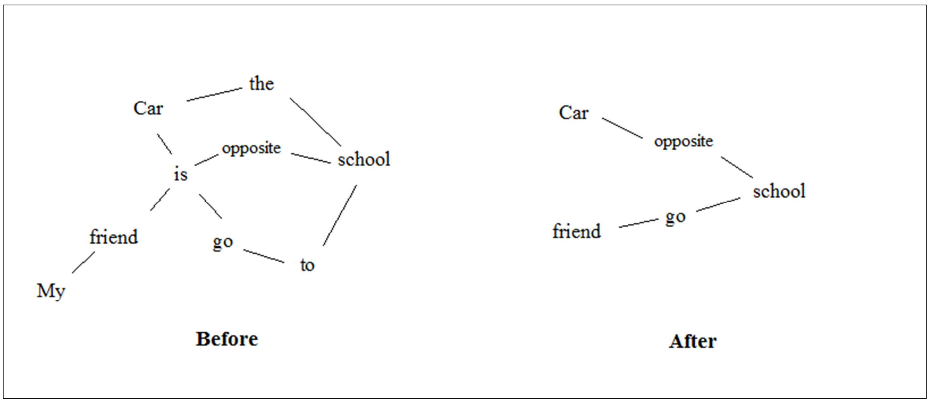


Figure 5.5: An Example of Two Sentences with and without the Presence of Function Words

From Figure 5.5, the chain of words between the word *Car* and the word *Friend* before taking function words out is *car-is-friend*, while the lexical chain on the right, eliminating those function words, yields the lexical chain *car-opposite-school-go-friend*. Therefore, including the word *is* makes the lexical chain between *Car* and *Friend* artificially short. Paradoxically, including the function words also leads to more complex and lushly-connected graphs, as there are thousands of sentences containing function words such as *is*, *to*, *and*, and so on. They make the lexical chain shorter than it should be. Therefore, eliminating function words is essential in this process to prevent complications. Furthermore, regarding the Thai language, the function words are acquired from the Thai National Corpus (Aroonmanakun, 2007).

For the main example words *Monk* and *Priest*, the lexical chain before and after removing function words is shown in Figure 5.6

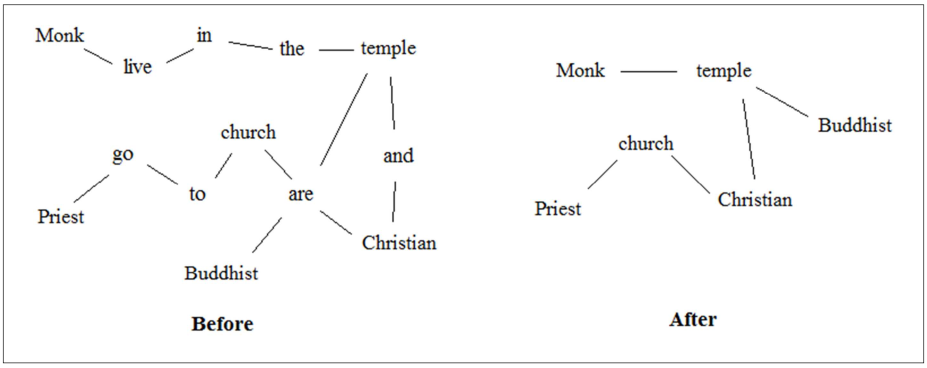


Figure 5.6: An Example of Lexical Chain before and after Remove Function Word

Before taking the function words out, the lexical chain between the word *Monk* and the word *Priest* is *Monk-live-in-the-temple-are-church-to-go-Priest*. However, after removing the function words, the lexical chain becomes *Monk-temple-Christian-church-Priest*.

5.2.4 STEP 3: Insert Lexical Chains into the Lexical Database

In this process, the links are attached to each word in every sentence as a lexical chain. Every time the word is inserted into the lexical database, it also counts the word frequency and so does the link frequency between words in every sentence. The link frequency is the frequency of two words in the sentence that are next to each other. Their link frequency is also counted when inserted into the lexical database. Word frequency and link frequency will be used to calculate the rating between two target words, which will be explained in more detail later in Section 5.2.6. This algorithm contains three main database tables, as shown in Figure 5.7.

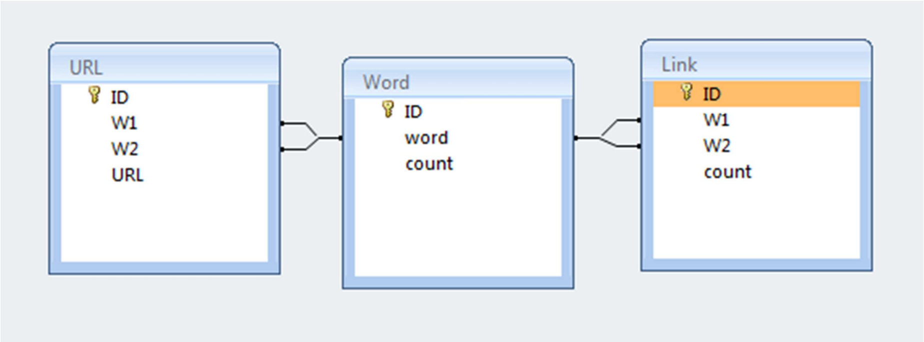


Figure 5.7: The Database

- The *Word* table records each word that is inserted into the database and the field *count* records word frequency.
- The *Link* table records each connection between two words in the lexical chain that is inserted into the database; field *count* records its link frequency.
- The *URL* table records each WebPage URL between two target words and is inserted into the database.

5.2.5 STEP 4: Searching for All Lexical Chain in the Lexical Database

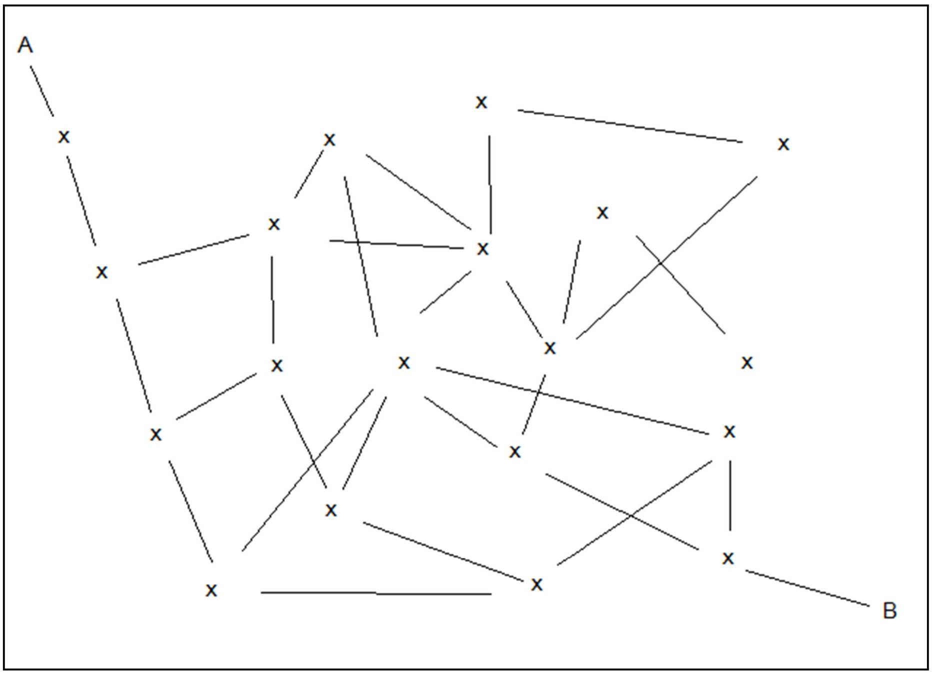


Figure 5.8: A Complex Chain of Words

More complex chains of words can be obtained from the lexical database after LCSS receives a new WebPage URL from the search engine. Figure 5.8 is a graph representing the lexical database; each *X* marks a position occupied by a word in the lexical database;

word *A* and word *B* represent the two target words. A high frequency word may be shared by hundreds of sentences making complex chains of words. There are two specific points that need to be considered. First, there are possibilities to find a lexical chain from word *A* to word *B*. Second, a boundary (the maximum range of the lexical chain) must be set on the distance from word *A* to word *B* to avoid the NP complete problem (Michael, 1979), which will be explained in Section 5.2.5.2. It is thus essential to expound the identification process.

5.2.5.1 STEP 4.1: Searching All Possible Lexical Chains between the Two Target Words

The algorithm will search and find all possible lexical chains between the two target words from the lexical database.

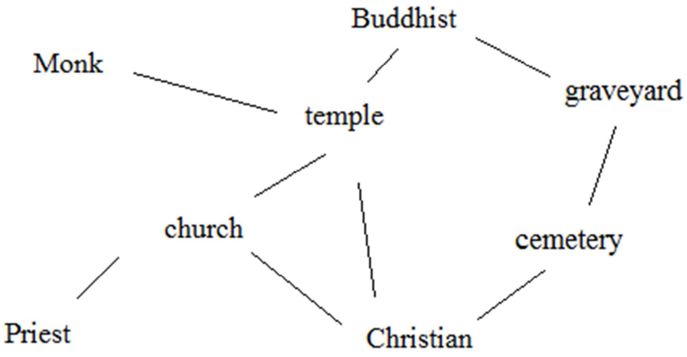


Figure 5.9: A Chain of Words between Word *Monk* and Word *Priest*

Figure 5.9 shows an example of a chain of words between the word *Monk* and the word *Priest*. There are three possible lexical chains between the word *Monk* and the word *Priest*:

- *Monk-temple-church-Priest*
- *Monk-temple-Christian-church-Priest*
- *Monk-temple-Buddhist-graveyard-cemetery-Christian-church-Priest*

These three lexical chains will be chosen and used to calculate the semantic similarity between the words *Monk* and *Priest*. In a case where LCSS cannot find any possible lexical chain (within the bound set) from the two target words, it can be assumed that the two target words are related very slightly or not related in meaning, as those two words are completely separate from each other. The algorithm will rate the word semantic similarity equal to 0 in a case where it cannot find any lexical chain from the two target words.

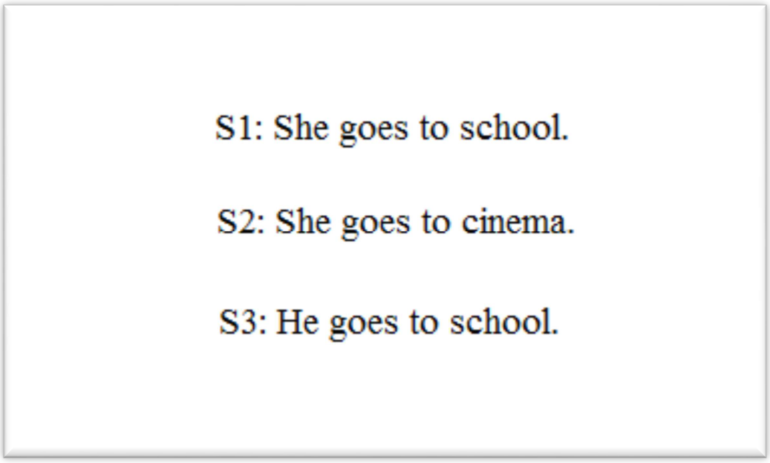
5.2.5.2 STEP 4.2: Maximum Range of Lexical Chain

To avoid the NP complete problem (Michael, 1979), the maximum range for the lexical chain is crucial. Also, if there is no maximum range (upper boundary) for the lexical chain, LCSS will obtain an infinitely large number of lexical chains for most pairs of target words as the algorithm obtains additional data most of the time the algorithm is used. Accordingly, on this basis, it is essential that a maximum range standard for the distance between the two target words is set. For the Thai language, the maximum range of the lexical chain is seven, including the two target words. This number is based on the average number of words in a sentence in the Thai language being 6.6 (Aroonmanakun, 2007)

In the example from Figure 5.9, there are three chains of words available. For the lexical chain *Monk-temple-Buddhist-graveyard-cemetery-Christian-church-Priest*, the two target words have to travel through six words to reach each. The number of words in total is eight as it includes the two target words. Hence, according to the high range number, it could be assumed that either the two words may be related to a low degree or they may not be related at all. Thus, the lexical chain *Monk-temple-Buddhist-graveyard-cemetery-Christian-church-Priest* will not be selected to calculate the rating for the word *Monk* and the word *Priest* as it exceeded the maximum range of seven.

5.2.6 STEP 5: Selecting the Best Available Lexical Chain

Once all lexical chains between the two target words have been obtained from the lexical database with the maximum range of no more than seven (in the Thai language), in the case of no lexical chain, the word semantic similarity rating of the two target words will be 0, as explained in Section 5.2.5.1. In the case where there is only one lexical chain for the two target words, that lexical chain will be automatically selected to calculate the word semantic similarity between the two target words in the next step. If there is more than one lexical chain, the extra step in LCSS is to select the best lexical chain available. To do this, the word frequency and the link frequency must be considered. Firstly, the word frequency is defined as the number of times the word occurred in past web searches and is recorded in the database. The higher the number is, the more frequency the words have. The example is given in Figure 5.10 showing three sentences.



S1: She goes to school.

S2: She goes to cinema.

S3: He goes to school.

Figure 5.10: An Example of Three Sentences

In Figure 5.10, *S1*, *S2* and *S3* are analysed together, the frequency of the words are two for the word *she*, three for *goes*, two for *school*, and one for *cinema*. It can therefore be concluded that the word *goes* has the highest frequency among the other words in this case. Secondly, the link between words must be identified prior to counting their frequency. A link is defined as the presence of two or more words which appear together in the lexical chain. According to Figure 5.10, there are *she-goes*, *he-goes*, *goes-school*, and *goes-cinema*, which are called the link between words. Consequently, the link frequency is how frequently they appear in sentences. When all *S1*, *S2* and *S3* are analysed together, it shows that the link *she-goes* counts two as the frequency, while *he-goes* is one. Similarly, while *goes-school* has a frequency of two, *goes-cinema* has one. Hence, in this case, the link *she-goes* and *goes–school* have a higher frequency than the other two links, *he-goes* and *goes–cinema*.

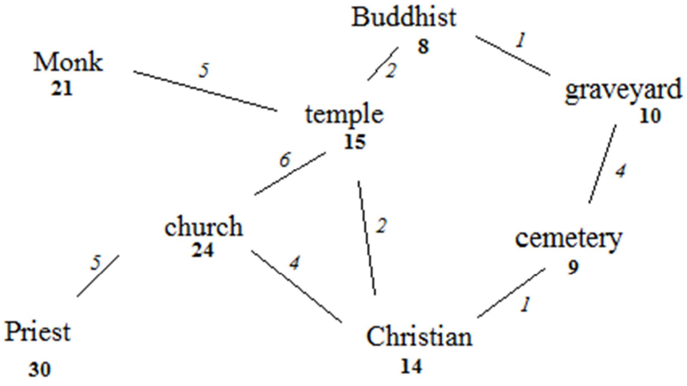


Figure 5.11: A Chain of Word between Word *Monk* and *Priest* with Word Frequency and Link Frequency

Accordingly, the frequency of the links between words needs to be pinpointed and established so that the lexical chain can be identified.

Figure 5.11 is the same as Figure 5.9 with the additional information. Figure 5.11 shows a chain of words between the words *Monk* and *Priest* with their word frequency and their link frequency. The bold numbers under the words represent the word frequency. The italic numbers appearing between words represent the link frequency, which is the count of the number of times the words were found to be adjacent during the graph construction between words.

After applying the upper boundary length, there were two lexical chains left for the words *Monk* and *Priest*. These are *Monk-temple-church-Priest* and *Monk-temple-Christian-church-Priest*, respectively. To determine which lexical chain to be selected in LCSS, it needs to calculate to find the Link Density of each lexical chain. Link Density can be used to determine which lexical chain is most representative of the word pair behaviour in the language and therefore should give the most meaningful measure of similarity.

The link density of the lexical chain is denoted as *LD* and it can be calculated from Equation 5.1:

$$Link\ Density\ (LD) = \frac{\sum LF}{\sum WF}$$ **Equation 5.1**

where

n is the total number of words in the lexical chain.

$\sum LF$ is the abbreviation for $\sum_{i=1}^{n-1} LF_{i \to i+1}$. It is a sum value of link frequency in a lexical chain. In other words, it is a sum value of each link frequency ($LF_{i \to i+1}$) between a pair of words *i*th and (*i*+*1*)th in the lexical chain where the *i* value can be [1, *n*-1].

$\sum WF$ is the abbreviation for $\sum_{i=1}^{n} WF_i$. It is a sum value of word frequency (WF_i) from word *i*th to *n*th in a lexical chain.

A number of experiments were conducted to find to the most suitable formula to use in a particular situation.

To calculate:

The link density of the lexical chain *Monk-temple-church-Priest* is:

$$= \quad \frac{LF_{Monk\to temple}+LF_{temple\to church}+LF_{church\to Priest}}{WF_{Monk}+WF_{temple}+WF_{church}+WF_{Priest}}$$

= (5+6+5) / (21+15+24+30)

= 0.178

Similarly, the link density of the lexical chain *Monk-temple-Christian-church-Priest* is:

= (5+2+4+5) / (21+15+14+24+30)

= 0.153

According to the example, it shows that the lexical chain *Monk-temple-church-Priest* is better than the lexical chain *Monk-temple-Christian-church-Priest* since it has a higher value of *LD*. As a result, the route *Monk-temple-church-Priest* will be selected to calculate the word semantic similarity rating in the next step.

5.2.7 STEP 6: Calculate the Semantic Similarity Rating between Two Target Words

This work adapts an equation for calculating similarity from the path length given in Li et al. (2003). This can be generalized to Equation 5.2, where *Alpha* is a constant and *Beta* can be calculated from Equation 5.3.

The semantic similarity from the lexical chain between two target words can be calculated by using two formulas in Equation 5.2 and Equation 5.3.

Two words are given, *w1* and *w2*; the semantic similarity $s(w1, w2)$ can be calculated from Equation 5.2:

$$s(w1, w2) \;=\; tanh(Alpha \times Beta) \qquad \textbf{Equation 5.2}$$

The *Alpha* parameter is a constant value which is decided depending on the language. The *Alpha* parameter of the Thai language will be discussed in the experiment section, Section 5.4. The value of *Beta* can be calculated from Equation 5.3:

$$Beta \;= \frac{\sum LF}{n} \times \left(\frac{\sum LF + \sum WF}{\sum WF}\right) \qquad \textbf{Equation 5.3}$$

Equation 5.2 is a well-known method used in STASIS (Li et al., 2003). A number of experiments were conducted to determine the most suitable equation for $Beta$.

From the example in Figure 5.11, as explained in Section 5.2.6, the lexical chain *Monk-temple-church-Priest* is selected as it returns a higher value of link density. Subsequently, LCSS calculates the rating by using Equation 5.2 and Equation 5.3 as:

$$\sum LF \quad = \quad 5+6+5 \quad = \quad 16$$

$$\sum WF \quad = \quad 21+15+24+30 \quad = \quad 90$$

$$n \quad = \quad 4$$

$$Beta \quad = \quad \frac{16}{4} \times \left(\frac{16+90}{90}\right) \quad = \quad 4.711$$

Then *s(w1, w2)* can be calculated from Equation 5.2:

Given that the *Alpha* value is 0.2:

$$s(w1, w2) \quad = \quad \tanh(0.2 \times 4.711)$$

$$= \quad 0.736$$

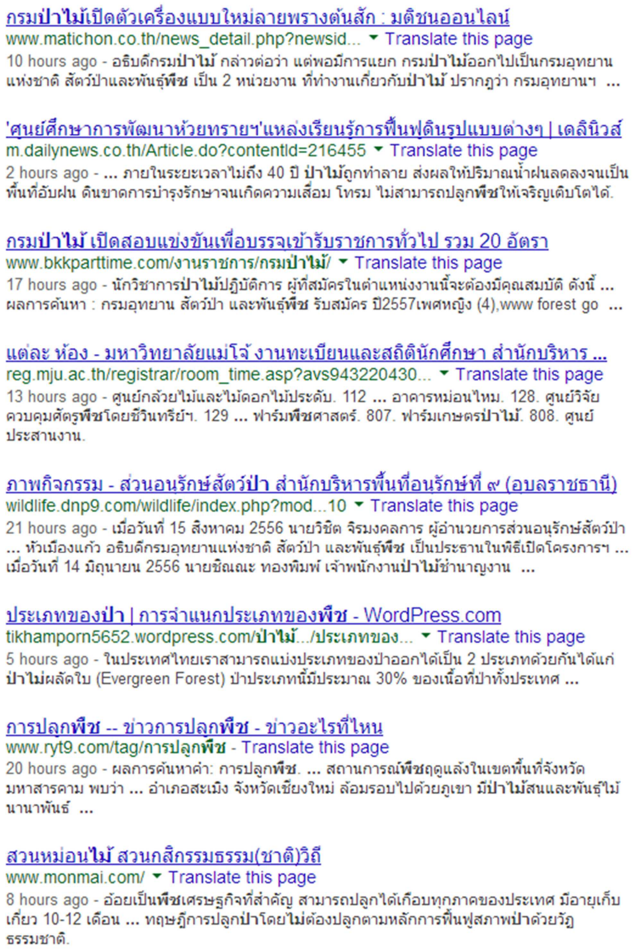
From the calculation, the semantic similarity rating of LCSS between the word *Monk* and the word *Priest* is 0.736.

5.2.8 An Example of LCSS usage with Thai Word Pairs

The aim of this section is to present an actual example of LCSS. Given two words, w_1 and w_2 as *พืช (Plant)* and *ป่าไม้ (Forest)*, the semantic similarity rating can be calculated in 6 steps.

STEP 1: Send a Request to Search Engine with "W1 (Space Bar) W2" as Input

The first step to calculate the rating between he words *พืช (Plant)* and *ป่าไม้ (Forest)* is to send a request to the search engine (Google). Figure 5.12 shows the top eight WebPages of searched words *พืช (Plant)* and *ป่าไม้ (Forest)*.



กรมป่าไม้เปิดตัวเครื่องแบบใหม่ลายพรางต้นสัก : มติชนออนไลน์
www.matichon.co.th/news_detail.php?newsid... ▾ Translate this page
10 hours ago - อธิบดีกรมป่าไม้ กล่าวต่อว่า แต่พอมีการแยก กรมป่าไม้ออกไปเป็นกรมอุทยานแห่งชาติ สัตว์ป่าและพันธุ์พืช เป็น 2 หน่วยงาน ที่ทำงานเกี่ยวกับป่าไม้ ปรากฎว่า กรมอุทยานฯ ...

'ศูนย์ศึกษาการพัฒนาห้วยทรายฯ'แหล่งเรียนรู้การฟื้นฟูดินรูปแบบต่างๆ | เดลินิวส์
m.dailynews.co.th/Article.do?contentId=216455 ▾ Translate this page
2 hours ago - ... ภายในระยะเวลาไม่ถึง 40 ปี ป่าไม้ถูกทำลาย ส่งผลให้ปริมาณน้ำฝนลดลงจนเป็นพื้นที่อับฝน ดินขาดการบำรุงรักษาจนเกิดความเสื่อม โทรม ไม่สามารถปลูกพืชให้เจริญเติบโตได้.

กรมป่าไม้ เปิดสอบแข่งขันเพื่อบรรจุเข้ารับราชการทั่วไป รวม 20 อัตรา
www.bkkparttime.com/งานราชการ/กรมป่าไม้/ ▾ Translate this page
17 hours ago - นักวิชาการป่าไม้ปฏิบัติการ ผู้ที่สมัครในตำแหน่งงานนี้จะต้องมีคุณสมบัติ ดังนี้ ... ผลการค้นหา : กรมอุทยาน สัตว์ป่า และพันธุ์พืช รับสมัคร ปี2557เพศหญิง (4),www forest go ...

แต่ละ ห้อง - มหาวิทยาลัยแม่โจ้ งานทะเบียนและสถิตินักศึกษา สำนักบริหาร ...
reg.mju.ac.th/registrar/room_time.asp?avs943220430... ▾ Translate this page
13 hours ago - ศูนย์กล้วยไม้และไม้ดอกไม้ประดับ. 112 ... อาคารหม่อนไหม. 128. ศูนย์วิจัยควบคุมศัตรูพืชโดยชีวินทรีย์ฯ. 129 ... ฟาร์มพืชศาสตร์. 807. ฟาร์มเกษตรป่าไม้. 808. ศูนย์ประสานงาน.

ภาพกิจกรรม - ส่วนอนุรักษ์สัตว์ป่า สำนักบริหารพื้นที่อนุรักษ์ที่ ๙ (อุบลราชธานี)
wildlife.dnp9.com/wildlife/index.php?mod...10 ▾ Translate this page
21 hours ago - เมื่อวันที่ 15 สิงหาคม 2556 นายวิชิต จิรมงคลการ ผู้อำนวยการส่วนอนุรักษ์สัตว์ป่า ... หัวเมืองแก้ว อธิบดีกรมอุทยานแห่งชาติ สัตว์ป่า และพันธุ์พืช เป็นประธานในพิธีเปิดโครงการฯ ... เมื่อวันที่ 14 มิถุนายน 2556 นายชิณณะ ทองพิมพ์ เจ้าพนักงานป่าไม้ชำนาญงาน ...

ประเภทของป่า | การจำแนกประเภทของพืช - WordPress.com
tikhamporn5652.wordpress.com/ป่าไม้.../ประเภทของ... ▾ Translate this page
5 hours ago - ในประเทศไทยเราสามารถแบ่งประเภทของป่าออกได้เป็น 2 ประเภทด้วยกันได้แก่ ป่าไม่ผลัดใบ (Evergreen Forest) ป่าประเภทนี้มีประมาณ 30% ของเนื้อที่ป่าทั้งประเทศ ...

การปลูกพืช -- ข่าวการปลูกพืช - ข่าวอะไรที่ไหน
www.ryt9.com/tag/การปลูกพืช - Translate this page
20 hours ago - ผลการค้นหาคำ: การปลูกพืช. ... สถานการณ์พืชฤดูแล้งในเขตพื้นที่จังหวัดมหาสารคาม พบว่า ... อำเภอสะเมิง จังหวัดเชียงใหม่ ล้อมรอบไปด้วยภูเขา มีป่าไม้สนและพันธุ์ไม้นานาพันธ์ ...

สวนหม่อนไม้ สวนกสิกรรมธรรม(ชาติ)วิถี
www.monmai.com/ ▾ Translate this page
8 hours ago - อ้อยเป็นพืชเศรษฐกิจที่สำคัญ สามารถปลูกได้เกือบทุกภาคของประเทศ มีอายุเก็บเกี่ยว 10-12 เดือน ... ทฤษฎีการปลูกป่าโดยไม่ต้องปลูกตามหลักการฟื้นฟูสภาพป่าด้วยวัฏธรรมชาติ.

Figure 5.12: An Example of Result from Google for words *พืช (Plant)* and *ป่าไม้ (Forest)*

STEP 2: Word Extraction from HTML Pages

For each sentence in the top eight WebPages from STEP 1, the Thai word extraction algorithm (Sornlertlamvanich, 2000) is used to separate the words in a sentence. For instance, in the paragraph:

“ประเทศไทยดินแดนแห่งความมหัศจรรย์ อุดมสมบูรณ์ไปด้วยพืชพันธุ์นานาชนิด มีความหลากหลายของพันธุกรรมและทรัพยากรธรรมชาติมากมายสมควรที่คนไทยควรตระหนัก และร่วมมือร่วมใจกันอนุรักษ์พันธุกรรมและทรัพยากรไทยให้คงอยู่ คู่กับคนไทยตลอดไป”

After each word in a sentence is separated and the function words are removed, the lexical chain for each sentence is created as follows:

- ประเทศไทย – ดินแดน – แห่ง - ความมหัศจรรย์
- อุดมสมบูรณ์ – ไป - พืช
- ความหลากหลาย – พันธุกรรม – ทรัพยากร – ธรรมชาติ – คนไทย - ตระหนัก
- ร่วมมือร่วมใจ – อนุรักษ์ – พันธุกรรม – ทรัพยากร - ไทย

- คู่ – คนไทย – ตลอด – ไป

STEP 3: Insert Lexical Chains into the Lexical Database

Each lexical chain that has been created in STEP 2 is inserted into the lexical database, in which the word frequency and the link frequency are also counted. However, starting with an empty lexical database is unnecessarily painstaking and so the lexical database is pre-populating by collecting data from Google following STEP 1-2 for all word pairs in TWS-65.

STEP 4: Searching for All Lexical Chains in the Lexical Database

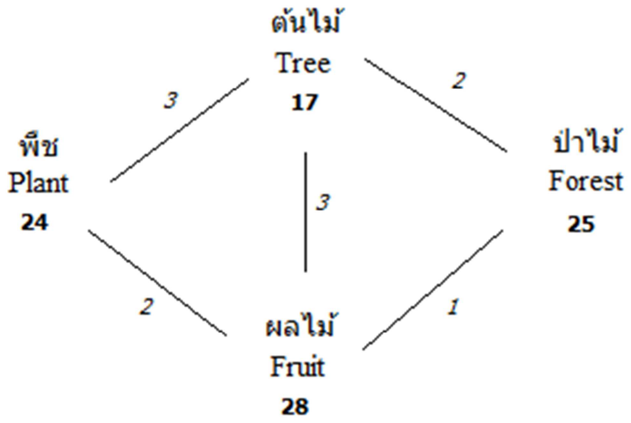


Figure 5.13: A Chain of Words between Word พืช (Plant) and ป่าไม้ (Woods) with Word Frequency and Link Frequency

The chain of words between words *พืช (Plant)* and *ป่าไม้ (Forest)* from the lexical database are shown in Figure 5.13. The bold numbers under the words represent the word frequency. The italicised numbers appearing between the words represent the link frequency

The possible lexical chains from the chain of words between the words *พืช (Plant)* to *ป่าไม้ (Forest)* in Figure 5.13 are as follows:

- *พืช-ต้นไม้-ป่าไม้ (Plant-Tree-Forest)*
- *พืช-ผลไม้-ป่าไม้ (Plant-Fruit-Forest)*
- *พืช-ต้นไม้-ผลไม้-ป่าไม้ (Plant-Tree-Fruit-Forest)*
- *พืช-ผลไม้-ต้นไม้-ป่าไม้ (Plant-Fruit-Tree-Forest)*

STEP 5: Selecting the Best Available Lexical Chain

Calculate the Link Density from each lexical chain from STEP 4 by using the Equation 5.2.

The link density of the lexical chain *พืช-ต้นไม้-ป่าไม้(Plant-Tree-Forest)* is:

$$= \quad \frac{LF_{Plant\rightarrow Tree}+LF_{Tree\rightarrow Forest}}{WF_{Plant}+WF_{Tree}+WF_{Forest}}$$

= (3+2) / (24+17+25)

= 0.075

The link density of the lexical chain *พืช-ผลไม้-ป่าไม้(Plant-Fruit-Forest)* is:

$$= \quad \frac{LF_{Plant\rightarrow Fruit}+LF_{Ftuit\rightarrow Forest}}{WF_{Plant}+WF_{Fruit}+WF_{Forest}}$$

= (2+1) / (24+28+25)

= 0.038

The link density of the lexical chain *พืช-ต้นไม้-ผลไม้--ป่าไม้(Plant-Tree-Fruit-Forest)* is:

$$= \quad \frac{LF_{Plant\rightarrow Tree}+LF_{Tree\rightarrow Fruit}+LF_{Ftuit\rightarrow Forest}}{WF_{Plant}+WF_{Tree}+WF_{Fruit}+WF_{Forest}}$$

= (3+3+1) / (24+17+28+25)

= 0.074

The link density of the lexical chain *พืช-ผลไม้-ต้นไม้-ป่าไม้(Plant-Fruit-Tree-Forest)* is:

$$= \quad \frac{LF_{Plant\rightarrow Fruit}+LF_{Fruit\rightarrow Tree}+LF_{Tree\rightarrow Forest}}{WF_{Plant}+WF_{Fruit}+WF_{Tree}+WF_{Forest}}$$

= (2+3+2) / (24+28+17+25)

= 0.074

The lexical chain *พืช-ต้นไม้-ป่าไม้ (Plant-Tree-Forest)* is chosen to calculate the word semantic similarity rating between the words *พืช (Plant)* to *ป่าไม้ (Forest)* since it has the highest value of *LD*.

STEP 6: Calculate the Semantic Similarity Rating between Two Target Words

Calculate the semantic similarity rating between the words *พืช (Plant)* to *ป่าไม้ (Forest)* from the lexical chain *พืช-ต้นไม้-ป่าไม้ (Plant-Tree-Forest)* from STEP 5 by using the Equation 5.2 and Equation 5.3 as:

$\sum LF$ = 3+2 = 5

$\sum WF$ = 24+17+25 = 66

n = 3

Beta = $\frac{5}{3} \times (\frac{5 + 66}{66})$ = 1.075

Then *s(w1, w2)* can be calculated from:

Given that the *Alpha* value is 0.2:

s(w1, w2) = *tanh (0.2 × 1.075)*

= *0.212*

From the calculation, the semantic similarity rating of LCSS between the words *พืช (Plant)* to *ป่าไม้ (Forest)* is 0.212.

5.3 Training Dataset and Testing Dataset to Evaluate LCSS

Prior to the present work, there has been no attempt to create a benchmark dataset for Thai word similarity measures. Prior work in English has established the high cost of creating a gold standard similarity dataset which is of sufficient size for training Machine Learning classifiers, although evidence has been found that a small carefully designed dataset can provide an acceptable evaluation (O’Shea at el., 2013).

To provide a meaningful evaluation, LCSS must be tested with a benchmark dataset. However, LCSS also has a trainable parameter and independent data are required for training and testing and an acceptable value can be established for this parameter with a relatively small dataset.

The aim of this section is to describe the methodology for creating training and testing sets. Also, the training set will be used to experiment with the *Alpha* parameter in Equation 5.2 in Section 5.4 and the testing set used to evaluate LCSS in Section 5.5.

From Chapter 4 Section 4.4.5, an experiment was conducted to find whether TWS-30 and the TWS-65 are statistically significantly different. No evidence was found; thus, TWS-30 and TWS-65 are considered to represent the same population.

5.3.1 Training Dataset

TWS-30 from Chapter 3 will be used as a training set to find the most suitable value of *Alpha* parameter, which will be explained in Section 5.4. After the value of the *Alpha* parameter is assigned, the testing data set will be used to evaluate LCSS.

5.3.2 Testing Dataset

The TWS-65 is used as a testing set to evaluate the LCSS measure. As mentioned in Section 4.4.5, there are 14 word pairs in the table which contain the same meaning in both TWS-30 and TWS-65. Hence, those word pairs will not be used as a testing set to avoid bias when evaluating with LCSS. Therefore, there are 51 word pairs from TWS-65 that will be adopted as the testing set. Those 51 word pairs will be called TWS-51 and can be found in Table 5.1. Column *WP* is the number of the word pairs. Columns W_1 and W_2 are the word pairs.

Table 5.1: TWS-51

WP	W_1		W_2	
1	แก้ว	Glass	ข้ารับใช้	Serf
2	อาหาร	Food	ลายเซ็น	Signature
3	อัญมณี	Gem	ลายเซ็น	Signature
4	ฝั่งทะเล	Coast	รถยนต์	Car
5	สุนัข	Dog	เครื่องมือ	Tool
6	การเดินทาง	Journey	สุสาน	Graveyard
7	เที่ยงวัน	Midday	โรงละคร	Theatre
8	เที่ยงวัน	Midday	การท่องเที่ยว	Voyage
9	ยานพาหนะ	Automobile	เพชรพลอย	Jewel
10	เนินเขา	Hill	ผลไม้	Fruit
11	นักมายากล	Magician	ถ้วย	Cup
12	ป่าช้า	Cemetery	หมา	Dog
13	นักมายากล	Magician	เครื่องมือ	Tool
14	การเดินทาง	Journey	กลางวัน	Noon
15	นิตยสาร	Magazine	ป้า	Aunt
16	นักบวช	Priest	หนังสือ	Book

17	เด็กผู้ชาย	Boy	หมา	Dog
19	สุนัข	Dog	เด็กหนุ่ม	Lad
20	วัด	Temple	พงไพร	Woods
21	ทาส	Slave	หมา	Dog
22	อาหาร	Food	ถ้วย	Cup
23	ครู	Teacher	หนังสือ	Book
24	พืช	Plant	ผ้าไหม	Silk
25	เด็กผู้ชาย	Boy	อาจารย์	Lecturer
26	โรงภาพยนต์	Cinema	โบสถ์	Church
27	ทาส	Slave	เด็กหนุ่ม	Lad
28	เนินเขา	Hill	ชายฝั่ง	Shore
29	ยานพาหนะ	Automobile	เครื่องมือ	Tool
30	ผ้าฝ้าย	Cotton	ต้นไม้	Tree
31	อุปกรณ์	Implement	รถยนต์	Car
32	ลุง	Uncle	อาจารย์	Lecturer
33	ป่าไม้	Forest	ผลไม้	Fruit
34	ครู	Teacher	ป้า	Aunt
35	นักบวช	Priest	พ่อมด	Wizard
36	แก้ว	Glass	เพชรพลอย	Jewel
38	วัด	Temple	สุสาน	Graveyard
39	พืช	Plant	พงไพร	Woods
41	อาหาร	Food	ผลไม้	Fruit
42	แก้ว	Glass	ถ้วย	Cup
43	วัด	Temple	พระ	Monk
44	ลุง	Uncle	ป้า	Aunt
45	ป่าไม้	Forest	ต้นไม้	Tree
46	โรงภาพยนต์	Cinema	โรงละคร	Theatre
47	ผ้าฝ้าย	Cotton	ผ้าไหม	Silk
49	พืช	Plant	ต้นไม้	Tree
50	นิตยสาร	Magazine	หนังสือ	Book
57	เนินเขา	Hill	ภูเขา	Mountain
60	นักบวช	Priest	พระ	Monk
62	วัด	Temple	โบสถ์	Church
63	ครู	Teacher	อาจารย์	Lecturer
65	สุนัข	Dog	หมา	Dog

The methods for collecting and rating these pairs are described in Section 4.5.

5.4 Experiment to Find the Most Suitable *Alpha* Parameter in LCSS in Thai Language

The *Alpha* parameter is the constant in Equation 5.2. It needs to be assigned before calculating the LCSS rating. As stated in Section 5.2, LCSS can be used in other languages and a suitable (optimal) value should be established for each language. For the Thai language, TWS-30 will be used as a training set to find the most suitable value of the *Alpha* parameter.

5.4.1 Methodology

The following experiment was conducted to find the most suitable *Alpha* parameter in LCSS to use in the Thai language.

- The lexical database was pre-populated by collecting data from Google following STEP1-2 for all word pairs in TWS-30.
- The LCSS rating for each word pair in the training set was calculated by applying a value of the *Alpha* parameter from 0-1, plus 0.05 each step.
- The correlation coefficients (*r*) between Human rating and LCSS were calculated for each value of the *Alpha* parameter
- The optimal value of *Alpha* was chosen using the highest correlation coefficient (*r*) that was obtained.

The Pearson Product-Moment correlation coefficients (*r*) between the Thai Human rating and LCSS rating with the value of the *Alpha* parameter between 1-0 will be calculated and shown in Section 5.4.3

5.4.2 LCSS Rating with a Different *Alpha* Parameter

Column *WP* shows the word pairs in TWS-30. Column *Human* shows the Average human rating for each word pair. The rest of the columns show similarity ratings of LCSS with the value of the *Alpha* parameter between 1-0. The values of the *Alpha* parameter 0.45, 0.30, 0.20, 0.15, 0.10, and 0.05 are shown in Table 5.2 (these 6 values are selected for presentation as they are the most promising values from the set tested).

Table 5.2: The Average of Similarity Rating for the 14 Word Pairs in Both Datasets

WP	***Human***	**The value of *Alpha***					
		0.45	**0.30**	**0.20**	**0.15**	**0.10**	**0.05**
1	0.020	0.000	0.000	0.000	0.000	0.000	0.000
5	0.006	0.000	0.000	0.000	0.000	0.000	0.000
9	0.017	0.000	0.000	0.000	0.000	0.000	0.000
13	0.171	0.487	0.340	0.232	0.175	0.118	0.059
17	0.158	0.563	0.401	0.276	0.209	0.141	0.071
21	0.150	0.498	0.349	0.238	0.180	0.121	0.061
25	0.137	0.493	0.345	0.235	0.178	0.119	0.060
29	0.149	0.648	0.473	0.330	0.252	0.170	0.086
33	0.541	0.479	0.334	0.228	0.172	0.115	0.058
37	0.315	0.467	0.325	0.221	0.167	0.112	0.056

41	0.325	0.750	0.570	0.407	0.313	0.213	0.108
47	0.403	0.754	0.575	0.411	0.316	0.215	0.109
48	0.393	0.248	0.167	0.112	0.084	0.056	0.028
49	0.605	0.493	0.345	0.235	0.178	0.119	0.060
50	0.221	0.489	0.342	0.233	0.176	0.118	0.059
51	0.781	0.493	0.345	0.235	0.178	0.119	0.060
52	0.583	0.645	0.471	0.328	0.250	0.169	0.085
53	0.836	0.981	0.913	0.773	0.648	0.473	0.252
54	0.697	0.445	0.308	0.209	0.158	0.106	0.053
55	0.806	0.809	0.635	0.462	0.358	0.245	0.124
56	0.805	0.487	0.340	0.232	0.175	0.118	0.059
57	0.708	0.882	0.727	0.548	0.431	0.298	0.153
58	0.834	0.907	0.765	0.586	0.465	0.324	0.166
59	0.379	0.510	0.358	0.245	0.185	0.124	0.062
60	0.606	0.692	0.513	0.361	0.276	0.187	0.094
61	0.509	0.923	0.791	0.615	0.491	0.344	0.177
62	0.850	0.901	0.755	0.576	0.456	0.317	0.163
63	0.770	0.787	0.610	0.440	0.340	0.232	0.118
64	0.752	0.960	0.860	0.698	0.570	0.407	0.212
65	0.769	0.907	0.765	0.586	0.465	0.324	0.166

5.4.3 Results

The Pearson Product-Moment correlations coefficients for each value of *Alpha* assigned in LCSS are shown in Table 5.3. The best correlation coefficients (*r*) between the Human rating and LCSS similarity rating are obtained from using value of *Alpha* = 0.20 which are:

- Pearson’s *r* = 0.703 (P-Value < 0.01)

Table 5.3: The Pearson Product-Moment Correlations in Each Value of *Alpha*

LCSS	Correlation	P-Value
Alpha=0.45	0.696	< 0.01
Alpha=0.30	0.702	< 0.01
Alpha=0.20	0.703	< 0.01
Alpha=0.15	0.695	< 0.01
Alpha=0.10	0.688	< 0.01
Alpha=0.05	0.682	< 0.01

As a result, it can be assumed that the most suitable value of the *Alpha* parameter is 0.20 for the Thai language.

5.5 Evaluation of LCSS with the Testing Dataset (TWS-51)

The aim of this section is to describe the series of experiments conducted using the testing dataset to evaluate the LCSS.

5.5.1 Methodology

The TWS-51 from Section 5.3.3 was used to evaluate LCSS by comparing the Pearson Product-Moment correlation coefficient (*r*) between human ratings and LCSS ratings over the dataset. The methodology is as follows:

- The lexical database was pre-populated by collecting data from Google following STEP1-2 for all word pairs in TWS51.
- LCSS rating was obtained using calculations as described in Section 5.2, with the *Alpha* value of 0.2, established in Section 5.4.3.

The Pearson Product-Moment correlation coefficient (*r*) between Thai human rating and LCSS measure will be calculated and shown in Section 5.5.3.

5.5.2 LCSS Semantic Similarity Rating

Table 5.4 shows the semantic similarity ratings for the Thai word pairs. Column *WP* is the number of the word pairs from the TWS51. Column *Human* presents the Human rating for the Thai word pairs. Column *LCSS* displays the LCSS rating for the Thai word pairs described in Section 5.2. All of the measures have been scaled in the range 0 to 1 to aid comparison.

Table 5.4: Semantic Similarity Rating for Human Participants and LCSS

WP	***Human***	**LCSS**	***WP***	***Human***	**LCSS**
1	0.014	0.218	**28**	0.294	0.319
2	0.017	0.000	**29**	0.316	0.567
3	0.024	0.000	**30**	0.321	0.223
4	0.028	0.000	**31**	0.334	0.678
5	0.031	0.213	**32**	0.353	0.202
6	0.034	0.207	**33**	0.388	0.201
7	0.044	0.000	**34**	0.406	0.561
8	0.056	0.202	**35**	0.430	0.204
9	0.061	0.104	**36**	0.481	0.202
10	0.069	0.207	**38**	0.538	0.564
11	0.069	0.211	**39**	0.553	0.364
12	0.073	0.000	**40**	0.564	0.568
14	0.080	0.318	**41**	0.591	0.558
15	0.086	0.333	**42**	0.603	0.698
16	0.104	0.397	**43**	0.669	0.697

17	0.105	0.203	**44**	0.686	0.394
18	0.110	0.360	**45**	0.726	0.396
19	0.133	0.494	**46**	0.754	0.521
20	0.135	0.209	**49**	0.763	0.202
21	0.139	0.399	**55**	0.799	0.401
22	0.163	0.201	**60**	0.853	0.568
23	0.246	0.566	**61**	0.894	0.802
24	0.261	0.221	**63**	0.923	0.701
25	0.271	0.221	**64**	0.946	0.876
26	0.274	0.000	**65**	0.981	0.975
27	0.290	0.202			

5.5.3 Discussion

The experimental results in Section 5.5.2 suggest that the LCSS provides good results on TWS-51. The Pearson Product-Moment correlation coefficient (*r*) between the Thai human rating and the LCSS measure is:

- Pearson's *r* = 0.723 (P-Value < 0.01)

Table 5.5 illustrates the agreement of both the machine measures with human ratings, and the machine ratings over TWS-51 by calculating the correlation coefficients (*r*) between the human ratings and the machine ratings over the 51 word pairs. Also, the leave-one-out resampling technique is used to find the correlation coefficient of each participant with the rest of the group.

Table 5.5: The Pearson Product-Moment Correlation Coefficients (*r*)

	Correlation *r*	**P-value**
Thai human similarity rating and TWSS	0.752	< 0.01
Thai human similarity rating and LCSS	0.723	< 0.01
Average Thai native speaker and the rest of the group	0.865	-
Worst Thai native speaker participant and the rest of the group	0.708	-
Best Thai native speaker participant and the rest of the group	0.928	-

In Table 5.5, the LCSS measure performs better than the correlation between the worst performing human and the rest of the group (*r* = 0.708), which supports the notion that it could build up the basis of an effective algorithm.

The paired sample t-test was used to find whether or not the Human ratings and LCSS ratings over the dataset were significantly different (α=0.05) from the hypotheses:

- H_0: There is no statistically significant difference between Human ratings and LCSS ratings.

- H_1: There is a statistically significant difference between Human ratings and LCSS ratings.

The result is:

- $t = 0.104$, $df = 50$ (P-Value > 0.05)

As a result, it fails to reject the null hypothesis, which means the ratings produced by humans are not statistically significantly different from the rating of LCSS. This means the rating produced by LCSS is not statistically significantly different from the Human rating over the TWS-51. As a consequence, LCSS is considered to be part of the nTWSS.

To calculate Steiger's z-test between two measures requires the construction of a correlation triangle. In this case, we considered comparing the correlation between TWSS and the TWS-51 human rating with the correlation between LCSS and the TWS-51 human rating. Correlation triangles are formed as shown in Figure 5.14 and the specific triangle for this calculation is formed according to Figure 5.15.

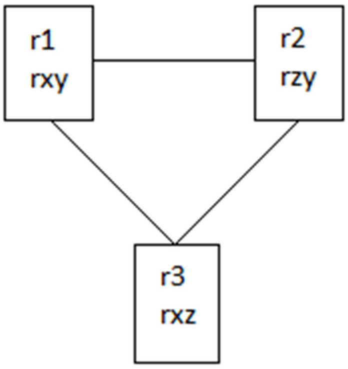


Figure 5.14: General Form of Correlation Triangle

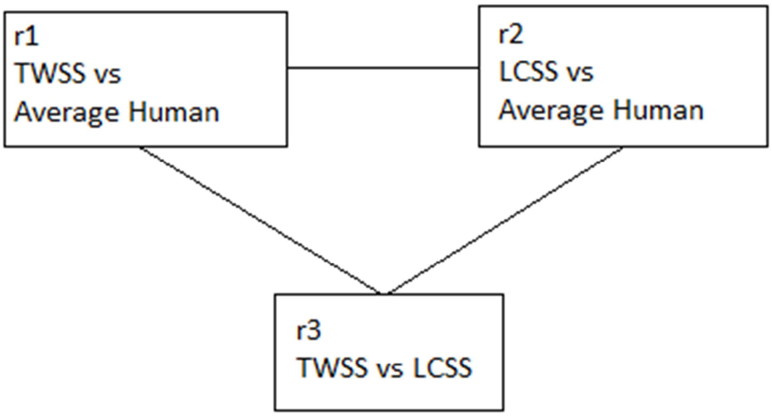


Figure 5.15: Specific Correlation Triangle for TWSS vs LCSS

From Table 5.4:

*r*1 rxy TWSS vs Average human	0.752
*r*2 rzy LCSS vs Average human	0.723
n = 51	

Calculate correlation:

*r*3 rxz TWSS vs LCSS	0.447

Applying the test gives the following results:

- *z* = 0.346, *df* = 48 (P-Value > 0.05)

As a result, the null hypothesis is accepted. This means TWSS and LCSS are not statistically significantly different from the TWS-51 dataset.

Table 5.6 shows the *Problem Word Pairs*, which are the same pairs shown in Table 4.7, Section 4.5.3. Those pairs are the pairs where the rating between TWSS and Thai human semantic similarity is different by more than 1.

Table 5.6: Problem Word Pairs

WP	***W_1***		***W_2***		***Human***	**TWSS**	**LCSS**
21	ทาส	Slave	หมา	Dog	0.555	1.781	1.595
27	ทาส	Slave	เด็กหนุ่ม	Lad	1.160	2.176	0.810
37	นักมายากล	Magician	พ่อมด	Wizard	2.010	3.964	2.137
38	วัด	Temple	สุสาน	Graveyard	2.150	0.387	2.255
41	อาหาร	Food	ผลไม้	Fruit	2.363	0.578	2.232
43	วัด	Temple	พระ	Monk	2.675	0.513	2.788
61	นักบวช	Priest	พระ	Monk	3.575	1.194	3.208
63	วัด	Temple	โบสถ์	Church	3.693	2.677	2.805
64	ครู	Teacher	อาจารย์	Lecturer	3.783	2.671	3.506

The most underperforming word pair of TWSS is word pair 61 for which the human rating is 3.575, since the word *นักบวช* (Priest) is a subset of the word *พระ* (Monk) in Thai. However, TWSS used WordNet as knowledge, based on English, which made TWSS underperform for this word pair, giving the rating 1.194, a difference of 2.381 from the Thai human rating. On the other hand, the rating from LCSS is 3.208 for word pair 61, which performs significantly better than TWSS, with a difference of only 0.367 from the Thai human rating. The Pearson Product-Moment correlation coefficients (*r*), Spearman's rank correlation coefficients (ρ), Kendall's tau rank correlation coefficients (τ) between the Thai human rating and TWSS results are:

- Pearson's *r* = 0.037 (P-Value > 0.05)
- Spearman's ρ = 0.083 (P-Value > 0.05)
- Kendall's tau τ = 0.111 (P-Value > 0.05)

Furthermore, the Pearson Product-Moment correlation coefficients (*r*), Spearman's rank correlation coefficients (ρ), Kendall's tau rank correlation coefficients (τ) between the Thai human rating and LCSS results are:

- Pearson's *r* = 0.900 (P-Value < 0.05)
- Spearman's ρ = 0.950 (P-Value < 0.05)
- Kendall's tau τ = 0.833 (P-Value < 0.05)

Steiger's z-test (Steiger, 1980) is used to find whether or not TWSS and LCSS are statistically significantly different (α=0.05) from the hypotheses:

- H_0: There are no statistically significant differences between two measures.
- H_1: There are statistically significant differences between two measures.

To calculate Steiger's z-test between two requires the construction of a correlation triangle. In this case, we considered comparing the correlation between TWSS and the *Problem Word Pairs* human rating with the correlation between LCSS and the *Problem Word Pairs* human rating. The specific triangle for this calculation is formed according to Figure 5.16.

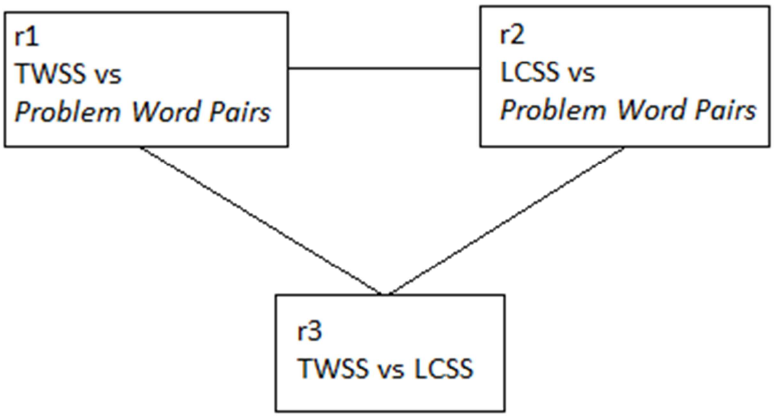


Figure 5.16: Specific Correlation Triangle for TWSS vs LCSS with the *Problem Word Pairs*

From Table 5.6:

*r*1 rxy TWSS vs *Problem Word Pairs* average human 0.037

*r*2 rzy LCSS vs *Problem Word Pairs* average human 0.900

n = 9

Calculate correlation:

*r*3 rxz TWSS vs LCSS -0.052

Applying the test gives the following results:

- z = -2.334, df = 6 (P-Value < 0.05)

With the result, the null hypothesis is rejected. This means TWSS and LCSS are statistically significantly different from the *Problem Word Pairs* human rating.

To compare TWSS and LCSS shows that the correlation coefficients between the *Problem Word Pairs* human rating and LCSS are significantly better than the correlation coefficients between the *Problem Word Pairs* human rating and TWSS with a difference of 0.863. It can be concluded that LCSS overcomes the problem in the Thai culture aspect as stated at the beginning of the chapter which TWSS cannot deal with very well.

5.6 Conclusion

This chapter cornered the creation of LCSS, a new Thai word similarity measure. It began by describing the details of the problem with the TWSS. Then it provided a step-by-step specification of the algorithm. Considerations included the removal of Thai function words, maximum length of the lexical chain, and the use of word frequency and link frequency. The training and testing datasets were also explained and prepared for the evaluation of LCSS. The evaluation of LCSS and the testing set (TWS-51) were subsequently discussed. Both the TWSS and LCSS perform well on their own, with TWSS significantly better and LCSS marginally better. However, the evidence shows that each contributes a different insight into the similarity process. Therefore, a combination may be more effective, which is the subject of Chapter 6.

Chapter 6

New Thai Word Semantic Similarity Measure (nTWSS)

6.1 Introduction

In Chapter 4, the problem of TWSS was stated and LCSS was shown to provide a different insight for words of particular importance in Thai culture in Chapter 5. Thus, the idea of creating a better version of TWSS was inspired. Combining the best feature of TWSS and LCSS could create a new version of TWSS (nTWSS) with better performance. Therefore, the research question investigated in this chapter is whether a combination of TWSS and LCSS can provide a better model of human perception of Thai word semantic similarity than either separately.

The contributions in this chapter are:

- Creation of nTWSS
- Evaluation of nTWSS.

The rest of this chapter is organized as follows: in Section 6.2 an experiment is designed to find the most suitable combination between TWSS and LCSS; Section 6.3 evaluates the new algorithm, nTWSS, with respect to human similarity ratings; and the conclusions are given in Section 6.4.

6.2 New Thai Word Semantic Similarity Measure (nTWSS)

As stated in Chapter 5, there are two Thai word measures, TWSS and LCSS. TWSS uses Li's word measure (Li et al., 2003) and translates Thai words into English, and the English WordNet (Miller, 1995) is used as its knowledge, as explained in Chapter 3. The LCSS algorithm in Chapter 5 uses information derived from a search engine to construct a chain of words graph, used to create a lexical chain from which word similarity is derived. In Chapter 5, a number of experiments were conducted to evaluate TWSS over the TWS-51 dataset, which shows that the Pearson Product-Moment correlation coefficients between Thai human rating and TWSS rating (r = 0.752, P-Value < 0.01) perform better than the correlation between the worst Thai native speaker and the rest of the group (r = 0.708). Likewise, the correlation coefficients between the Thai human rating and LCSS (r = 0.723, P-Value < 0.01) also perform better than the worst Thai native speaker and the rest of the group. This means that both TWSS and LCSS are viable measures in their own right. As mentioned in Chapters 4 and 5, the *Problem Word Pairs* in Table 5.6 are specifically grounded in Thai culture. It was found that TWSS performs poorly (r = 0.037) on the *Problem Word Pairs*. However, LCSS performs significantly better then TWSS with a correlation coefficient of 0.900 over the *Problem Word Pairs*, as discussed in Section

5.5.3. However, the overall performance of TWSS (*r* = 0.752) is still better than LCSS (*r* = 0.723) over the TWS-51 dataset. This leads to the conjecture that each of the two word similarity measures has a different fundamental insight into Thai word similarity. Accordingly, the idea of creation of nTWSS is inspired.

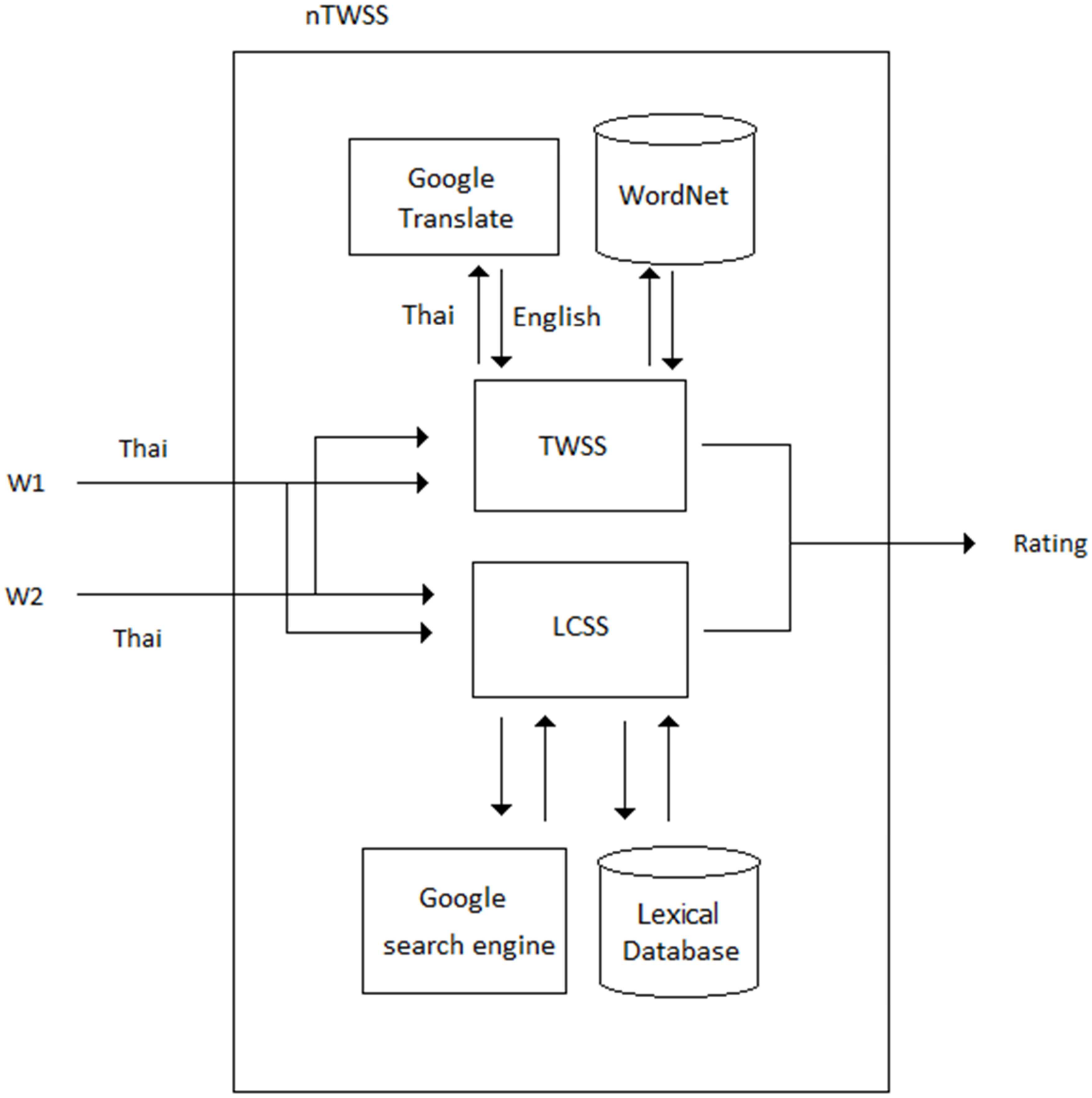


Figure 6.1: An Overview of nTWSS

Figure 6.1 shows an overview of how nTWSS, TWSS and LCSS will be combined to create nTWSS. The similarity rating between two words of nTWSS, *s(w1,w2)*, can be calculated as:

$$s(w1, w2) = \delta\, S_{TWSS} + (1 - \delta) S_{LCSS} \qquad \textbf{Equation 6.1}$$

where S_{TWSS} is the word similarity rating obtained from TWSS, S_{LCSS} is the word similarity rating obtained from LCSS, and δ is a constant in the rage $0 < \delta < 1$ which adjusts the proportion between the two components. An experiment was conducted to find the most suitable value of δ between TWSS and LCSS in Section 6.3. This is a similar approach to STASIS, but STASIS uses δ to adjust the rating between semantic component and word order component.

6.3 Experiment to Find the Most Suitable δ Parameter

The aim of this experiment is to find the most suitable δ parameter for a combination between TWSS and LCSS by using TWS-30 as a training set described in Section 5.3.1. The δ parameter is a constant in Equation 6.1.

6.3.1 Methodology

The training set and testing set are independent as the testing set removes 14 word pairs that consist of the same meaning with the training set, as explained in Section 5.3. This training set was also used in the experiment to find the most suitable value of the *Alpha* parameter for the Thai language in Section 5.4, in which the result of the experiment assigned the value of the *Alpha* parameter as 0.2. Thus, in this experiment the *Alpha* parameter will be assigned as 0.2. The values of the δ parameter used in this experiment are in the range 0-1.

For each value of the δ parameter, the procedure is as follows:

- Calculate the TWSS rating for each word pair in TWS-30 (Thai-English translation included).
- The lexical database was pre-populated by collecting data from Google following STEP1-2 in Chapter 5 for all word pairs in TWS-30.
- Calculate the LCSS rating for each of the word pairs in TWS-30 by applying the value of *Alpha* parameter as 0.2.
- Calculate the nTWSS by combining ratings from TWSS and LCSS for each proportion.
- Calculate the correlation coefficients (r) between Human rating and nTWSS for each proportion.

Choose the δ parameter from the highest value of correlation coefficients (r).

The Pearson Product-Moment correlation coefficients (r) between Human rating and nTWSS for each proportion will be calculated and shown in Section 6.3.2.

6.3.2 nTWSS Rating with Different *δ* Parameter

Table 6.1 displays the semantic similarity ratings for the nTWSS. Column *WP* shows the number of the word pairs in the training set. Column *Human* presents the human rating for the Thai word pairs. Column *TWSS* lists the TWSS rating which can be calculated as described in Chapter 3. Column *LCSS* shows the LCSS rating, where the *Alpha* value = 0.2. The columns *δ parameter* present a series of values of *δ* (the 5 values are selected for presentation as they are the most promising values from the set tested). All of the measures have been scaled in the range 0 to 1 to aid comparison.

Table 6.1: Semantic Similarity Rating of nTWSS

WP	*Human*	*TWSS*	*LCSS*	***δ* Parameter**				
				0.7	0.6	0.5	0.4	0.3
1	0.020	0.097	0.000	0.068	0.058	0.048	0.039	0.029
5	0.006	0.070	0.000	0.049	0.042	0.035	0.028	0.021
9	0.017	0.016	0.000	0.011	0.010	0.008	0.006	0.005
13	0.171	0.110	0.232	0.146	0.159	0.171	0.183	0.195
17	0.158	0.322	0.276	0.308	0.303	0.299	0.294	0.290
21	0.150	0.365	0.238	0.327	0.314	0.301	0.289	0.276
25	0.137	0.176	0.235	0.194	0.200	0.206	0.212	0.218
29	0.149	0.145	0.330	0.200	0.219	0.237	0.256	0.274
33	0.541	0.322	0.228	0.293	0.284	0.275	0.265	0.256
37	0.315	0.298	0.221	0.275	0.268	0.260	0.252	0.244
41	0.325	0.365	0.407	0.377	0.382	0.386	0.390	0.394
47	0.403	0.448	0.411	0.437	0.433	0.429	0.426	0.422
48	0.393	0.991	0.112	0.727	0.639	0.552	0.464	0.376
49	0.605	1.000	0.235	0.770	0.694	0.618	0.541	0.465
50	0.221	0.214	0.233	0.220	0.222	0.224	0.225	0.227
51	0.781	0.818	0.235	0.643	0.585	0.527	0.468	0.410
52	0.583	0.996	0.328	0.796	0.729	0.662	0.596	0.529
53	0.836	0.544	0.773	0.613	0.636	0.659	0.682	0.704
54	0.697	0.819	0.209	0.636	0.575	0.514	0.453	0.392
55	0.806	0.816	0.462	0.710	0.674	0.639	0.604	0.568
56	0.805	0.801	0.232	0.630	0.573	0.516	0.460	0.403
57	0.708	0.978	0.548	0.849	0.806	0.763	0.720	0.677
58	0.834	0.816	0.586	0.747	0.724	0.701	0.678	0.655
59	0.379	1.000	0.245	0.773	0.698	0.622	0.547	0.471
60	0.606	0.811	0.361	0.676	0.631	0.586	0.541	0.496
61	0.509	0.816	0.615	0.755	0.735	0.715	0.695	0.675
62	0.850	0.999	0.576	0.872	0.830	0.788	0.745	0.703
63	0.770	1.000	0.440	0.832	0.776	0.720	0.664	0.608
64	0.752	1.000	0.698	0.909	0.879	0.849	0.819	0.788
65	0.769	0.999	0.586	0.875	0.834	0.793	0.752	0.710

6.3.3 Results

The Pearson product-moment correlation coefficients of a series of proportions between TWSS and LCSS ratio over TWS-30 dataset are shown in Table 6.2

Table 6.2: Correlation Coefficients

δ Parameter	0.7	0.6	0.5	0.4	0.3
Correlation	0.865	0.875	0.879	0.875	0.857
P-Value	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01

As seen above, Table 6.2 shows that the best correlation of nTWSS can be obtained by using $\delta = 0.5$; i.e. the best correlation can be obtained by using half of the TWSS rating and half of the LCSS rating. The Pearson Product-Moment correlation coefficients (*r*) between the Thai human rating and nTWSS result is:

- Pearson's r = 0.879 (P-Value < 0.01).

According to this result, the nTWSS will calculate the word semantic similarity between two target words by using $\delta = 0.5$.

6.4 Evaluation of TWS-51 with nTWSS

The aim of this section is to describe a series of experiments that were conducted using the Testing set (TWS-51) from Section 5.3.2 to evaluate the nTWSS.

6.4.1 Methodology

The Testing dataset (TWS-51) from Section 5.3.2 is used to evaluate nTWSS. The nTWSS rating is calculated by using $\delta = 0.5$, found to be the most suitable value (as described in Section 6.3). The *Alpha* parameter will be assigned as 0.2 as explained in Section 5.4.

The procedure for this experiment is as follows:

- Translate all word pairs in TWS-51 in to English using Google Translate.
- Calculate TWSS rating for each word pair in TWS-51.
- The lexical database was pre-populated by collecting data from Google following STEP1-2 in Chapter 5 for all word pairs in TWS-51.
- Calculate the LCSS rating for each word pair in TWS-51 by applying the value of *Alpha* parameter as 0.2.
- Calculate the nTWSS by using Equation 6.1.
- Calculate the correlation coefficients (*r*) between Human rating and nTWSS.

The Pearson Product-Moment correlation coefficients (*r*) between Thai human rating and nTWSS measure is shown in Section 6.4.3.

6.4.2 Results

Table 6.3 shows the semantic similarity ratings for the nTWSS. Column *WP* present the number of the word pairs of TWS-65. Column *Human* displays the human rating for the Thai word pairs. Column *TWSS* shows the TWSS rating for the Thai word pairs described in Section 3.2. Column *LCSS* shows the LCSS rating for the Thai word pairs described in Section 5.2. Column *nTWSS* shows the nTWSS rating, half from TWSS half from the LCSS rating, described in Section 6.2. All of the measures have been scaled in the range 0 to 1 to aid comparison.

Table 6.3: nTWSS Semantic Similarity Rating

WP	Human	TWSS	LCSS	nTWSS
1	0.014	0.144	0.218	0.181
2	0.017	0.216	0.000	0.108
3	0.024	0.176	0.000	0.088
4	0.028	0.144	0.000	0.072
5	0.031	0.044	0.213	0.128
6	0.034	0.014	0.207	0.110
7	0.044	0.097	0.000	0.048
8	0.056	0.044	0.202	0.123
9	0.061	0.132	0.104	0.118
10	0.069	0.144	0.207	0.175
11	0.069	0.157	0.211	0.184
12	0.073	0.079	0.000	0.040
14	0.080	0.097	0.318	0.207
15	0.086	0.053	0.333	0.193
16	0.104	0.128	0.397	0.262
17	0.105	0.145	0.203	0.174
18	0.110	0.108	0.360	0.234
19	0.133	0.108	0.494	0.301
20	0.135	0.197	0.209	0.203
21	0.139	0.445	0.399	0.422
22	0.163	0.360	0.201	0.280
23	0.246	0.105	0.566	0.335
24	0.261	0.360	0.221	0.291
25	0.271	0.365	0.221	0.293
26	0.274	0.448	0.000	0.224
27	0.290	0.544	0.202	0.373
28	0.294	0.520	0.319	0.419
29	0.316	0.244	0.567	0.406
30	0.321	0.548	0.223	0.386
31	0.334	0.445	0.678	0.562

32	0.353	0.244	0.202	0.223
33	0.388	0.216	0.201	0.208
34	0.406	0.244	0.561	0.403
35	0.430	0.365	0.204	0.284
36	0.481	0.263	0.202	0.233
38	0.538	0.097	0.564	0.330
39	0.553	0.547	0.364	0.455
40	0.564	0.322	0.568	0.445
41	0.591	0.144	0.558	0.351
42	0.603	0.668	0.698	0.683
43	0.669	0.128	0.697	0.413
44	0.686	0.445	0.394	0.420
45	0.726	0.586	0.396	0.491
46	0.754	0.818	0.521	0.669
49	0.763	0.664	0.202	0.433
55	0.799	0.670	0.401	0.536
60	0.853	0.716	0.568	0.642
61	0.894	0.298	0.802	0.550
63	0.923	0.669	0.701	0.685
64	0.946	0.668	0.876	0.772
65	0.981	1.000	0.975	0.988

6.4.3 Discussion

The experimental results in Section 6.4.2 suggest that the nTWSS measure and semantic similarity of human rating provides good results. As can be seen in Figure 6.2, most of the data points are near the linear line (dotted line). The Pearson Product-Moment correlations obtained from these results are:

- Pearson's r = 0.867 (P-Value = 0.000).

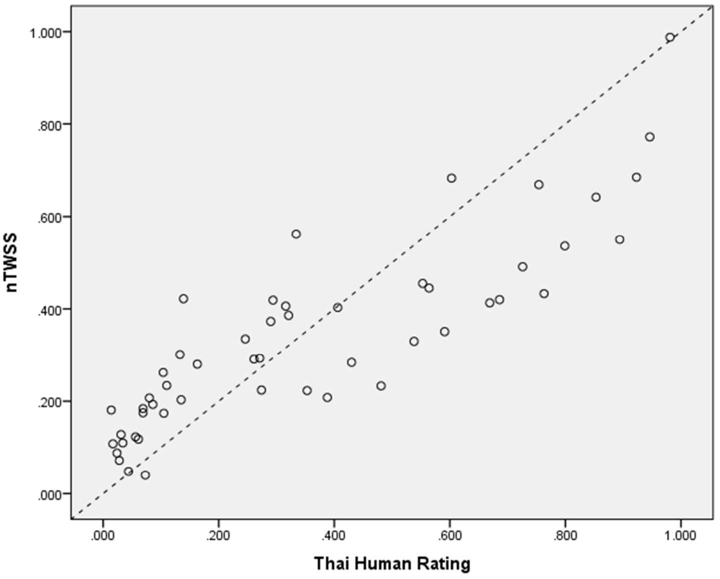


Figure 6.2: The Correlation between TWS-51 and nTWSS

Table 6.4 illustrates the agreement of three measures with human ratings by calculating the correlation coefficients (*r*) between the human ratings and the machine ratings over the testing dataset. Also, human performance using the leave-one-out resampling technique to find the correlation coefficient between each participant and the rest of the group is shown in this table.

Table 6.4: The Pearson Product-Moment Correlation Coefficients (*r*)

	Correlation *r*	**P-Value**
Thai human similarity rating and TWSS	0.752	0.000
Thai human similarity rating and LCSS	0.723	0.000
Thai human similarity rating and nTWSS	0.867	0.000
Average Thai native speaker and the least of the group	0.865	-
Worst Thai native speaker participant and the least of the group	0.708	-
Best Thai native speaker participant and the least of the group	0.928	-

The nTWSS measure performed better than TWSS and LCSS alone, with a difference of 0.115 between the TWSS and nTWSS correlation coefficients and a difference of 0.144 between the LCSS and nTWSS correlation coefficients.

The paired sample t-test was used to find whether or not Human ratings and nTWSS ratings over the dataset is statistically significantly different (α=0.05) from the hypotheses:

- H_0: There is no statistically significant difference between Human ratings and nTWSS ratings.
- H_1: There is a statistically significant difference between Human ratings and nTWSS ratings.

The result is:

- t = 0.819, df = 50 (P-Value > 0.05).

From this result, it fails to reject the null hypothesis, meaning that there is no evidence to support the position that the ratings produced by nTWSS are not statistically significantly different from the Human ratings.

Applying Steiger's z-test (Steiger, 1980) to find whether or not TWSS and nTWSS are statistically significantly different (α=0.05), the hypotheses are:

- H_0: There is no statistically significant difference between two measures.
- H_1: There is a statistically significant difference between two measures.

To calculate Steiger’s z-test between two measures requires the construction of a correlation triangle. In this case, we considered comparing the correlation between TWSS and the TWS-51 human rating with the correlation between nTWSS and the TWS-51 human rating. The specific triangle for this calculation is formed according to Figure 6.3.

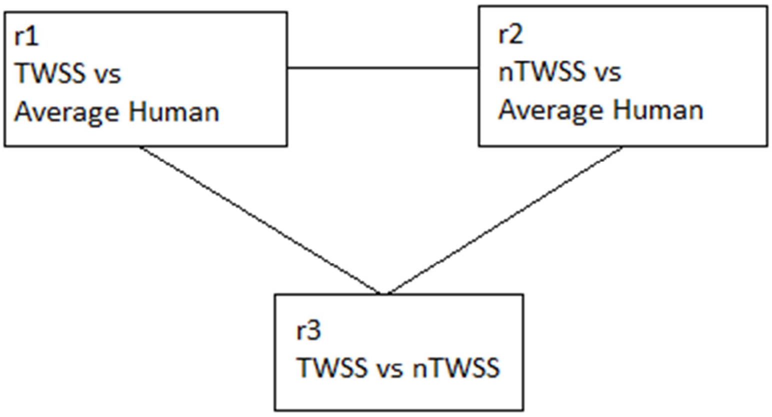


Figure 6.3: Specific Correlation Triangles for TWSS vs nTWSS

From Table 5.4:

r1 rxy TWSS vs Average human	0.752
r2 rzy nTWSS vs Average human	0.863
n = 51	

Calculate correlation:

r3 rxz TWSS vs nTWSS	0.849

Applying the test gives the following result:

- z = 2.736, df = 48 (P-Value < 0.01).

From its result, the null hypothesis is rejected, meaning TWSS and nTWSS are statistically significantly different. This means nTWSS (r = 0.867) performs significantly better than TWSS (r = 0.752) with the TWS-51 dataset.

Moreover, nTWSS performs close to the best performance of the humans and the rest of the group (r = 0.928) with a difference of 0.061 between the two correlation coefficients, which supports the view that this is an effective algorithm and it should be useful to any future research on Thai semantic similarity.

6.5 Conclusion

This chapter described how the nTWSS measure works by combining the best features of TWSS and LCSS; showed the experiments conducted to evaluate nTWSS; and discussed the experimental results. The nTWSS showed a positive result with the correlation coefficients (r = 0.867). Moreover, the subsequent procedure of this research is to create a Thai short text semantic similarity measure that uses nTWSS to calculate the rating between two target words. To do this, the preliminary stage is to obtain a sentence benchmark dataset, which will be discussed in detail in the next chapter.

Chapter 7

A 65 Sentence Thai Benchmark Dataset (TSS-65)

7.1 Introduction

In Chapter 6, the word measure (nTWSS) gave a promising result (*r* = 0.867). However, the overall aim of this research is to propose a sentence similarity measure suitable for Thai Conversational Agents which utilizes nTWSS. To do this, a Thai sentence benchmark dataset is needed for the evaluation. The aim of this chapter is to create the first sentence semantic similarity dataset (TSS-65) based on the Thai language. As there is no prior Thai sentence semantic similarity dataset, a methodology is required to create one. This will be the first of its kind and will contribute substantially to future research on the Thai language. The proposed methodology adapts procedures previously shown to be effective in other languages for the Thai language. Hence, this research will create a Thai sentence semantic similarity dataset adapting the methodology used to create the STSS-65 dataset (O'Shea et al., 2008; O'Shea et al., 2010) and Thai culture will be taken into account during its creation. This Thai sentence semantic similarity benchmark dataset (TSS-65) will contain 65 sentence pairs with human ratings. TSS-65 consists of pairs corresponding to TWS-65 (i.e. each sentence pair from TSS-65 is created from a word pair in TWS-65). The creation of TSS-65 is explained step-by-step as follows:

- The Methodology for creating TSS-65
- The Application of the methodology to rating TSS-65
- Evaluation of TWS-65 with TSS-65.

The rest of this chapter is organized as follows: Section 7.2 describes the methodology for creating TSS-65; Section 7.3 describes the methodology to produce TSS-65; Section 7.4 discusses TSS-65; and Section 7.5 is the conclusion.

7.2 Creation of the TSS-65 Dataset

The aim of this section is to describe the methodology for finding a set of Thai sentence pairs for TSS-65 which follows the procedure from the STSS-65 dataset (O'Shea et al., 2008, 2010). The reason the STSS-65 procedure was chosen is that, according to O’Shea et al. (2008), STSS-65 is specifically created to evaluate STSS measures. The STSS-65 dataset is adopted by a number English STSS researchers to evaluate or compare their algorithms. This methodology was used to create a Gold Standard STSS-65 dataset (O'Shea et al., 2010) utilizing word pairs from the R&G dataset (Rubenstein and Goodenough, 1965). The sentence pairs in the TSS-65 dataset will correspond to the word pairs TWS-65 from Chapter 4 which is the dataset based on the Thai language. TWS-65 contains 48 nouns arranged in various combinations to make up the 65 word pairs. The

TSS-65 dataset is built from TWS-65, by adopting the single Royal Institute dictionary definition (Thai Royal Institute, 2011) of the 65 word pairs plus 4 more definitions generated by 4 native Thai speakers as the materials. The Royal Institute dictionary was chosen over other dictionaries to provide the definitions because this is the official dictionary for the Thai language. Also, it is used to teach students at primary and high school (Thai Royal Institute, 2011). In addition, for each noun, four native Thai speakers were asked to provide a sentence that contains that noun to represent the definition. The four definitions from native Thai speakers serve as a substitute in the case that the definition from the dictionary is too complicated or not commonly used. An experiment was conducted to find the most suitable definition for each noun in TWS-65. TSS-65 will be presented in Section 7.3.4.

7.2.1 Experiment to Find the Most Suitable Definition for Each Noun in TWS-65

Following O’Shea et al. (2006), the aim of this experiment is to find the most suitable definition for each noun in TWS-65, which will be used to create TSS-65.

7.2.1.1 Participants

The definition for each noun in TWS-65 was chosen by 20 native Thai speakers to create TSS-65. The participants have 12 Art/Humanities and 8 Science/Engineering backgrounds. They consisted of 8 undergraduates and 12 postgraduates studying at 6 different Thai universities. The average age of the participants was 26, with 8 males and 12 females.

7.2.1.2 Materials

There are 48 nouns in TWS-65. There are five possible definitions for each noun by adopting from one definition from the Thai-Thai dictionary (The Royal Institute, 2011) and four more definitions by native Thai speakers.

The data collection instrument was a questionnaire in which each noun in TWS-65 was printed with five definitions for selection in a standard Thai font (see Appendix 3.2) and a small amount of personal data (Name, Confirmation of being a native Thai speaker, Age, Gender, and Academic background) to ensure a representative sample of the Thai population. Examples of experimental materials used are:

- Appendix 1.5 The Person Data Collection Sheet
- Appendix 3.1 The Instruction Sheet
- Appendix 3.2 Sample Question Sheet.

7.2.1.3 Procedure

The participants were asked to perform the following procedure:

1. Please read through all definitions for each noun.
2. Please select the definition that you feel is the best definition to represent that noun. You can only select one definition for each noun.

The definitions were shuffled into a random order when presented to the participants.

7.2.1.4 Results

Tables 7.1-7.48 show definitions of each noun and number of participants choosing that definition. Column *Definition* shows the five definitions. Column *Noun* shows the original word from TWS-65 and five definitions. Column *Number* shows number of participants choosing the definition.

Table 7.1: Noun *ลายมือชื่อ* (Autograph)

Definition	Noun *ลายมือชื่อ*	Number
1	ลายมือชื่อคือสัญลักษณ์แทนเจ้าตัว An autograph symbolizes the person.	12
2	ลายมือชื่อคือลายลักษณ์อักษรเพื่อแทนตน An autograph is handwriting representing the person.	4
3	ลายมือชื่อคือชื่อที่ถูกเขียนเพื่อเป็นสัญลักษณ์แทนผู้เขียน An autograph is a name written to represent the writer.	3
4	ลายมือชื่อคือชื่อของบุคคลซึ่งเขียนลงไว้เพื่อรับรองว่าตนเป็นผู้ทำหนังสือ An autograph is a name of individual written in order to verify their authenticity of the document.	1
5	ลายมือชื่อคืออักษรชื่อที่เขียนด้วยลายมือชื่อของผู้ทำธุรกรรม An autograph is a hand-written letter of the person making the transaction.	0

Table 7.2: Noun *ลายเซ็น* (Signature)

Definition	Noun *ลายเซ็น*	Number
1	ลายเซ็นคือสัญลักษณ์แทนเจ้าตัว A signature symbolizes the person.	12
2	ลายเซ็นคือสัญลักษณ์ที่ถูกเขียนเพื่อรับรองผู้เขียน A signature is a symbol written to verify the writer.	4
3	ลายเซ็นคือสัญลักษณ์แทนลายมือชื่อ A signature symbolizes an autograph.	3
4	ลายเซ็นคือลายมือชื่อที่เขียนหวัด A signature is a scribbled autograph.	1
5	ลายเซ็นคือลายมือชื่อ A signature is an autograph.	0

Table 7.3: Noun *เด็กผู้ชาย (Boy)*

Definition	Noun *เด็กผู้ชาย*	Number
1	เด็กผู้ชายคือมนุษย์วัยเยาว์เพศผู้	6
	A boy is a young male human.	
2	เด็กผู้ชายคือมนุษย์เพศผู้ที่อยู่ระหว่างการเกิดและวัยแรกรุ่น	4
	A boy is an early age male between birth and teenage.	
3	เด็กผู้ชายคือบุคคลชายอายุเกินกว่าเจ็ดปีบริบูรณ์แต่ยังไม่เกินกว่าสิบห้าปีบริบูรณ์	4
	A boy is a male individual aged over 7 years but not exceeding 15 years old.	
4	เด็กผู้ชายคือมนุษย์เพศผู้ก่อนวัยเจริญพันธุ์	3
	A boy is a male before the reproductive age.	
5	เด็กผู้ชายคือผู้ชายที่มีอายุไม่ถึง ๑๕ ปีบริบูรณ์	3
	A boy is a male aged not over 15 years old.	

Table 7.4: Noun *เด็กหนุ่ม (Lad)*

Definition	Noun *เด็กหนุ่ม*	Number
1	เด็กหนุ่มคือผู้ชายอายุน้อย	10
	A lad is a young man.	
2	เด็กหนุ่มคือวัยรุ่นชาย	8
	A lad is a male teenager.	
3	เด็กหนุ่มคือชายที่มีอายุพ้นวัยเด็ก	1
	A lad is a male whose age is over childhood.	
4	เด็กหนุ่มคือชายที่ยังดูไม่แก่ตามวัย	1
	A lad is a man whose look is not as old as his age.	
5	เด็กหนุ่มคือชายที่มีอายุตั้งแต่ ๑๕-๓๐ ปี	0
	A lad is a man aged between 15-30 years old.	

Table 7.5: Noun *ฝั่งทะเล (shore)*

Definition	Noun *ฝั่งทะเล*	Number
1	ฝั่งทะเลคือชายฝั่งที่ติดทะเล	5
	A shore is a coast close to the sea.	
2	ฝั่งทะเลคือที่ดินติดริมทะเล	4
	A shore is land close to the sea	
3	ฝั่งทะเลคือหาดทะเล	4
	A shore is a beach.	
4	ฝั่งทะเลคือบริเวณบกที่ประชิดกับทะเล	4
	A shore is an area close to the sea.	
5	ฝั่งทะเลคือริมหาดทะเล	3
	A shore is the edge of the beach.	

Table 7.6: Noun *ชายฝั่ง (coast)*

Definition	Noun *ชายฝั่ง*	Number
1	ชายฝั่งคือชายทะเล	6
	A coast is a beach.	
2	ชายฝั่งคือแนวแผ่นดินจากทะเลเป็นต้นไปบนบก	5
	A coast is a land line from the sea onwards.	
3	ชายฝั่งคือฝั่งบนบกในแนวทะเล	3
	A coast is an area on the land alongside the sea line.	
4	ชายฝั่งคือแถบแผ่นดินนับจากแนวชายทะเลขึ้นไปบนบกจน	3
	A coast is a land line from the sea line to the land.	
5	ชายฝั่งคือแนวชายทะเลขึ้นไปบนบกจนถึงบริเวณ	3
	A coast is a sea line up to the land.	

Table 7.7: Noun *ป่าช้า (Cemetery)*

Definition	Noun *ป่าช้า*	Number
1	ป่าช้าคือสถานที่ฝังศพ	15
	A cemetery is a place to bury corpses.	
2	ป่าช้าคือสถานที่รวบรวมหลุมศพ	3
	A cemetery is a place that gathers corpses.	
3	ป่าช้าคือสถานที่จัดเตรียมไว้เพื่อทำพิธีกรรมหลังความตาย	1
	A cemetery is a place organized for the after death ritual.	
4	ป่าช้าคือสถานที่รวบรวมศพผู้ตาย	1
	A cemetery is the place that gathers dead people.	
5	ป่าช้าคือป่าหรือที่ซึ่งจัดไว้เป็นที่ฝังหรือเผาศพ	0
	A cemetery is a forest or a place used for burying or burning corpses.	

Table 7.8: Noun *สุสาน (graveyard)*

Definition	Noun *สุสาน*	Number
1	สุสานคือสถานที่เก็บศพ	14
	A graveyard is a place to store corpses.	
2	สุสานคือที่ฝังหรือเผาศพ	4
	A graveyard is place for burying or burning corpses.	
3	สุสานคือที่ฝังศพผู้ตาย	2
	A graveyard is a place for burying corpses.	
4	สุสานคือสถานที่ทำลายศพผู้ตาย	0
	A graveyard is a place to destroy corpses.	
5	สุสานคือหลุมศพ	0
	A graveyard is a grave.	

Table 7.9: Noun *การเดินทาง* (Journey)

Definition	**Noun *การเดินทาง***	**Number**
1	การเดินทางคือการเคลื่อนย้ายจากสถานที่หนึ่งไปยังที่หนึ่ง	11
	A journey is to travel from one place to another.	
2	การเดินทางคือการสัญจร	5
	A journey is to travel.	
3	การเดินทางคือการย้ายตำแหน่งของสิงมีชีวิต	2
	A journey is a movement of a living thing.	
4	การเดินทางคือการไปที่โน่นที่นี่เรื่อยไป	1
	A journey is to travel around.	
5	การเดินทางคือการเคลื่อนที่เพื่อจุดประสงค์ใดจุดประสงค์หนึ่ง	1
	A journey is a movement for a specific purpose.	

Table 7.10: Noun *การท่องเที่ยว* (Travelling)

Definition	**Noun *การท่องเที่ยว***	**Number**
1	การท่องเที่ยวคือการเดินทางในช่วงขณะหนึ่งเพื่อพักผ่อน	8
	Travelling is to travel at a certain time for leisure.	
2	การท่องเที่ยวคือการเดินทางเผื่อผ่อนคลาย	6
	Travelling is a journey for leisure.	
3	การท่องเที่ยวคือการเตร็ดเตร่ไปเพื่อหาความสนุกเพลิดเพลินตามที่ต่างๆ	4
	Travelling is to wander around for leisure and entertainment in places.	
4	การท่องเที่ยวคือการทัศนาจร	1
	Travelling is an excursion.	
5	การท่องเที่ยวคือการเปลี่ยนตำแหน่งชั่วคราวเพื่อการผักผ่อน	1
	Travelling is a temporary movement for leisure.	

Table 7.11: Noun *ทาส* (Slave)

Definition	**Noun *ทาส***	**Number**
1	ทาสคือข้ารับใช้	14
	A slave is a thrall.	
2	ทาสคือขี้ข้าหรือบริวาร	4
	A slave is a thrall or servant.	
3	ทาสคือผู้ไม่มีอิสระในตัว	2
	A slave is a person with no freedom.	
4	ทาสคือผู้ที่อุทิศตนแก่สิ่งที่เลื่อมใสศรัทธา	2
	A slave is a person who devotes himself to what he has faith in.	
5	ทาสคือผู้ที่ยอมตนให้ตกอยู่ในอำนาจสิ่งใดสิ่งหนึ่ง	0
	A slave is a person who accepts being under the power of something.	

Table 7.12: Noun ***ข้ารับใช้*** *(Serf)*

Definition	**Noun *ข้ารับใช้***	**Number**
1	ข้ารับใช้คือผู้รับใช้	8
	A serf is a servant.	
2	ข้ารับใช้คือคนใช้ของเจ้านาย	8
	A servant is a serf of the master.	
3	ข้ารับใช้คือผู้ดูแล	4
	A servant is a care-taker.	
4	ข้ารับใช้คือผู้แวดล้อมหรือผู้ติดตาม	0
	A serf is an entourage.	
5	ข้ารับใช้คือผู้ไม่มีอิสระในตัว	0
	A servant is a person with no freedom.	

Table 7.13: Noun ***อุปกรณ์*** *(Equipment)*

Definition	**Noun *อุปกรณ์***	**Number**
1	อุปกรณ์คือเครื่องมืออำนวยความสะดวก	8
	Equipment is facilities.	
2	อุปกรณ์คือสิ่งของสำหรับใช้ในการงาน	7
	Equipment is tools used for work.	
3	อุปกรณ์คือวัตถุที่นำมาใช้งาน	2
	Equipment is an object used for work.	
4	อุปกรณ์คือเครื่องช่วยหรือเครื่องประกอบ	2
	Equipment is a helping tool.	
5	อุปกรณ์คือเครืองใช้ต่างๆ	1
	Equipment is tools.	

Table 7.14: Noun ***เครื่องมือ*** *(Tool)*

Definition	**Noun *เครื่องมือ***	**Number**
1	เครื่องมือคือสิ่งที่ใช้เพื่อทุนแรง	7
	Tools are used for labour-saving devices.	
2	เครื่องมือคือสิ่งของอำนวยความสะดวก	7
	Tools are facilities.	
3	เครื่องมือคืออุปกรณ์เพื่อช่วยเหลือหรือช่วยเสริม	3
	Tools are helping equipment.	
4	เครื่องมือคืออุปกรณ์หรือบุคคลที่ถูกใช้	3
	Tools are equipment or people being used.	
5	เครื่องมือคือสิ่งของสำหรับใช้ในการงาน	0
	Tools are objects used for work.	

Table 7.15: Noun *เที่ยงวัน* (Midday)

Definition	Noun *เที่ยงวัน*	Number
1	เที่ยงวันคือเวลาสิบสองนาฬิกา	6
	Midday is the time at noon.	
2	เที่ยงวันคือเวลาเที่ยงตรง	5
	Midday is at noon.	
3	เที่ยงวันคือจุดตรงกลางระหว่างวัน	4
	Midday is the middle point of the day.	
4	เที่ยงวันคือจุดหนึ่งของเวลากลางวัน	3
	Midday is a point during the daytime.	
5	เที่ยงวันคือเวลาในกลางวัน	2
	Midday is the time during the day.	

Table 7.16: Noun *กลางวัน* (Noon)

Definition	Noun *กลางวัน*	Number
1	กลางวันคือระยะเวลาราว ๆ เที่ยงวัน	16
	Noon is the time around twelve o'clock.	
2	กลางวันคือเวลารุ่งสางถึงโพล้เพล้	2
	Noon is dawn to the sunset.	
3	กลางวันคือเวลาที่พระอาทิตย์ขึ้น	1
	Noon is the time when the sun rises.	
4	กลางวันคือระยะเวลาตั้งแต่ย่ำรุ่งถึงย่ำค่ำ	1
	Noon is the duration from dawn to late evening.	
5	กลางวันคือส่วนของวันตั้งแต่รุ่งถึงค่ำ	0
	Noon is part of the day from dawn to early night.	

Table 7.17: Noun *อัญมณี* (Gem)

Definition	Noun *อัญมณี*	Number
1	อัญมณีคือแร่ธรรมชาติที่มีมูลค่า	7
	Gem is a natural mineral of value.	
2	อัญมณีคือรัตนชาติที่เจียระไนแล้ว	4
	Gem is the precious jewels.	
3	อัญมณีคือแก้วมณีอื่นๆนอกจากเพชรพลอย	4
	Gem is other precious stone other than jewels.	
4	อัญมณีคือเพชรนิลจินดา	3
	Gem is jewellery.	
5	อัญมณีคือมวลของแข็งที่ประกอบไปด้วยแร่ชนิดเดียวกัน	2
	Gem is a solid form of mineral.	

Table 7.18: Noun *เพชรพลอย* (Jewel)

Definition	Noun *เพชรพลอย*	Number
1	เพชรพลอยคือเครื่องประดับที่มีมูลค่า	7
	Jewel is accessories of value.	
2	เพชรพลอยคือชื่อแก้วที่แข็งมีน้ำแวววาว	7
	Jewel is a sparkling precious stone.	
3	เพชรพลอยคือเครื่องเพชรพลอย	4
	Jewel is precious stone	
4	เพชรพลอยคือเครื่องประดับมีมูลค่า	2
	Jewel is accessories of value.	
5	เพชรพลอยคืออัญมณี	0
	Jewel is gem.	

Table 7.19: Noun *เนินเขา* (Hill)

Definition	Noun *เนินเขา*	Number
1	เนินเขาคือลักษณะภูมิประเทศสูงขึ้นไปเล็กน้อย	6
	Hill is a little high-up terrain.	
2	เนินเขาคือโคกขนาดใหญ่ที่ค่อยลาดสูงขึ้นจากระดับเดิม	6
	Hill is a high-up slope.	
3	เนินเขาคือพืนที่ลาดสูงขึ้นจากระดับเดิม	4
	Hill is a high-up place from the same level.	
4	เนินเขาคือที่สูงหรือที่ดอน	2
	Hill is the high place.	
5	เนินเขาคือพื้นที่ที่มีระดับสูงขึ้นจากบริเวณรอบๆแต่ไม่สูงมากเท่าภูเขา	2
	Hill is the area high up from other places but not as high as a mountain.	

Table 7.20: Noun *ภูเขา* (Mountain)

Definition	Noun *ภูเขา*	Number
1	ภูเขาคือลักษณะภูมิประเทศสูงขึ้นไปมาก	11
	Mountain is very high-up terrain.	
2	ภูเขาคือเนินหินที่สูงขึ้นเป็นจอมใหญ่	3
	Mountain is a pile of rocks.	
3	ภูเขาคือเขาขนาดใหญ่หรือสูง	3
	Mountain is a big or high hill.	
4	ภูเขาคือพื้นที่สูงชัน	2
	Mountain is the steep area.	
5	ภูเขาคือพื้นที่ที่มีระดับสูงขึ้นจากบริเวณรอบๆ ตั้งแต่ ๖๐๐ เมตรขึ้นไป	1
	Mountain is the place higher than other areas up to 600 metres.	

Table 7.21: Noun *ป่าไม้* (Forest)

Definition	Noun *ป่าไม้*	Number
1	ป่าไม้คืออาณาเขตซึ่งอุดมไปด้วยต้นไม้	9
	Forest is an area of abundant trees.	
2	ป่าไม้คือดินแดนที่เต็มไปด้วยพรรณต้นไม้	6
	Forest is land full of plants.	
3	ป่าไม้คือที่ที่มีต้นไม้ต่างๆ ขึ้นมา	2
	Forest is a place with many plants.	
4	ป่าไม้คือที่ดินที่ยังมิได้มีบุคคลได้มาตามประมวลกฎหมายที่ดิน	2
	Forest is land which has not been acquired by anyone according to the law.	
5	ป่าไม้คือที่ดินที่ไม่มีบุคคลใดบุคคลหนึ่งครอบครอง	1
	Forest is land with no owner.	

Table 7.22: Noun *พงไพร* (Woods)

Definition	Noun *พงไพร*	Number
1	พงไพรคือป่าไม่ชนิดนึง	6
	Woods are one kind of forest.	
2	พงไพรคือคงหญ้าหรือดงไม้ที่รวมกันเป็นผืนป่า	5
	Woods are bushes formed into a forest.	
3	พงไพรคือพื้นที่ป่า	4
	Woods are forest areas.	
4	พงไพรคือพื้นที่ที่อุดมไปด้วยพรรณไม้	4
	Woods are the areas of abundant plants.	
5	พงไพรคือป่ารกชัฏ	1
	Woods are overgrown forests.	

Table 7.23: Noun *ยานพาหนะ* (Vehicle)

Definition	Noun *ยานพาหนะ*	Number
1	ยานพาหนะคือเครื่องจักรใช้ในการเดินทาง	9
	A vehicle is a machine used for transportation.	
2	ยานพาหนะคือสัตว์สำหรับขี่บรรทุกหรือลากเข็น	7
	A vehicle is animals for riding and carrying.	
3	ยานพาหนะคือเครื่องขับขี่	3
	A vehicle is a riding machine.	
4	ยานพาหนะคือเครื่องนำไป	1
	A vehicle is the guiding machine.	
5	ยานพาหนะคือเครื่องขับขี่มีรถและเรือเป็นต้น	0
	A vehicle is a driving machine such as car and ship.	

Table 7.24: Noun *รถยนต์ (Car)*

Definition	Noun *รถยนต์*	Number
1	รถยนต์คือยานพาหนะสี่ล้อ	17
	A car is a four-wheeled vehicle.	
2	รถยนต์คือพาหนะชนิดหนึ่ง	1
	A car is a vehicle.	
3	รถยนต์ที่มีล้อตั้งแต่3 ล้อและเดินด้วยกำลังเครื่องยนต์	1
	A car with more than 3 wheels and driven by a motor.	
4	รถยนต์คือยานพาหนะที่ขับเคลื่อนด้วยเครื่องยนต์	1
	A car is a vehicle driven by a motor.	
5	รถยนต์คือยานพาหนะทุกชนิดที่ใช้ในการขนส่งทางบก	0
	A car is all kinds of vehicle used in land transportation.	

Table 7.25: Noun *อาหาร (Food)*

Definition	Noun *อาหาร*	Number
1	อาหารคือสิ่งที่สิ่งมีชีวิตรับประทาน	8
	Food is what living creatures eat.	
2	อาหารคือเครื่องหล่อเลี้ยงชีวิต	6
	Food is what makes people survive.	
3	อาหารคือสสารใด ๆซึ่งบริโภคเพื่อเสริมโภชนาการให้แก่ร่างกาย	5
	Food is any substances consumed for nutrients.	
4	อาหารคือของกิน	1
	Food is edible things.	
5	อาหารคือเครื่องค้ำจุนชีวิต	0
	Food is what makes people live.	

Table 7.26: Noun *ผลไม้ (Fruit)*

Definition	Noun *ผลไม้*	Number
1	ผลไม้คืออาหารที่ได้จากต้นไม้	6
	Fruit is food derived from a tree.	
2	ผลไม้คือลูกไม้	4
	Fruit is fruit.	
3	ผลไม้คือลูกหรือผลของต้นไม้	4
	Fruit is fruit from tree.	
4	ผลไม้คือผลผลิตจากพืชเพื่อการขยายพันธุ์	4
	Fruit is a product of a plant for reproduction.	
5	ผลไม้คือผลที่เกิดจากการขยายพันธุ์ของต้นไม้	2
	Fruit is a product of plant reproduction.	

Table 7.27: Noun *แก้ว* (Glass)

Definition	Noun *แก้ว*	Number
1	แก้วคือภาชนะบรรจุของเหลว	14
	A glass is a container containing liquid.	
2	แก้วคือภาชนะที่ทำด้วยแก้วสำหรับใส่น้ำกิน	3
	Glass is a container for water, made of glass.	
3	แก้วคือวัสดุแข็งที่มีรูปลักษณะอยู่ตัวและเป็นเนื้อเดียว	2
	Glass is a solid material.	
4	แก้วคือหินแข็งใสแลลอดเข้าไปข้างในได้	1
	Glass is a transparent stone.	
5	แก้วคือของที่ได้จากการใช้ทรายขาว	0
	Glass derives from white sand.	

Table 7.28: Noun *ถ้วย* (Bowl)

Definition	Noun *ถ้วย*	Number
1	ถ้วยคือภาชนะทรงโค้งหงายใช้เพื่อบรรจุของเหลว	7
	A bowl is a rounded container containing liquid.	
2	ถ้วยคือภาชนะก้นลึกมีรูปต่างๆ สำหรับใส่น้ำหรือของบริโภค	5
	A bowl is container for water or anything for consumption.	
3	ถ้วยคืออุปกรณ์ก้นลึกบรรจุของต่างๆ	3
	A bowl is an equipment containing object.	
4	ถ้วยคือชามขนาดเล็กมีรูปต่างๆ	3
	A bowl is a small bowl with different shapes.	
5	ถ้วยคือลักษณนามเรียกถ้วยที่มีสิ่งของบรรจุ	2
	A bowl is a classifier used to call a bowl with something in it.	

Table 7.29: Noun *นักบวช* (Priest)

Definition	Noun *นักบวช*	Number
1	นักบวชคือผู้ถือศีลทางศาสนาคริสต์	16
	A priest is a person observing the precepts in Christianity.	
2	นักบวชคือผู้สืบทอดศาสนา	3
	A priest is a person pertaining to religion.	
3	นักบวชคือบรรพชิต	1
	A priest is a priest.	
4	นักบวชคือผู้ทรงศีล	0
	A priest is a person observing the precepts.	
5	นักบวชคือผู้ถือบวช	0
	A priest is a priest.	

Table 7.30: Noun *พระ* (Monk)

Definition	Noun *พระ*	Number
1	พระคือผู้ถือศีลอาศัยอยู่ในวัด	7
	A monk is a person living in a temple.	
2	พระคือนักบวชในศาสนา	6
	A monk is a priest in relations.	
3	พระคือพระสงฆ์	3
	A monk is a monk.	
4	พระคือพระพุทธรูป	2
	A monk is a statue of Buddha.	
5	พระคือคำใช้แทนชื่อเรียกภิกษุสงฆ์	2
	A monk is a word used to represent monk.	

Table 7.31: Noun *นักมายากล* (Magician)

Definition	Noun *นักมายากล*	Number
1	นักมายากลแสดงราวกับใช้เวทมนต์	6
	A magician performs as if using magic.	
2	นักมายากลคือนักเล่นกล	5
	A magician is a magician.	
3	นักมายากลคือนักแสดงกล	5
	A magician performs magic.	
4	นักมายากลคือผู้มีความสามารถพิเศษ	3
	A magician is a person with special abilities.	
5	นักมายากลคือผู้แสดงที่ลวงตาให้เห็นเป็นจริง	1
	A magician is a person performing as if it is real.	

Table 7.32: Noun *พ่อมด* (Wizard)

Definition	Noun *พ่อมด*	Number
1	พ่อมดคือผู้ใช้เวทมนต์เพศผู้	7
	A wizard is male person using magic.	
2	พ่อมดคือผู้วิเศษ	4
	A wizard is a magic person.	
3	พ่อมดคือผู้ใช้คาถาอาคม	4
	A wizard is a person using magic.	
4	พ่อมดคือผู้ใช้อำนาจเหนือธรรมชาติ	3
	A wizard is a person using super power.	
5	พ่อมดคือชายที่ใช้อำนาจทำอะไรได้ผิดธรรมดา	2
	A wizard is a male using great power.	

Table 7.33: Noun *ผ้าฝ้าย (Muslin)*

Definition	Noun *ผ้าฝ้าย*	Number
1	ผ้าฝ้ายคือผ้าที่ผลิตมาจากดอกฝ้าย	8
	Muslin is a cloth produced from cotton.	
2	ผ้าฝ้ายคือสิ่งทอจากดอกฝ้าย	4
	Muslin is a cloth produced from cotton.	
3	ผ้าฝ้ายคือเครื่องนุ่งห่มทำด้วยดอกฝ้าย	3
	Muslin is clothes made of cotton.	
4	ผ้าฝ้ายคือสิ่งที่ทำด้วยเยื่อใยโดยวิธีทอหรืออัดดอกฝ้าย	3
	Muslin is made of the tissue of cotton.	
5	ผ้าฝ้ายคือผลิตภัณฑ์ชนิดหนึงจากดอกฝ้าย	2
	Muslin is a product made of cotton.	

Table 7.34: Noun *ผ้าไหม (Silk)*

Definition	Noun *ผ้าไหม*	Number
1	ผ้าไหมคือผ้าที่ผลิตมาจากเส้นใยจากตัวไหม	8
	Silk is a cloth produced from silkworm fibres.	
2	ผ้าไหมคือสิ่งทอจากใยจากตัวไหม	5
	Silk is clothes from silkworm fibres.	
3	ผ้าไหมคือเครื่องนุ่งห่มทำด้วยใยจากตัวไหม	3
	Silk is clothes from silkworm.	
4	ผ้าไหมคือสิ่งที่ทำด้วยเยื่อใยโดยวิธีทอหรืออัดใยจากตัวไหม	2
	Silk is made of tissue from silkworm fibres.	
5	ผ้าไหมคือผลิตภัณฑ์ชนิดหนึงจากใยจากตัวไหม	2
	Silk is a product produced from silkworm fibres.	

Table 7.35: Noun *ครู (Teacher)*

Definition	Noun *ครู*	Number
1	ครูคือผู้สอนในโรงเรียน	14
	A teacher is someone teaching at school.	
2	ครูคือผู้สั่งสอนศิษย์	4
	A teacher is a person teaching students.	
3	ครูคือผู้ถ่ายทอดความรู้ให้แก่ศิษย์	2
	A teacher is a person sharing knowledge with students.	
4	ครูคือผู้มีความหนักแน่น	0
	A teacher is a steady person.	
5	ครูคือผู้ควรแก่การเคารพ	0
	A teacher is a person worth respect.	

Table 7.36: Noun *อาจารย์* *(Lecturer)*

Definition	Noun *อาจารย์*	Number
1	อาจารย์คือผู้สอนในมหาวิทยาลัย	15
	A Lecturer is a person teaching at university.	
2	อาจารย์คือคำที่ใช้เรียกนำหน้าชื่อบุคคลเพื่อแสดงความยกย่องว่ามีความรู้ในทางใดทางหนึ่ง	5
	A Lecturer is a word used as the title of person to show respect as an expert in a field.	
3	อาจารย์คือคุณครู	0
	A Lecturer is a teacher.	
4	อาจารย์คือนักปราชญ์	0
	A Lecturer is a philosopher.	
5	อาจารย์คือผู้สอนวิชาและความประพฤติ	0
	A Lecturer is a person teaching subjects and behaviours.	

Table 7.37: Noun *นิตยสาร* *(Magazine)*

Definition	Noun *นิตยสาร*	Number
1	นิตยสารคือหนังสือออกรายสัปดาห์หรือรายเดือน	8
	A magazine is a book published weekly or monthly.	
2	นิตยสารคือหนังสือพิมพ์ที่ออกเป็นรายคาบ	6
	A magazine is a newspaper published in a period.	
3	นิตยสารคืองานเขียนที่ออกเป็นรายคาบ	4
	A magazine is writing published in a period.	
4	นิตยสารคือรายงานหรือบันทึกที่ออกเป็นรายคาบ	2
	A magazine is a report or record published in a period.	
5	นิตยสารคือสิ่งพิมพ์รายคาบที่ออกเป็นระยะสำหรับผู้อ่านทั่วไป	0
	A magazine is printed matter published in a period for general readers.	

Table 7.38: Noun *หนังสือ* *(Book)*

Definition	Noun *หนังสือ*	Number
1	หนังสือคือสิ่งบันทึกตัวอักษร	6
	A book is a recorder of letters.	
2	หนังสือคืองานเขียน	6
	A book is writing.	
3	หนังสือคือรายงานหรือบันทึก	4
	A book is a report or record.	
4	หนังสือคือสิ่งพิมพ์	2
	A book is printed matter.	
5	หนังสือคือสิ่งพิมพ์เก็บความรู้	2
	A book is printed matter recording knowledge.	

Table 7.39: Noun *วัด* (Temple)

Definition	Noun *วัด*	Number
1	วัดคือที่สถานที่พักพิงทางศาสนาพุทธ	11
	A temple is a place for Buddhism.	
2	วัดคือสถานที่ทางศาสนา	3
	A temple is a place for religions.	
3	วัดคือที่อยู่ของสงฆ์หรือนักบวช	3
	A temple is a place for monks or priests.	
4	วัดคือสอบขนาดหรือปริมาณของสิ่งต่างๆ	2
	A temple is to measure the size or quantity of things.	
5	วัดคืออาราม	1
	A temple is a temple.	

Table 7.40: Noun *โบสถ์* (Church)

Definition	Noun *โบสถ์*	Number
1	โบสถ์คือสถานที่ทางศาสนาคริสต์	12
	A church is a Christian place	
2	โบสถ์คือสถานที่ที่นักบวชประชุม	3
	A church is a place where priests have meetings.	
3	โบสถ์คือสถานที่ระกอบพิธีกรรมศักดิ์สิทธิ	2
	A church is a place for sacred rituals.	
4	โบสถ์คือสถานที่ประกอบพิธีกรรมของศาสนาอื่นๆที่ไม่ใช่ศาสนาพุทธ	2
	A church is a place for sacred rituals of other religions which are not Buddhism.	
5	โบสถ์คือสถานที่สำหรับนักบวชใช้ประชุม	1
	A church is a place where the priests have meetings.	

Table 7.41: Noun *ลุง* (Uncle)

Definition	Noun *ลุง*	Number
1	ลุงคือพี่ชายของบิดาหรือมารดา	7
	An uncle is a brother of father or mother.	
2	ลุงคือลักษณะผู้ชายที่ทำตัวแก่เกินวัยทั้งใบหน้า การแต่งกาย ทัศนคติ และการวางตัว	5
	Uncle is a characteristic of a male who acts and looks older in terms of appearance, attitudes and manner.	
3	ลุงคือคำเรียกชายที่ไม่รู้จักแต่มักจะมีอายุแก่กว่าพ่อหรือแม่	4
	Uncle is a term used to call a male stranger, usually older than parents.	
4	ลุงคือชายที่มีวัยไล่เลี่ยแต่แก่กว่าพ่อหรือแม่	3
	Uncle is a man of similar and older age to parents.	
5	ลุงคือสามีของป้า	1
	Uncle is aunt's husband.	

Table 7.42: Noun *ป้า (Aunt)*

Definition	Noun *ป้า*	Number
1	ป้าคือพี่สาวของบิดาหรือมารดา	7
	An aunt is a sister of father or mother.	
2	ป้าคือลักษณะผู้หญิงที่ทำตัวแก่เกินวัยทั้งใบหน้า การแต่งกาย ทัศนคติ และการวางตัว	6
	Aunt is a characteristic of a female who acts and looks older in terms of appearance, attitudes and manner.	
3	ป้าคือคำเรียกหญิงที่ไม่รู้จักแต่มักจะมีอายุแก่กว่าพ่อหรือแม่	3
	Aunt is a term used to call a female stranger, usually older than parents.	
4	ป้าคือหญิงที่มีวัยไล่เลี่ยแต่แก่กว่าพ่อหรือแม่	3
	Aunt is a woman of similar and older age to parents.	
5	ป้าคือภรรยาของลุง	1
	Aunt is uncle's husband.	

Table 7.43: Noun *สุนัข (Dog)*

Definition	Noun *สุนัข*	Number
1	สุนัขคือหมาใช้ในภาษาทางการ	14
	A dog is a dog in official language.	
2	สุนัขคือหมา	3
	A dog is a dog	
3	สุนัขคือชื่อสัตว์เลี้ยงลูกด้วยนมชนิดหนึ่ง	2
	A dog is one kind of mammal.	
4	สุนัขคือสัตว์ที่เฝ้าบ้าน	1
	A dog is a domestic animal.	
5	สุนัขคือสัตว์เลี้ยงลูกด้วยนมมีเขี้ยว 2 คู่ ตีนหน้ามี 5 นิ้ว ตีนหลังมี 4 นิ้ว	0
	A dog is a mammal with two fangs, five-finger forelegs and four-finger back legs.	

Table 7.44: Noun *หมา (Dog)*

Definition	Noun *หมา*	Number
1	หมาคือสัตว์เลี้ยงลูกด้วยนมชนิดหนึ่ง	9
	Dog is one kind of mammal.	
2	หมาคือสัตว์เลี้ยงลูกด้วยนมมีเขี้ยว 2 คู่ ตีนหน้ามี 5 นิ้ว ตีนหลังมี 4 นิ้ว	4
	A dog is a mammal with two fangs, five-finger forelegs and four-finger back legs.	
3	หมาคือสุนัขในภาษาชาวบ้าน	3
	A dog is a dog in general language.	
4	หมาคือสัตว์ที่เฝ้าบ้าน	3
	A dog is a domestic animal.	
5	หมาคือสุนัข	1
	A dog is a dog.	

Table 7.45: Noun *โรงภาพยนต์ (Cinema)*

Definition	Noun *โรงภาพยนต์*	Number
1	โรงภาพยนตร์คือสถานที่จัดแสดงภาพยนตร์	5
	A cinema is a place to show movies.	
2	โรงภาพยนตร์คือสถานที่ผักผ่อนหย่อนใจชนิดหนึ่ง	5
	A cinema is a place for leisure.	
3	โรงภาพยนตร์คือสถานที่ฉายภาพยนตร์	5
	A cinema is a place to show movies.	
4	โรงภาพยนตร์คือโรงหนังภาพยนตร์	3
	A cinema is a cinema.	
5	โรงภาพยนตร์คือสถานที่เฉพาะสำหรับฉายภาพยนตร์	2
	A cinema is a place only for showing movies.	

Table 7.46: Noun *โรงละคร (Theatre)*

Definition	Noun *โรงละคร*	Number
1	โรงละครคือสถานที่จัดแสดงละคร	6
	A theatre is a place for shows.	
2	โรงละครคือสถานที่ผักผ่อนหย่อนใจชนิดหนึ่ง	6
	A theatre is a place for leisure.	
3	โรงละครคือสถานที่ฉายละคร	4
	A theatre is a place showing dramas.	
4	โรงละครคือสถานที่เล่นละคร	3
	A theatre is the place for playing dramas.	
5	โรงละครคือสถานที่เฉพาะสำหรับฉายละคร	1
	A theatre is the place only for showing dramas.	

Table 7.47: Noun *พืช (Plant)*

Definition	Noun *พืช*	Number
1	พืชคือสิ่งมีชิวิตสีเขียว	8
	A plant is a green living thing.	
2	พืชคือเมล็ดพันธุ์ไม้สิ่งที่จะเป็นพันธุ์ต่อไป	4
	A plant is a seed to be reproduced.	
3	พืชคือพรรณไม้ที่งอกอยู่ตามที่ต่างๆ	4
	A plant is plant growing in places.	
4	พืชคือต้นไม้ต่างๆ	2
	A plant is trees.	
5	พืชคือส่วนใดส่วนหนึ่งของพืชที่แยกแล้วก็ยังสามารถจะงอกได้	2
	A plant is a part of plant which, even cut off, is still able to grow.	

Table 7.48: Noun *ต้นไม้ (Tree)*

Definition	Noun *ต้นไม้*	Number
1	ต้นไม้คือพืชคือสิ่งมีชิวิตสีเขียวชนิดนึง	8
	A tree is a type of plant.	
2	ต้นไม้คือคำรวมเรียกพืชทั่วไปโดยปกติชนิดมีลำต้น	6
	A tree is a general term to call a plant, normally having a trunk.	
3	ต้นไม้คือไม้ยืนต้นขนาดใหญ่	2
	A tree is a big perennial tree.	
4	ต้นไม้คือพืชที่มีอายุยืนยาว	2
	A tree is a long-living plant.	
5	ต้นไม้คือพืชชนิดที่มีลำต้นใหญ่มีกิ่งแยกออกไป	2
	A tree is a plant with a huge trunk and branches.	

The definitions are those which were chosen by the highest number of participants. However, if there were two or more definitions that obtained the same number of participants, the definition for that word was randomly chosen from the most popular definitions.

7.2.1.5 TSS-65 Sentence Pairs

TSS-65 is created by replacing the words from the TWS-65 with the most suitable definition from Section 7.2.1.4. Table 7.49 shows the TSS-65 sentence pairs. Column *SP* is the sentence pair number. Column *TSS-65* is the sentence pair corresponding with the word pair in TWS-65 in Column *TWS-65*.

Table 7.49: TSS-65 Sentence Pairs

SP	TWS-65		TSS-65	
	***W*₁**	***W*₂**	***S*₁**	***S*₂**
1	แก้ว	ข้ารับใช้	แก้วคือภาชนะบรรจุของเหลว	ข้ารับใช้คือผู้รับใช้
	Glass	Serf	A glass is a container containing liquid	A serf is a servant.
2	อาหาร	ลายเซ็น	อาหารคือสิ่งที่สิ่งมีชีวิตรับประทาน	ลายเซ็นคือสัญลักษณ์แทนเจ้าตัว
	Food	Signature	Food is what living creatures eat.	A signature symbolizes the person.
3	อัญมณี	ลายเซ็น	อัญมณีคือแร่ธรรมชาติที่มีมูลค่า	ลายเซ็นคือสัญลักษณ์แทนเจ้าตัว
	Gem	Signature	Gem is a natural mineral of value.	A signature symbolizes the person.
4	ฝั่งทะเล	รถยนต์	ฝั่งทะเลคือชายฝั่งที่ติดทะเล	รถยนต์คือยานพาหนะสี่ล้อ
	Coast	Car	A shore is a coast close to the sea.	A car is a four-wheeled vehicle.
5	สุนัข	เครื่องมือ	สุนัขคือหมาใช้ในภาษาทางการ	เครื่องมือคือสิ่งที่มนุษย์ใช้เพื่อทุนแรง
	Dog	Tool	A dog is a dog in official language.	Tool is used for a labour-saving device.
6	การเดินทาง	สุสาน	การเดินทางคือการเคลื่อนย้ายจากสถานที่หนึ่งไปยังที่หนึ่ง	สุสานคือสถานที่เก็บศพ
	Journey	Graveyard	Journey is to travel from one place to another.	A graveyard is a place to store corpses.
7	เที่ยงวัน	โรงละคร	เที่ยงวันคือเวลาสิบสองนาฬิกา	โรงละครคือสถานที่จัดแสดงละคร
	Midday	Theatre	Midday is the time at noon.	A theatre is a place for shows.
8	เที่ยงวัน	การท่องเที่ยว	เที่ยงวันคือเวลาสิบสองนาฬิกา	การท่องเที่ยวคือการเดินทางในช่วงขณะหนึ่งเพื่อพักผ่อน
	Midday	Voyage	Midday is the time at noon.	Journey is to travel at a certain time for leisure.
9	ยานพาหนะ	เพรชพลอย	ยานพาหนะคือเครื่องจักรใช้ในการเดินทาง	เพชรพลอยคือเครื่องประดับที่มีมูลค่า
	Automobile	Jewel	A vehicle is a machine used for transportation.	Jewel is accessories of value.
10	เนินเขา	ผลไม้	เนินเขาคือลักษณะภูมิประเทศสูงขึ้นไปเล็กน้อย	ผลไม้คืออาหารที่ได้จากต้นไม้
	Hill	Fruit	Hill is a little high-up terrain.	Fruits are food derived from a tree.
11	นักมายากล	ถ้วย	นักมายากลแสดงราวกับใช้เวทมนต์	ถ้วยคือภาชนะทรงโค้งหงายใช้เพื่อบรรจุของเหลว
	Magician	Cup	A magician performs as if using magic.	A bowl is a rounded container containing liquid.
12	ป่าช้า	หมา	ป่าช้าคือสถานที่ฝังศพ	หมาคือสัตว์เลี้ยงลูกด้วยนมชนิดหนึ่ง
	Cemetery	Dog	A cemetery is a place to bury corpses.	Dog is one kind of mammal.

13	ฝั่งทะเล	พงไพร	ฝั่งทะเลคือชายฝั่งที่ติดทะเล	พงไพรคือป่าไม่ชนิดนึง
	Coast	Woods	A shore is a coast close to the sea.	Woods are one kind of forest.
14	นักมายากล	เครื่องมือ	นักมายากลแสดงราวกับใช้เวทมนต์	เครื่องมือคือสิ่งที่มนุษย์ใช้เพื่อทุนแรง
	Magician	Tool	A magician performs as if using magic.	Tool is used for a labour-saving device.
15	การเดินทาง	กลางวัน	การเดินทางคือการเคลื่อนย้ายจากสถานที่หนึ่งไปยังที่หนึ่ง	กลางวันคือระยะเวลาราว ๆ เที่ยงวัน
	Journey	Noon	Journey is to travel from one place to another.	Noon is around twelve o’clock.
16	นิตยสาร	ป้า	นิตยสารคือหนังสือออกรายสัปดาห์หรือรายเดือน	ป้าคือพี่สาวของบิดาหรือมารดา
	Magazine	Aunt	A magazine is a book published weekly or monthly.	An aunt is a sister of father or mother.
17	นักบวช	หนังสือ	นักบวชคือผู้ถือศีลทางศาสนาคริสต์	หนังสือคือสิ่งบันทึกตัวอักษร
	Priest	Book	A priest is a person observing the precepts in Christianity.	A book is a recorder of letters.
18	เด็กผู้ชาย	หมา	เด็กผู้ชายคือมนุษย์วัยเยาว์เพศผู้	หมาคือสัตว์เลี้ยงลูกด้วยนมชนิดหนึ่ง
	Boy	Dog	A boy is a young male human.	Dog is one kind of mammal.
19	สุนัข	เด็กหนุ่ม	สุนัขคือหมาใช้ในภาษาทางการ	เด็กหนุ่มคือผู้ชายอายุน้อย
	Dog	Lad	A hound is a dog in official language.	A lad is a young man.
20	วัด	พงไพร	วัดคือที่สถานที่พักพิงทางศาสนาพุทธ	พงไพรคือป่าไม่ชนิดนึง
	Temple	Woods	A temple is a place for Buddhism.	Woods are one kind of forest.
21	ทาส	หมา	ทาสคือข้ารับใช้	หมาคือสัตว์เลี้ยงลูกด้วยนมชนิดหนึ่ง
	Slave	Dog	A slave is a thrall.	Dog is one kind of mammal.
22	อาหาร	ถ้วย	อาหารคือสิ่งที่สิ่งมีชีวิตรับประทาน	ถ้วยคือภาชนะทรงโค้งหงายใช้เพื่อบรรจุของเหลว
	Food	Cup	Food is what living creatures eat.	A bowl is a rounded container containing liquid.
23	ครู	หนังสือ	ครูคือผู้สอนในโรงเรียน	หนังสือคือสิ่งบันทึกตัวอักษร
	Teacher	Book	A teacher is someone teaching at school.	A book is a recorder of letters.
24	พืช	ผ้าไหม	พืชคือสิ่งมีชิวิตสีเขียว	ผ้าไหมคือผ้าที่ผลิตมาจากเส้นใยจากตัวไหม
	Plant	Silk	A plant is a green living thing.	Silk is a cloth produced from silkworm fibres.
25	เด็กผู้ชาย	อาจารย์	เด็กผู้ชายคือมนุษย์วัยเยาว์เพศผู้	อาจารย์คือผู้สอนในมหาวิทยาลัย
	Boy	Lecturer	A boy is a young male human.	A teacher is a person teaching at university.
26	โรงภาพยนต์	โบสถ์	โรงภาพยนตร์คือสถานที่จัดแสดงภาพยนตร์	โบสถ์คือสถานที่ทางศาสนาคริสต์

	Cinema	Church	A cinema is a place to show movies.	A church is a Christian place.
27	ทาส	เด็กหนุ่ม	ทาสคือข้ารับใช้	เด็กหนุ่มคือผู้ชายอายุน้อย
	Slave	Lad	A slave is a thrall.	A lad is a young man.
28	เนินเขา	ชายฝั่ง	เนินเขาคือลักษณะภูมิประเทศสูงขึ้นไปเล็กน้อย	ชายฝั่งคือชายทะเล
	Hill	Shore	Hill is a little high-up terrain.	A coast is a beach.
29	ยานพาหนะ	เครื่องมือ	ยานพาหนะคือเครื่องจักรใช้ในการเดินทาง	เครื่องมือคือสิ่งที่มนุษย์ใช้เพื่อทุนแรง
	Automobile	Tool	A vehicle is a machine used for transportation.	Tool is used for labour-saving device.
30	ผ้าฝ้าย	ต้นไม้	ผ้าฝ้ายคือผ้าที่ผลิตมาจากดอกฝ้าย	ต้นไม้คือพืชชนิดนึง
	Cotton	Tree	Muslin is a cloth produced from cotton.	A tree is a type of plant.
31	อุปกรณ์	รถยนต์	อุปกรณ์คือเครื่องมืออำนวยความสะดวก	รถยนต์คือยานพาหนะสี่ล้อ
	Implement	Car	Equipment is a tool used as facilities.	A car is a four-wheeled vehicle.
32	ลุง	อาจารย์	ลุงคือพี่ชายของบิดาหรือมารดา	อาจารย์คือผู้สอนในมหาวิทยาลัย
	Uncle	Lecturer	An uncle is a brother of father or mother.	A teacher is a person teaching at university.
33	ป่าไม้	ผลไม้	ป่าไม้คืออาณาเขตซึ่งอุดมไปด้วยต้นไม้	ผลไม้คืออาหารที่ได้จากต้นไม้
	Forest	Fruit	Forest is an area of abundant trees.	Fruit is food derived from a tree.
34	ครู	ป้า	ครูคือผู้สอนในโรงเรียน	ป้าคือพี่สาวของบิดาหรือมารดา
	Teacher	Aunt	A teacher is someone teaching at school.	An aunt is a sister of father or mother.
35	นักบวช	พ่อมด	นักบวชคือผู้ถือศีลทางศาสนาคริสต์	พ่อมดคือผู้ใช้เวทมนต์เพศผู้
	Priest	Wizard	A priest is a person observing the precepts in Christianity.	A wizard is a male person using magic.
36	แก้ว	เพชรพลอย	แก้วคือภาชนะบรรจุของเหลว	เพชรพลอยคือเครื่องประดับที่มีมูลค่า
	Glass	Jewel	A glass is a container containing liquid.	Jewel is accessories of value.
37	นักมายากล	พ่อมด	นักมายากลแสดงราวกับใช้เวทมนต์	พ่อมดคือผู้ใช้เวทมนต์เพศผู้
	Magician	Wizard	A magician performs as if using magic.	A wizard is a male person using magic.
38	วัด	สุสาน	วัดคือที่สถานที่พักพิงทางศาสนาพุทธ	สุสานคือสถานที่เก็บศพ
	Temple	Graveyard	A temple is a place for Buddhism.	A graveyard is a place to store corpses.
39	พืช	พงไพร	พืชคือสิ่งมีชิวิตสีเขียว	พงไพรคือป่าไม่ชนิดนึง
	Plant	Woods	A plant is a green living thing.	Woods are one kind of forest.

40	ป่าไม้	ภูเขา	ป่าไม้คืออาณาเขตซึ่งอุดมไปด้วยต้นไม้	ภูเขาคือลักษณะภูมิประเทศสูงขึ้นไปมาก
	Forest	Mountain	Forest is an area of abundant trees.	Mountain is very high-up terrain.
41	อาหาร	ผลไม้	อาหารคือสิ่งที่สิ่งมีชีวิตรับประทาน	ผลไม้คืออาหารที่ได้จากต้นไม้
	Food	Fruit	Food is what living creatures eat.	Fruit is food derived from a tree.
42	แก้ว	ถ้วย	แก้วคือภาชนะบรรจุของเหลว	ถ้วยคือภาชนะทรงโค้งหงายใช้เพื่อบรรจุของเหลว
	Glass	Cup	A glass is a container containing liquid.	A bowl is a rounded container containing liquid.
43	วัด	พระ	วัดคือที่สถานที่พักพิงทางศาสนาพุทธ	พระคือผู้ถือศีลอาศัยอยู่ในวัด
	Temple	Monk	A temple is a place for Buddhism.	A monk is a person living in a temple.
44	ลุง	ป้า	ลุงคือพี่ชายของบิดาหรือมารดา	ป้าคือพี่สาวของบิดาหรือมารดา
	Uncle	Aunt	An uncle is a brother of father or mother.	An aunt is a sister of father or mother.
45	ป่าไม้	ต้นไม้	ป่าไม้คืออาณาเขตซึ่งอุดมไปด้วยต้นไม้	ต้นไม้คือพืชคือสิ่งมีชิวิตสีเขียวชนิดนึง
	Forest	Tree	Forest is an area of abundant trees.	A tree is a type of plant.
46	โรงภาพยนต์	โรงละคร	โรงภาพยนตร์คือสถานที่จัดแสดงภาพยนตร์	โรงละครคือสถานที่จัดแสดงละคร
	Cinema	Theatre	A cinema is a place to show movies.	A theatre is a place for shows.
47	เนินเขา	ภูเขา	เนินเขาคือลักษณะภูมิประเทศสูงขึ้นไปเล็กน้อย	ภูเขาคือลักษณะภูมิประเทศสูงขึ้นไปมาก
	Hill	Mountain	Hill is a little high-up terrain.	Mountain is very high-up terrain.
48	เด็กผู้ชาย	เด็กหนุ่ม	เด็กผู้ชายคือมนุษย์วัยเยาว์เพศผู้	เด็กหนุ่มคือผู้ชายอายุน้อย
	Boy	Lad	A boy is a young male human.	A lad is a young man.
49	ผ้าฝ้าย	ผ้าไหม	ผ้าฝ้ายคือผ้าที่ผลิตมาจากดอกฝ้าย	ผ้าไหมคือผ้าที่ผลิตมาจากเส้นใยจากตัวไหม
	Cotton	Silk	Muslin is a cloth produced from cotton.	Silk is a cloth produced from silkworm fibres.
50	ยานพาหนะ	รถยนต์	ยานพาหนะคือเครื่องจักรใช้ในการเดินทาง	รถยนต์คือยานพาหนะสี่ล้อ
	Automobile	Car	A vehicle is a machine used for transportation.	A car is a four-wheeled vehicle.
51	ฝั่งทะเล	ชายฝั่ง	ฝั่งทะเลคือชายฝั่งที่ติดทะเล	ชายฝั่งคือชายทะเล
	Coast	Shore	A shore is a coast close to the sea.	A coast is a beach.
52	อุปกรณ์	เครื่องมือ	อุปกรณ์คือเครื่องมืออำนวยความสะดวก	เครื่องมือคือสิ่งที่ใช้เพื่อทุนแรง
	Implement	Tool	Equipment is a tool used as facilities.	Tool is used for a labour-saving device.
53	ทาส	ข้ารับใช้	ทาสคือข้ารับใช้	ข้ารับใช้คือผู้รับใช้

	Slave	Serf	A slave is a thrall.	A serf is servant.
54	การเดินทาง	การท่องเที่ยว	การเดินทางคือการเคลื่อนย้ายจากสถานที่หนึ่งไปยังที่หนึ่ง	การท่องเที่ยวคือการเดินทางในช่วงขณะหนึ่งเพื่อพักผ่อน
	Journey	Voyage	Journey is to travel from one place to another.	Journey is to travel at a certain time for leisure.
55	นิตยสาร	หนังสือ	นิตยสารคือหนังสือออกรายสัปดาห์หรือรายเดือน	หนังสือคือสิ่งบันทึกตัวอักษร
	Magazine	Book	A magazine is a book published weekly or monthly.	A book is a recorder of letters.
56	ลายมือชื่อ	ลายเซ็น	ลายมือชื่อคือสัญลักษณ์แทนเจ้าตัว	ลายเซ็นคือสัญลักษณ์แทนเจ้าตัว
	Autograph	Signature	An autograph symbolizes the person.	A signature symbolizes the person.
57	เที่ยงวัน	กลางวัน	เที่ยงวันคือเวลาสิบสองนาฬิกา	กลางวันคือระยะเวลาราว ๆ เที่ยงวัน
	Midday	Noon	Midday is the time at noon.	Noon is around twelve o’clock.
58	ป่าไม้	พงไพร	ป่าไม้คืออาณาเขตซึ่งอุดมไปด้วยต้นไม้	พงไพรคือป่าไม่ชนิดนึง
	Forest	Woods	Forest is an area of abundant trees.	Woods are one kind of forests.
59	อัญมณี	เพชรพลอย	อัญมณีคือแร่ธรรมชาติที่มีมูลค่า	เพชรพลอยคือเครื่องประดับที่มีมูลค่า
	Gem	Jewel	Gem is a natural mineral of value.	Jewel is accessories of value.
60	พืช	ต้นไม้	พืชคือสิ่งมีชิวิตสีเขียว	ต้นไม้คือพืชชนิดหนึ่ง
	Plant	Tree	Plant is a green living thing.	A tree is a type of plant.
61	นักบวช	พระ	นักบวชคือผู้ถือศีลทางศาสนาคริสต์	พระอาศัยอยู่ในวัด
	Priest	Monk	A priest is a person observing the precepts in Christianity.	A monk is a person living in a temple.
62	ป่าช้า	สุสาน	ป่าช้าคือสถานที่ฝังศพ	สุสานคือสถานที่เก็บศพ
	Cemetery	Graveyard	A cemetery is a place to bury corpses.	A graveyard is a place to store corpses.
63	วัด	โบสถ์	วัดคือที่สถานที่พักพิงทางศาสนาพุทธ	โบสถ์คือสถานที่ทางศาสนาคริสต์
	Temple	Church	A temple is a place for Buddhism.	A church is a Christian place.
64	ครู	อาจารย์	ครูคือผู้สอนในโรงเรียน	อาจารย์คือผู้สอนในมหาวิทยาลัย
	Teacher	Lecturer	A teacher is someone teaching at school.	A teacher is a person teaching at university
65	สุนัข	หมา	สุนัขคือหมาใช้ในภาษาทางการ	หมาคือสัตว์เลี้ยงลูกด้วยนมชนิดหนึ่ง
	Dog	Dog	A dog is a dog in official language.	Dog is one kind of mammal.

7.3 Methodology for Rating TWS-65

The aim of this section is to describe the methodology for rating the TSS-65 dataset that was collected in Section 7.2. This methodology is the same one used with TWS-65. Moreover, this section aims to present TSS-65.

7.3.1 Participants

Similarity ratings were collected from 40 native Thai speakers to complete the benchmark dataset. The participants had an equal number of Art/Humanities and Science/Engineering backgrounds. They consisted of 22 undergraduates and 18 postgraduates studying at 4 different Thai universities. The average age of the participants was 22 and standard deviation was 2.4, with 23 males and 17 females.

7.3.2 Materials

Following the previous practice of O'Shea et al. (2008), the representative subset of 65 sentence pairs chosen is shown in Table 7.49. Each sentence pair was printed on a separate card using a standard Thai font. A questionnaire (see Appendix 4.3) was produced containing instructions for recording similarity ratings and a small amount of personal data. Semantic anchors were also provided to guide the participants. The examples of experimental materials are:

- Appendix 1.5 The Person Data Collection Sheet
- Appendix 1.6 Semantic Anchors
- Appendix 4.1 The Ethics Statement
- Appendix 4.2 The Instruction Sheet
- Appendix 4.3 A Sample Card
- Appendix 4.4 Sample Rating Recording Sheet.

7.3.3 Procedure

The participants were asked to perform the following procedure:

1. Please sort the cards into four groups in a rough order of the similarity of meaning of the sentence pair.
2. After sorting the cards into groups, order the cards in each group according to similarity of meaning (i.e. the card that contains the lowest similarity of meaning is at the top of the group).

3. Please recheck the cards in every group. You may change a pair of sentences to other groups at this stage.
4. Please rate the semantic similarity rating of each pair of sentences by writing a number between 0.0 (minimum similarity) and 0.9 for first group; 1.0 and 1.9 for second group; 2.0 to 2.9 for third group; and 3.0 and 4.0 (maximum similarity) for fourth group on the recording sheet. You can use the first decimal place (for example, 2.5) to show finer degrees of similarity. You also may assign the same value to more than one pair.

The cards were shuffled into a random order before being given to the participants.

7.3.4 TSS-65

The average Human rating for TSS-65 is shown in Table 7.50. These are the rating for the sentence pairs in Table 7.49. Column *SP* is the number of the sentence pairs in Table 7.49. Column *Human* is the average of similarity rating from 40 native Thai speakers. Column *SD* is the standard deviation.

Table 7.50: TSS-65 Sentence Pairs with Human Rating

SP	*Human*	*SD*	*SP*	*Human*	*SD*
1	0.655	0.650	**34**	0.525	0.597
2	0.078	0.097	**35**	1.678	1.034
3	0.090	0.209	**36**	0.525	0.723
4	0.623	0.665	**37**	2.923	0.792
5	0.863	0.848	**38**	2.645	1.098
6	0.385	0.490	**39**	2.023	0.949
7	0.318	0.437	**40**	1.430	0.935
8	0.463	0.411	**41**	1.828	1.115
9	0.155	0.308	**42**	3.820	0.228
10	1.068	0.782	**43**	2.145	0.990
11	0.425	0.535	**44**	3.425	0.532
12	0.225	0.484	**45**	3.270	0.583
13	1.290	0.967	**46**	3.248	0.825
14	0.648	0.646	**47**	2.950	0.763
15	0.468	0.578	**48**	3.535	0.515
16	0.288	0.376	**49**	3.215	0.640
17	0.733	0.721	**50**	2.778	0.718
18	0.835	0.731	**51**	3.178	0.734
19	0.795	0.777	**52**	3.323	0.599
20	1.223	0.998	**53**	3.818	0.192
21	0.975	0.886	**54**	2.833	1.046
22	0.455	0.538	**55**	2.588	0.548
23	1.618	1.146	**56**	3.768	0.420
24	0.433	0.458	**57**	3.185	0.847

25	1.403	1.009	**58**	3.073	0.724
26	1.353	1.125	**59**	2.723	0.941
27	0.543	1.021	**60**	3.003	0.798
28	1.360	0.884	**61**	2.678	0.979
29	2.170	0.949	**62**	3.553	0.608
30	1.975	1.111	**63**	3.210	0.755
31	2.325	1.060	**64**	3.545	0.508
32	1.123	0.904	**65**	2.755	1.273
33	2.410	0.918			

7.4 Discussion of the TSS-65

The fundamental conjecture was that if two words have a particular degree of word similarity, their definitions ought to have a consistent degree of sentence similarity. The Pearson Product-Moment correlation coefficients are adopted to demonstrate consistency of the word and sentence pair similarities over the two datasets.

In calculating these for the 65 pairs of similarity ratings (words vs. sentences), the results are:

- Pearson's $r = 0.896$ (P-Value < 0.01)

For *r*, a value of +1 indicates perfect correlation, 0 indicates no relationship and -1 indicates a perfect negative correlation. P-values indicate the likelihood of obtaining the result by chance.

The ANOVA test was used to find whether or not the Human rating of TWS-65 and TSS-65 were statistically significantly different ($\alpha=0.05$) from the hypotheses:

- H_0: There is no statistically significant difference between two datasets.
- H_1: There is a statistically significant difference between two datasets.

The result is:

- $f = 0.174$, $df = 1$ (P-Value > 0.05)

With its result, it fails to reject the null hypothesis, which means the human ratings procedure in TWS-65 and TSS-65 are not statistically significantly different.

Table 7.51: Correlation Coefficients with Mean Human Judgment

	Correlation *r*
Average of the correlation of all participants	0.840
Best participant	0.902
Worst Participant	0.752

Table 7.51 shows the Pearson Product-Moment correlation coefficients of 40 participants; the leave-one-out resampling technique is used to find the correlation coefficient of each participant with the rest of the group.

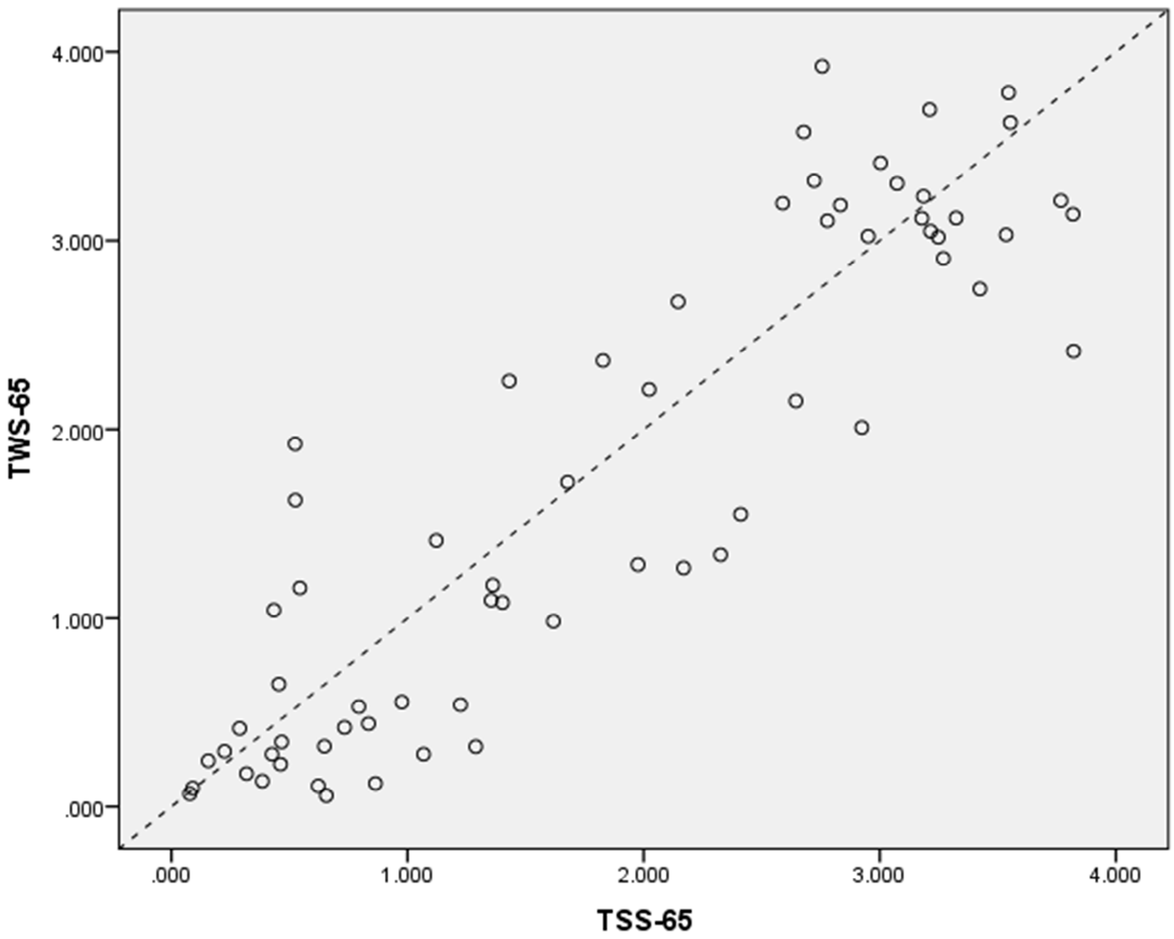


Figure 7.1: The Correlation between TWS-65 and TSS-65

Figure 7.1 shows the data point between TWS-65 and TSS-65. The most outliers from Figure 7.1 are obtained from sentence pairs the pair 42 and 65.

Table 7.52: The Odd Pair

Pair	W_1	W_2	S_1	S_2
42	แก้ว	ถ้วย	แก้วคือภาชนะบรรจุของเหลว	ถ้วยคือภาชนะทรงโค้งหงายใช้เพื่อบรรจุของเหลว
	Glass	Cup	A glass is a container containing liquid.	A bowl is a rounded container containing liquid.
65	สุนัข	หมา	สุนัขคือหมาใช้ในภาษาทางการ	หมาคือสัตว์เลี้ยงลูกด้วยนมชนิดหนึ่ง
	Dog	Dog	A dog is a dog in official language.	Dog is one kind of mammal.

Table 7.52 shows the word pairs 42 and 65 and sentence pairs 42 and 65; the Human rating for the word pair 65 (Dog-Dog) is 3.923, which is the highest rating that was obtained in TWS-65. Basically, the word *สุนัข* (Dog) and the word *หมา* (Dog) mean the same, which is 'dog'. However, the Human rating for the sentence pair 65 is 2.756. This is because the definitions disambiguate two different word senses for sentence pair 65. The word *สุนัข* (Dog) is normally used formally, which its definition "สุนัขคือหมาใช้ในภาษาทางการ" explains well in English. On the other hand, the word *หมา* (Dog), whose definition is 'หมาคือสัตว์เลี้ยงลูกด้วยนมชนิดหนึ่ง', means 'Dog is one kind of mammal'. Another odd pair is pair 42 (Glass-Cup); the human rating for this word pair is 2.413, while the human rating for this sentence pair is 3.821. There are two reasons why the human rating for this sentence pair is significantly higher than the word pair. Firstly, in Thai, the word *แก้ว* (Glass) is normally used to describe 'glass' but in some cases, it can also be used to describe 'crystal'. This explains why the human rating for word pair is not high, as the humans make a subjective judgment regarding which word sense to use. Secondly, the human rating of the sentence pair is very high because the definition of the word *แก้ว* (Glass) used in the sentence is 'แก้วคือภาชนะบรรจุของเหลว' means 'A glass is a container containing liquid'. Therefore, it cannot refer to 'crystal'. Moreover, both definitions of the sentence pair describe *แก้ว* (Glass) and *ถ้วย* (Cup) as used to contain liquid. Therefore, it is reasonable to say that native Thai speakers see more similarity in the sentence pairs than in the word pairs.

7.5 Conclusion

This chapter presented the method of selecting definitions to create TSS-65. The sentence pairs in TSS-65 correspond with the TWS-65. The methodology to rate TSS-65 was also described in this chapter. This chapter presened the first Thai sentence benchmark dataset that can be used to evaluate Thai sentence similarity measures. This paves the way for the development of the new Thai sentence similarity measure, the subject of Chapter 8.

Chapter 8

Thai Short Text Semantic Similarity Measures (TSTS)

8.1 Introduction

Chapter 1 set out the aim of this thesis to propose a Thai sentence semantic similarity measure (TSTS). A validated Thai word similarity measure and a Thai sentence semantic similarity benchmark dataset are needed to produce the Thai sentence similarity measure. The Thai word similarity measure (nTWSS) and Thai sentence dataset (TSS-65) were presented in Chapters 6 and 7, respectively. nTWSS will be used as a component of TSTS and TSS-65 will be used to evaluate TSTS. The aim of this chapter is to investigate the research question: Can a Thai word measure be used to develop a Thai sentence similarity measure?

The contributions in this chapter are:

- Creation of TSTS
- Evaluation of TSTS with TSS-65
- An illustration of the use of TSTS with representative dialogue utterances for a future Thai Conversational Agents.

The rest of this chapter is organized as follows: Section 8.2 sets out the design and implementation of the Thai sentence semantic similarity measure works; Section 8.3 discusses human and machine sentence similarity ratings; Section 8.4 illustrates the use of TSTS for a future Thai Conversational Agents; and Section 8.5 concludes.

8.2 A Thai Sentence Semantic Similarity Measure (TSTS)

As mentioned in Chapter 2 Section 2.3.4, Semantic Similarity based on Semantic Nets and Corpus Statistics, or ‘STASIS’ (Li et al, 2006), is chosen to be used as a prototype to develop TSTS. STASIS uses three elements for the determination of sentence similarity, which are word similarity, statistical information (such as word frequency), and word order similarity. However, as TSTS is a measure based on the Thai language, word order similarity will not be used. In Thai, there are a number of cases whereby the order of the words in a sentence can be changed whist retaining the meaning, as described in detail in Section 2.5.2. Therefore, there are two components that will be used to develop TSTS: word similarity and statistical information. The word similarity measure (nTWSS) used to calculate the word similarity and the word frequency are referred from the Thai National Corpus (Aroonmanakun, 2007).

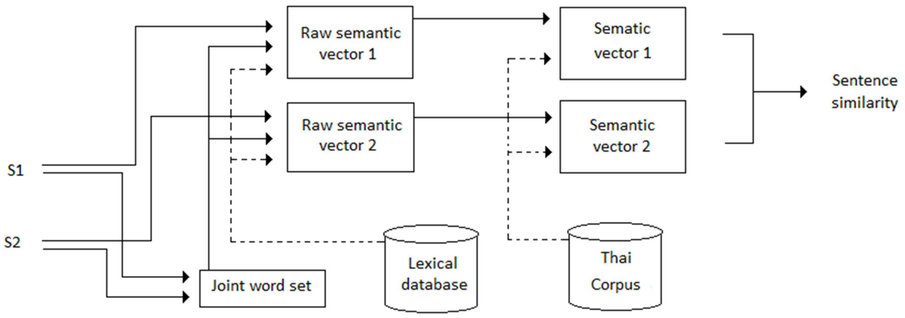


Figure 8.1: An Overview of TSTS

Figure 8.1 shows an overview of TSTS. This algorithm can be separated into three steps, as follows:

- Construction of the Joint Word Set
- Formation of the Lexical Semantic Vector
- Calculation of the Sentence Similarity.

This similarity measure is implemented for the Thai language; however, for more clarity, an English example is used to illustrate the algorithm.

8.2.1 Construction of the Joint Word Set

Equation 8.1 describes a joint word set T derived from all the unique words in two sentences: T_1 and T_2.

$$T = T_1 \cup T_2 = \{w_1, w_2, ..., w_m\} \qquad \textbf{Equation 8.1}$$

Given two sentences T_1 and $T_{2,}$ a joint word set is formed using Equation 8.1:

T_1: The lion is the king of the jungle

T_2: Lion is a mammal

A joint word set, T is

T = {The lion is the king of the jungle a mammal}

8.2.2 Formation of the Lexical Semantic Vectors

The lexical semantic vector for each sentence, denoted by $\check{s}$, is derived from the joint word set. The m equals the number of words in the joint word set. Each entry, $\check{s}_i$ (where i=1, 2, ..., m) is determined by the semantic similarity of the corresponding word in the joint word set to a word in the sentence.

For each word in the joint set, there are two possible outcomes when the joint set is scanned:

- Case 1: $\check{s}_i$ is set to 1, if w_i appears in the sentence,
- Case 2: If w_i is not contained in T_1, a semantic similarity score is computed between w_i and each word in the short text T_1, using the nTWSS word measure. The most similar word in T_1 to w_i is that with the highest similarity score. If the highest score exceeds a preset threshold, then $\check{s}_i$ is equal to the highest score; if not, $\check{s}_i$ is 0.

The threshold is used because it is assumed that the values below the threshold are merely contributing noise (Li et al., 2006). It is set as 0.2; this value is the same as an original value in STASIS (Li et al, 2006). The choice was made because nTWSS uses WordNet as a component, in common with STASIS. The value could be optimized for Thai when larger machine learning Thai word similarity datasets become available.

Equation 8.2 shows how the words are weighted according to their information content (Resnik, 1999), on the assumption that word frequency influences the contribution of the individual words to the overall similarity. Entropy measures are calculated using the Thai National Corpus (Aroonmanakun, 2007):

$$s_i = \check{s} \times I(w_i) \times I(w_j) \qquad \textbf{Equation 8.2}$$

Given that w_i is a word in the joint word set, and w_j is its associated word in the sentence. $I(w_i)$ is the information content of w_i in the corpus. The value of $I(w_i)$ can be [0,1] and defined as:

$$I(w_i) = \frac{-\log(p(w_i))}{\log(N+1)} \qquad \textbf{Equation 8.3}$$

where $p(w_i)$ is the probability of a word w_i, and N is the total number of words in the corpus. $p(w_i)$ can be calculated as:

$$p(w_i) = \frac{n+1}{N+1} \qquad \textbf{Equation 8.4}$$

where n is the word frequency of the word w in the corpus.

8.2.3 Calculation of the Sentence Similarity

Lastly, the semantic similarity between T_1 and T_2, $s(T_1,T_2)$, is calculated using a cosine similarity measure between two vectors, as shown in Equation 8.5:

$$s(T_1, T_2) = \frac{s_1 \times s_2}{\|s_1\| \times \|s_2\|} \qquad \textbf{Equation 8.5}$$

8.3 Evaluation of the Thai Short Text Semantic Similarity Measure

Semantic similarity is a product of human perception, grounded in consciousness. Therefore, the only way to evaluate the TSTS algorithm is against a dataset of Thai sentence pairs with human similarity ratings. The accepted measure of agreement between human and machine rating is the Pearson Product-Moment correlation coefficients (*r*). The aim of this section is to describe a series of experiments that were conducted using TSS-65 from Chapter 7 to evaluate the TSTS measure described in Section 8.2.

8.3.1 Methodology

To evaluate the TSTS measure, the Thai sentence benchmark dataset is required. There is only one Thai sentence benchmark dataset that is available, which TSS-65 from Chapter 7. The TSS-65 dataset is used to evaluate. The TSTS rating can be obtained by calculating the sentence pairs from the dataset, as explained in Section 8.2. The Pearson Product-Moment correlation coefficients (*r*) between the Thai human rating and TWSS will be calculated and shown in Section 8.3.3.

8.3.2 Results

Table 8.1 shows the semantic similarity ratings for the translated word pairs. Column *SP* is the number of the sentence pair, as shown in Table 8.1. Column *Human* is the human rating for the Thai sentence pairs. Column *STASIS* is the STASIS rating for the Thai sentence pairs translated in to English by Google translation (Och, 2005). Column *TSTS* is the machine rating for the Thai sentence pairs using the algorithm described in Section 8.2. All the measures have been scaled in the range 0 to 1 to aid comparison.

Table 8.1: Semantic Similarity of Human Rating, STASIS, and TSTS

SP	*Human*	*STASIS*	*TSTS*	*SP*	*Human*	*STASIS*	*TSTS*
1	0.164	0.311	0.425	**34**	0.131	0.462	0.419
2	0.019	0.525	0.346	**35**	0.419	0.613	0.565
3	0.023	0.671	0.624	**36**	0.131	0.665	0.326
4	0.156	0.531	0.537	**37**	0.731	0.904	0.702
5	0.216	0.669	0.538	**38**	0.661	0.662	0.589
6	0.096	0.445	0.415	**39**	0.506	0.339	0.635
7	0.079	0.384	0.387	**40**	0.358	0.326	0.535
8	0.116	0.218	0.444	**41**	0.457	0.644	0.525
9	0.039	0.470	0.620	**42**	0.955	0.825	0.947
10	0.267	0.480	0.409	**43**	0.536	0.424	0.757
11	0.106	0.369	0.250	**44**	0.856	0.877	0.946
12	0.056	0.196	0.653	**45**	0.818	0.600	0.858
13	0.323	0.548	0.385	**46**	0.812	0.805	0.876
14	0.162	0.646	0.364	**47**	0.738	0.675	0.921
15	0.117	0.327	0.225	**48**	0.884	0.920	0.895
16	0.072	0.334	0.228	**49**	0.804	0.779	0.724
17	0.183	0.389	0.547	**50**	0.694	0.860	0.707
18	0.209	0.571	0.561	**51**	0.794	0.630	0.715
19	0.199	0.793	0.685	**52**	0.831	0.331	0.748
20	0.306	0.376	0.485	**53**	0.954	0.564	0.912
21	0.244	0.467	0.474	**54**	0.708	0.589	0.821
22	0.114	0.598	0.552	**55**	0.647	0.668	0.620
23	0.404	0.404	0.676	**56**	0.942	0.457	0.943
24	0.108	0.549	0.682	**57**	0.796	0.562	0.624
25	0.351	0.429	0.643	**58**	0.768	0.638	0.845
26	0.338	0.566	0.579	**59**	0.681	0.791	0.785
27	0.136	0.603	0.562	**60**	0.751	0.619	0.624
28	0.340	0.472	0.577	**61**	0.669	0.346	0.575
29	0.543	0.684	0.736	**62**	0.888	0.559	0.867
30	0.494	0.705	0.624	**63**	0.803	0.622	0.777
31	0.581	0.831	0.672	**64**	0.886	0.788	0.944
32	0.281	0.392	0.574	**65**	0.689	0.673	0.779
33	0.603	0.747	0.564				

8.3.3 Discussion

Figure 8.2 is a scatter plot plotting the rating for the Thai sentence pairs calculated by TSTS against their corresponding Thai human ratings. The closer the points to the line of best fit, the better the correlation.

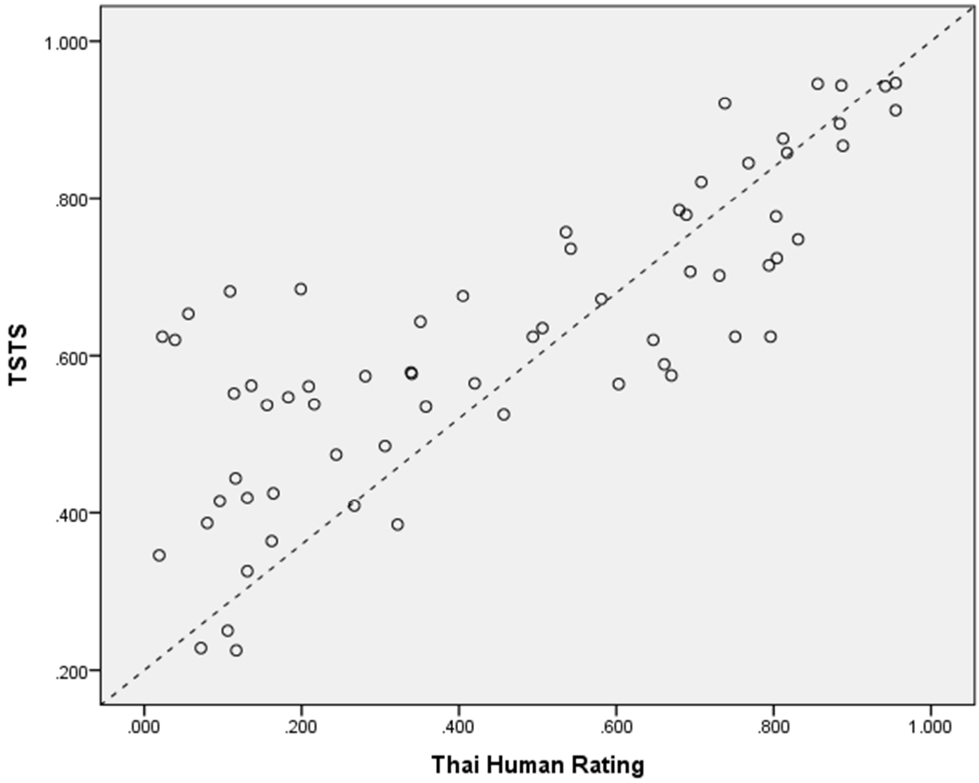


Figure 8.2: The Correlation between TSS-65 and TSTS

The Pearson Product-Moment correlations obtained from these results were:

- Pearson's r = 0.809 (P-Value < 0.01)

Table 8.2 illustrates the agreement of machine measure with human ratings by calculating the Pearson Product-Moment correlations coefficient (r) between the human ratings and the TSTS over the TSS-65. Also, the correlation coefficient of each participant with rest of the group over TSS-65 from Table 7.50 in Section 7.4 is presented for comparison.

Table 8.2: Correlation Coefficients

	Correlation *r*	**P-Value**
TSS-65 and TSTS	0.809	< 0.01
TSS-65 and STASIS	0.510	< 0.01
Average of the correlation of all participants	0.840	-
Worst Thai native speaker participant and the least of the group	0.752	-
Best Thai native speaker participant and the least of the group	0.902	-

The Thai sentence measure performs better than the correlation between the worst performing human and the least of the group (r = 0.752). This supports the view that it could form the basis of an effective sentence semantic similarity measure for two Thai sentences.

As the TSTS is the first sentence similarity measure for two Thai sentences, there is no other Thai sentence similarly measure for comparison. However, omparison with an English measure might give an idea how well TSTS performs. The STASIS ratings shown in Table 8.1 are obtained by calculating the rating with all the sentence pairs in TSS-65 and those sentence pairs are translated from Thai into English by Google Translate. A correlations coefficient of 0.510 (P-value < 0.01) was obtained, which was markedly below the TSTS value when compared with TSS-65.

Steiger's z-test (Steiger, 1980) was applied to find whether or not TSTS and STASIS were statistically significantly different (α=0.05) from the hypotheses:

- H_0: There is no statistically significant difference between two measures.
- H_1: There is a statistically significant difference between two measures.

To calculate Steiger's z-test between two measures requires the construction of a correlation triangle. In this case, we considered comparing the correlation between the TSTS and TSS-65 human rating with the correlation between STASIS and the TSS-65 human rating. The specific triangle for this calculation is formed according to Figure 8.3.

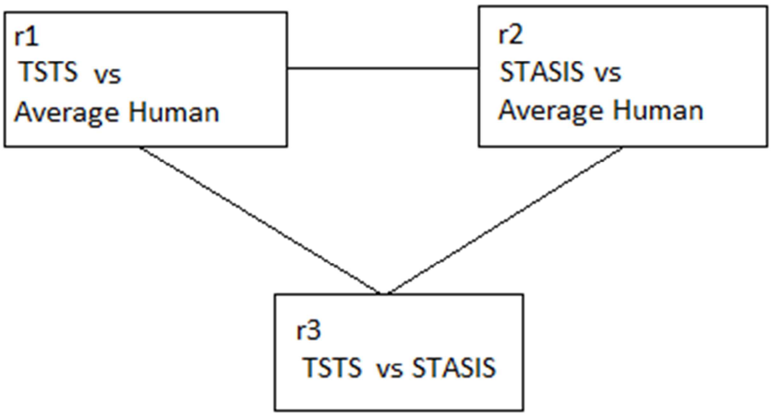


Figure 8.3: Specific Correlation Triangles for TWSS vs nTWSS

From Table 8.3:

r1 rxy TSTS vs Average human	0.809
r2 rzy STASIS vs Average human	0.510

n = 65

Calculate correlation:

r3 rxz TSTS vs STASIS	0.511

Applying the test gives the following results:

- z = 3.695, df = 62 (P-Value < 0.01)

From this result, the null hypothesis is rejected; this means TSTS and STASIS are statistically significantly different. This means TSTS (r = 0.809) performs significantly better than STASIS (r = 0.510) with the TSS-65 dataset.

One of the reasons that STASIS has not behaved as well as might be expected is the “sentence” translation (Thai to English). The Thai and English language structures are mostly different, as mentioned in Chapter 2. According to Aiken (2011), the Google sentences translations between European languages are usually good, whilst the Asian languages are often relatively poor. The Thai-English sentence translation does not correspond to human expectation (Kritsuthikul, 2006). For instance, for the sentence ‘หมาคือสัตว์เลี้ยงลูกด้วยนมชนิดหนึ่ง’, literally “Dog is one kind of mammal” in sentence pairs 19, 21 and 65, the translation from Google was ‘Puppies are okapi’. The meanings before translation and after translation are clearly different. However, after cutting the word ‘ชนิดหนึ่ง’, meaning ‘a kind of’, the result from Google Translate is ‘Dog is a mammal’, which is the expected meaning. This example shows one of the problems with the sentence translation.

Nevertheless, TSS-65 is also translated by a native Thai speaker who has a BA in Language and Culture from Chulalongkorn University, one of the best universities in Thailand. Those translated sentence pairs are given the similarity rating by STASIS, a correlations coefficient of 0.754 (P-value < 0.01), which was an improvement, but still below the TSTS value (r = 0.809, P-value < 0.01). This also supports the view that TSTS could be a basis for an effective sentence semantic similarity measure for two Thai sentences.

8.4 The Evaluation of TSTS usage with Conversational Agent Log Files

The aim of the evaluation is to find how well TSTS performs with Thai sentence pairs. This following section will provide detail of the experimental methodology, results and discussion.

8.4.1 Methodology

To achieve the aim of the evaluation of the TSTS, 15 specific sentence pairs are selected. The 15 sentence pairs can be separated into three groups: High similarity group, Medium similarity group, and Low similarity group. These 15 sentence pairs are chosen from English Conversational Agent debt adviser log files. These log files come from a real-life system. The semantic similarity rating calculated from TSTS is also separated into three groups as follows:

- High similarity group (rating between 0.750-1.00)
- Medium similarity group (rating between 0.25-0.749)
- Low similarity group (rating between 0.00-0.249).

The Medium similarity group has a bigger range because the Medium similarity group contains both Medium-High similarity and Medium-Low similarity groups.

The results are shown in Section 8.4.3.

8.4.2 Materials

The chosen 15 sentence pairs are shown in Table 8.3. These 15 sentence pairs are translated into Thai by a native Thai speaker who has a BA in Language and Culture. Column *SP* is the sentence pair number. Column *Prediction* is the prediction similarity group of sentence pairs (based on my personal judgement).

Table 8.3: The Chosen 15 Sentence Pairs

SP	***S_1***	***S_2***	***Prediction***
1	ฉันสับสน	ฉันยังสับสนอยู่	High
	I am confused	I am still confused	
2	ฉันไม่เข้าใจ	คุณไม่เข้าใจ	High
	I cannot understand this	You do not understand	
3	ฉันเป็นคนติดการพนัน	ฉันเป็นคนติดพนัน	High
	I am a gambling addicy	I have a gambling addiction	
4	ฉันต้องการเงิน	ฉันต้องการเงินสด	High
	I want money	I need cash	
5	ก็ได้ฉันจะจ่ายสามส่วน	ฉันจ่ายได้แค่สามส่วน	High
	All right I can pay third	I can only pay third	
6	ฉันยังต้องจ่ายหนี้อยู่ไหม	ฉันไม่มีปัญหาหนี้สิน	Medium
	Do I still have to pay my debt	I do not have a debt problem	
7	ฉันสับสนกับตัวเลือกพวกนี้ไปหมดแล้ว	ฉันรู้สึกว่าเหตุผลพวกนี้สับสน	Medium
	I am confused by all these choices	I find all the reasons confusing	
8	ไม่สามารถจ่ายเงินได้	เงินค่าที่อยู่อาศัย	Medium
	Cannot afford payment	My accommodation payment	

9	นิยามการวัดผลให้ฉันที	ฉันยังไม่ได้รับการวัดผลเลย	Medium
	Define assessment for me	I did not get my assessment yet	
10	ฉันยังไม่ได้เงิน	ฉันต้องการเงิน	Medium
	I did not get the money	I want money	
11	ฉันไม่อยากบอกชื่อกับคุณ	ฉันยื่นเรื่องขอรับเงินไป	Low
	I do not want to tell you my name	I applied for the money	
12	ฉันซื้อรถมา	ไม่สามารถจ่ายเงินได้	Low
	I bought a car	Cannot afford payment	
13	ฉันต้องจ่ายที่ไหน	ฉันเป็นคนติดพนัน	Low
	Where do I pay	I have a gambling addiction	
14	ฉันต้องการความช่วยเหลือ	ฉันไม่แน่ใจว่าควรทำอย่างไร	Low
	I need your help	I am not sure what I should do	
15	ก็ได้ฉันจะจ่ายสามส่วน	คุณไม่เข้าใจ	Low
	All right I can pay third	You do not understand	

8.4.3 Results

The experiment result is shown in Table 8.4. Column *SP* is the sentence pair number. Column *TSTS* is the machine rating for the Thai sentence pairs using the algorithm described in Section 8.2. Column *Group* is the similarity group of TSTS rating. Column *Prediction* is the prediction similarity group of sentence pair from Table 8.3. Column *Result* is the result of the prediction for each sentence pair.

Table 8.4: The Results of the Chosen 15 Sentence Pairs

SP	***TSTS***	***Group***	***Prediction***	***Result***
1	0.922	High	High	Correct
2	0.894	High	High	Correct
3	0.908	High	High	Correct
4	0.953	High	High	Correct
5	0.843	High	High	Correct
6	0.527	Medium	Medium	Correct
7	0.612	Medium	Medium	Correct
8	0.224	Low	Medium	**Wrong**
9	0.467	Medium	Medium	Correct
10	0.726	Medium	Medium	Correct
11	0.186	Low	Low	Correct
12	0.309	Medium	Low	**Wrong**
13	0.226	Low	Low	Correct
14	0.324	Medium	Low	**Wrong**
15	0.178	Low	Low	Correct

8.4.4 Discussion

According to the results shown in Table 8.4, TSTS predicts the High similarity group correctly. For the Medium similarity Group, TSTS predicts 4 from 5 Medium sentence pairs accurately, while the Low similarity Group TSTS predicts 3 from 5 Low sentence pairs accurately. Thus, TSTS predicts 12 from 15 sentence pairs accurately; i.e. 80%. The Spearman rank correlation between human prediction and TSTS prediction is:

- Spearman’s ρ = 0.864 (P-Value < 0.01)

One of the wrongly predicted sentence pairs from Table 8.4 is sentence pair 8; this sentence pair is meant to be in the Medium similarity group. However, TSTS produced the rating for sentence pair 8 as 0.224, which is in the Low similarity group. This happens because after translating into Thai, the word ‘payment’ in the sentence ‘Cannot afford payment’ is translated into word ‘เงิน’ (Money). Therefore, the word ‘payment’ in the sentence ‘My accommodation payment’ is translated into ‘เงินค่า’ (Payment). This makes the meanings in Thai and English more different because the word ‘เงินค่า’ (Payment) in Thai can only be used in some specific content, while the word ‘เงิน’ (Money) is generally used .

A Conversational Agent has the capacity to go back, correct and disambiguate use of any incorrect sense misunderstandings through dialogue, which is why 80% accuracy is approaching workabl,e whereas it would be too low for applications such as information retrieval. Therefore, this supports the view that TSTS could be used to create the Thai semantic-based Conversational Agents.

8.5 Conclusion

The aim of the research is to propose a Thai sentence semantic similarity measure (TSTS). This chapter described how the first sentence semantic similarity measures works and also discussed the experiment result and how correlations coefficient of 0.809 (P-Value < 0.01) was obtained. This measure performs better than STASIS when used with the Thai sentence pairs and should be useful in Thai semantic similarity. TSTS is considered to be a starting point for a Thai sentence measure which can be used to create semantic-based Conversational Agents in future. However, there are a number of aspects concerning the algorithm which can still be improved, discussed in Chapter 9.

Chapter 9

Conclusion and Future Work

9.1 Introduction

This chapter summarises the work and contributions in relation to the research aims and objectives of this thesis. The contributions of the thesis are also concluded. Finally, recommendations for the direction of future research are given.

9.2 Summary of the Work

This research has proposed three Thai word similarity measures, two Thai word benchmark datasets, one Thai sentence similarity measure, and one Thai sentence benchmark dataset. These are the outcomes of the work to answer the research questions as follows:

- Can a semantic-based Conversational Agent be developed in Thai?

Chapter 2 established that this research question cannot be given an immediate answer ‘YES’. The Thai language simply does not yet have the resources to support this. Therefore, the main focus of this work is to create a suitable framework to support future work in the development of Conversational Agents. This chapter provided a background to this thesis that introduced related research, including English word similarity measures, non-English similarity measures, English sentence similarity measures, non-English sentence similarity measures, the fundamentals of the Thai language and the current state of research in Thai WordNet. Also, the potential for a Thai similarity measure was reviewed and discussed. This found no research about Thai similarity measures. Therefore, as a starting point to develop a new Thai similarity measure, the STASIS architecture was selected.

- Can an English word similarity measure be developed for the Thai language by translating Thai words into English?

Chapter 3 proposed the first Thai word measure (TWSS), which was developed directly from Li’s measure (Li et al., 2003). This work answers this research question. TWSS was created based on the conversion of Thai words to English for Li’s measure to be applied. Moreover, a 30 Thai word pair benchmark dataset (TWS-30) was also presented in this chapter. In an evaluation of TWSS with TWS-30, a correlation coefficient of 0.823 (P-value < 0.01) was obtained, providing supporting evidence for the research question. This result was promising. However, this measure could not be used to fully predict those word pairs that relate to Thai culture as TWS-30 was built based on an English dataset

(Rubenstein and Goodenough, 1965). Thus, to experiment on words relating to the Thai culture, a more effective evaluation is needed.

- Can a WordNet based English word similarity measure produce a similarity rating between words based on Thai culture?

Chapter 4 presents the methodology for the creation of a 65 Thai word pairs benchmark dataset based on Thai culture (TWS-65) which addresses this research question. The evaluation of a subset of TWS-30 and TWS-65 human ratings has shown that both datasets are not significantly different. In addition, a correlation coefficient of 0.807 was obtained between TWS-65 human ratings and TWSS ratings. TWSS uses the English-based WordNet to perform the rating. Hence, it results in an inefficient rating performance of these word pairs which are related mainly to the Thai culture. Therefore, the limitations of TWSS mean it should be considered as a pathway to a final Thai word similarity measure.

- Can a search engine provide an alternative natural language resource for a Thai word similarity measure?

Chapter 5 presented the investigations undertaken in considering this research question. This chapter proposed a word similarity measure based on a lexical chain that was created from a mini corpus produced by a search engine (LCSS). The aim of this algorithm is to overcome the problem with TWSS. A training dataset (TWS-30) and a testing dataset (TWS-51) were also presented. The training dataset was used to find the most suitable *Alpha* parameter in LCSS. The testing dataset was used to evaluate the LCSS algorithm. A correlation coefficient of 0.723 (P-value < 0.01) was obtained. Both the TWSS and LCSS perform quite well on their own. However, evidence shows that each contributes a different insight into the similarity process. Therefore, a combination of TWSS and LCSS may be more effective.

- Can a combination of TWSS and LCSS provide a better model of human perception of Thai word semantic similarity than either separately?

Chapter 6 proposed a word measure that was created from a combination of TWSS and LCSS, called nTWSS, to addresses this research question. The correlation coefficient between nTWSS ratings and TWS-51 human ratings was r = 0.867 (P-Value < 0.01), a significant improvement on TWSS or LCSS alone. Accordingly, nTWSS can be used to develop a Thai sentence similarity measure.

- Can a Thai word measure be used to develop a Thai sentence similarity measure?

Chapter 7 and Chapter 8 presented the investigations undertaken in considering the research question

Chapter 7 presents the first Thai sentence benchmark dataset (TSS-65). Following O'Shea's procedure (O'Shea, 2008), TSS-65 was created by replacing the words with a definition, shown in Section 7.2.1. Comparing TSS-65 and TWS-65, a correlation coefficient of 0.896 (P-value < 0.01) was obtained indicating general consistency between derived sentences and the word. This paved the way for the development of the new Thai sentence similarity measure.

Chapter 8 proposed the first Thai sentence semantic similarity measure (TSTS). The sentence measure is developed from STASIS; nTWSS is used to calculate the semantic similarity between two words. Word order is not taken into account. In this measure, the correlation coefficient between TSS-65 and TSTS was $r = 0.809$ (P-value < 0.01).

- Is the developed Thai sentence similarity measure feasible for use in developing Thai Conversational Agents

In Chapter 8, an experiment was conducted to answer the research question: 'Is the developed Thai sentence measure feasible to use to develop Thai Conversational Agents?' In this simple experiment, TSTS predicted the categories low, medium, and high similarity with 80% accuracy between sentence pairs from a Conversational Agent log file. Furthermore, the crucial function of STSS in a Conversational Agent is to find rules that capture attributes accurately; therefore, the higher performance of TSTS in this circumstance is important. Medium and Low similarity matches against rules normally leading to disambiguation of user meaning or the firing of off-topic 'chat' rules. Because Conversational Agents inherently do multiple interactions to disambiguate misunderstandings, this is approaching a usable performance. Therefore, this supports the view that TSTS could be used to create Thai semantic-based Conversational Agents.

9.3 Summary of Contribution

The contributions in this thesis are:

- Review and discussion of Thai natural language resources
- Creation of the first Thai word semantic similarity measure (TWSS)
- Methodology for creating the first Thai word semantic similarity benchmark dataset (TWS-30)

- Application of the methodology to rating TWS-30
- Evaluation of TWSS with TWS-30
- Creation of a 65 Thai word pairs benchmark dataset (TWS-65)
- Application of the methodology to rating TWS-65
- Evaluation of TWS-30 with TWS-65
- Evaluation of TWSS with TWS-65
- Creation of a word similarity measure based on a lexical chain created from a search engine (LCSS)
- Creation of a testing dataset (TWS-51)
- Evaluation of LCSS with TWS-51
- Creation of a new word measure specifically for the Thai language (nTWSS)
- Evaluation of nTWSS with TWS-51
- Creation of a 65 Thai sentence pairs benchmark dataset (TSS-65)
- The application of the methodology to rating TSS-65
- Evaluation of TWS-65 with TSS-65
- Creation of the first Thai sentence similarity measure (TSTS)
- Evaluation of TSTS with TSS-65
- An illustration of the use of TSTS with representative dialogue utterances for a future Thai Conversational Agents.

These contributions are expected to provide a substantial starting point for research in the new fields of Thai word semantic similarity, Thai semantic sentence similarity, and Thai Conversational Agents.

9.4 Future Work

There are a number of potential directions in which this research could be continued in the future. These directions are as follows:

- Creation of Thai semantic-based Conversational Agents that use the TSTS algorithm.
- For nTWSS, non-linear approaches may be more appropriate when combining two different measures of similarity; this was suggested by O’Shea (O’Shea et al., 2008). As nTWSS was created by a linear combination of TWSS and LCSS using the δ parameter, replacing this with an artificial neural network trained to combine

the two components should improve the overall performance. However, a larger Thai word dataset is needed to do this.

- In LCSS, a Thai sentence extraction algorithm (Sornlertlamvanich, 2000) is used as mentioned in Section 5.2.4. This algorithm extracts Thai words from a Thai sentence with the accuracy of 85%. Unfortunately, this is the most precise algorithm available at that time. If the accuracy of the algorithm can be improved, the nTWSS is also likely to be improved.
- Machine Learning can be applied to predicting the next word in a lexical chain in the LCSS algorithm. The current algorithm is programmed to find any lexical chain that is available from the lexical database. This process take a long time, causing the algorithm to perform slowly. The next word in a Lexical chain can be predicted by Machine Learning, which will save time to process. Also, the best Lexical chain may be chosen without a calculation.
- According to O'Shea et al. (2013), STSS-131 used the best practice established from STSS-65 to rate a more representative set of English sentences. This can also be done with TSS-65. A more representative set of Thai sentences should provide a more meaningful evaluation of Thai sentence measure.

References

Achananuparp, P., Hu, X., Zhou, X., & Zhang, X. (2008). *Utilizing semantic, syntactic, and question category information for automated digital reference services*. In Digital Libraries: Universal and Ubiquitous Access to Information (pp. 203-214). Springer Berlin Heidelberg.

Agirre, E., Diab, M., Cer, D., & Gonzalez-Agirre, A. (2012, June). Semeval-2012 task 6: *A pilot on semantic textual similarity*. In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (pp. 385-393). Association for Computational Linguistics.

Aiken, M., & Balan, S. (2011). *An analysis of Google Translate accuracy*. Translation Journal, 16(2).

Almarsoomi, F. A., O'Shea, J. D., Bandar, Z. A., & Crockett, K. A. (2012). *Arabic word semantic similarity*. In Proceedings of World Academy of Science, Engineering and Technology (No. 70). World Academy of Science, Engineering and Technology.

Almarsoomi, F. A., OShea, J. D., Bandar, Z., & Crockett, K. (2013, October). *AWSS: An Algorithm for Measuring Arabic Word Semantic Similarity*. In Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on (pp. 504-509). IEEE.

Anger, S. (1998). *Thai language, alphabet and pronunciation.* Available: http://www.omniglot.com/writing/thai.htm. Last accessed 22th Jan 2013.

Aroonmanakun, W. (2007). *Creating the Thai National Corpus*. Manusaya. Special Issue, 13, 4-17.

Aroonmanakun, W., Tansiri, K., & Nittayanuparp, P. (2009, August). *Thai National Corpus: a progress report*. In Proceedings of the 7th Workshop on Asian Language Resources (pp. 153-158). Association for Computational Linguistics.

Baayen, R. H., Piepenbrock, R., & van H, R. (1993). *The {CELEX} lexical data base on {CD-ROM}*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Barzilay, R., & McKeown, K. R. (2005). *Sentence fusion for multidocument news summarization*. Computational Linguistics, 31(3), 297-328.

Battig, W. F., & Montague, W. E. (1969). *Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms*. Journal of Experimental Psychology, 803(p2), 1.

Blalock, H. M. (1960). *Social statistics* (Vol. 1221). New York: McGraw-Hill.

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2007). *Measuring semantic similarity between words using web search engines*. www, 7, 757-766.

Brin, S., & Page, L. (1998). *The anatomy of a large-scale hyper textual Web search engine*. Computer networks and ISDN systems, 30(1), 107-117.

Burnard, L. (Ed.). (1995). *British National Corpus: Users Reference Guide British National Corpus Version 1.0*. Oxford Univ. Computing Service.

Charles, W. G. (2000). *Contextual correlates of meaning*. Applied Psycholinguistics, 21(04), 505-524.

Copeland, B.J. (2000), *The Turing Archive for the History of Computing, the Turing Test*, Available: http://www.alanturing.net/turing_archive/pages/Reference%20Articles/TheTuringTest.html. Last accessed 20th Feb 2010.

Dong, Z., & Dong, Q. (2006). *HowNet and the Computation of Meaning*. Singapore: World Scientific.

Ehsani, F., Bernstein, J., & Najmi, A. (2000). *An interactive dialog system for learning Japanese*. Speech Communication, 30(2), 167-177.

Eisele, A., & Chen, Y. (2010, May). *MultiUN: A Multilingual Corpus from United Nation Documents*. InProc. of LREC, Valletta, Malta.

Elkateb, S., Black, W., Rodríguez, H., Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006). *Building a wordnet for arabic*. In Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006).

Fattah, M. A., & Ren, F. (2009). *GA, MR, FFNN, PNN and GMM based models for automatic text summarization*. Computer Speech & Language, 23(1), 126-144.

Fenton, N. E., & Pfleeger, S. L. (1998). *Software metrics: a rigorous and practical approach*. PWS Publishing Co.

Ferri, F., Grifoni, P., & Paolozzi, S. (2007, January). *Multimodal sentence similarity in human-computer interaction systems*. In Knowledge-Based Intelligent Information and Engineering Systems (pp. 403-410). Springer Berlin Heidelberg.

Firth, J. R. (1957). *A synopsis of linguistic theory, 1930-1955*. London: Oxford University Press.

Francis, W. N., & Kucera, H. (1979). *Brown corpus manual*. Brown University Department of Linguistics.

Guan, Y., Wang, X. L., Kong, X. Y., & Zhao, J. (2002). *Quantifying semantic similarity of Chinese words from HowNet*. In Machine Learning and Cybernetics, 2002. Proceedings. 2002 International Conference on (Vol. 1, pp. 234-239). IEEE.

Guo, W., & Diab, M. (2012, July). *Modelling sentences in the latent space*. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 (pp. 864-872). Association for Computational Linguistics.

Göker, M., Roth-Berghofer, T., Bergmann, R., Pantleon, T., Traphöner, R., Wess, S., & Wilke, W. (1998). *The development of HOMER a case-based CAD/CAM help-desk support tool*. In Advances in Case-Based Reasoning (pp. 346-357). Springer Berlin Heidelberg.

Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E. G., & Milios, E. (2006). *Information retrieval by semantic similarity*. International journal on semantic Web and information systems (IJSWIS), 2(3), 55-73.

Hodges, A., Monk, R., & Raphael, F. (1997). *Turing: a natural philosopher*. London: Phoenix.

Huang, Y., Zheng, F., Xu, M., Yan, P., & Wu, W. (2000, October). *Language understanding component for Chinese dialogue system*. In INTERSPEECH (pp. 1053-1056).

Ibrahim, A., & Johansson, P. (2002, October). *Multimodal dialogue systems for interactive tv applications*. In Proceedings of the 4th IEEE International Conference on Multimodal Interfaces (p. 117). IEEE Computer Society.

Inkpen, D. (2007). *Semantic similarity knowledge and its applications*. Studia Universitatis Babes-Bolyai Informatica, 52(1), 11-22.

Islam, A., & Inkpen, D. (2008). *Semantic text similarity using corpus-based word similarity and string similarity*. ACM Transactions on Knowledge Discovery from Data (TKDD), 2(2), 10.

Jarmasz, M., & Szpakowicz, S. (2004). *Roget's Thesaurus and Semantic Similarity1*. Recent Advances in Natural Language Processing III: Selected Papers from RANLP, 2003, 111.

Jiang, J. J., & Conrath, D. W. (1997). *Semantic similarity based on corpus statistics and lexical taxonomy*. In ROCLING'97, 1997.

Jin, H., & Chen, H. (2008). *SemreX: Efficient search in a semantic overlay for literature retrieval*. Future Generation Computer Systems, 24(6), 475-488.

Kendall, M. G. (1938). *A new measure of rank correlation*. Biometrika.

Kennedy, A., & Szpakowicz, S. (2008). *Evaluating Roget's thesauri*. In Proceedings of the 46th AnnualMeeting of the Association for Computational Linguistics: Human Language Technologies, pages 416–424.

Kimura, Y., Araki, K., & Tochinai, K. (2007). *Identification of spoken questions using similarity-based TF· AoI*. Systems and Computers in Japan, 38(10), 81-94.

Kritsuthikul, N., Thammano, A., & Supnithi, T. (2006, October). *English-Thai Example-Based Machine Translation using n-gram model*. In Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on (Vol. 5, pp. 4386-4390). IEEE.

Kopp, S., Gesellensetter, L., Krämer, N. C., & Wachsmuth, I. (2005, January). *A conversational agent as museum guide–design and evaluation of a real-world application*. In Intelligent Virtual Agents (pp. 329-343). Springer Berlin Heidelberg.

Kozima, H., & Furugori, T. (1993, April). *Similarity between words computed by spreading activation on an English dictionary*. In Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics (pp. 232-239). Association for Computational Linguistics.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). *An introduction to latent semantic analysis*. Discourse processes, 25(2-3), 259-284.

Lee, M. D., Pincombe, B. M., & Welsh, M. B. (2005). *An empirical evaluation of models of text document similarity*. In CogSci2005, pages 1254–1259, 2005.

Lemon, O., & Liu, X. (2006, April). *DUDE: a dialogue and understanding development environment, mapping business process models to information state update dialogue systems*. In Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations (pp. 99-102). Association for Computational Linguistics.

Lesk, M. (1986, June). *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone*. In Proceedings of the 5th annual international conference on Systems documentation (pp. 24-26). ACM.

Lewis, M. P. (ed.) (2009). *Ethnologue: Languages of the World, Sixteenth edition*. Dallas, Tex.: SIL International.

Li, Y., Bandar, Z. A., & McLean, D. (2003). *An approach for measuring semantic similarity between words using multiple information sources*. Knowledge and Data Engineering, IEEE Transactions on, 15(4), 871-882.

Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., & Crockett, K. (2006). *Sentence similarity based on semantic nets and corpus statistics*. Knowledge and Data Engineering, IEEE Transactions on, 18(8), 1138-1150.

Li, R., Li, S., & Zhang, Z. (2009, September). *The semantic computing model of sentence similarity based on Chinese FrameNet*. In Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 03 (pp. 255-258). IEEE Computer Society.

Lin, D. (1998, July). *An information-theoretic definition of similarity*. In ICML(Vol. 98, pp. 296-304).

Mayor, M. (ed.). (2009). *Longman dictionary of contemporary English*. Pearson Education India.

McSherry, D. (2001). *Interactive case-based reasoning in sequential diagnosis*. Applied Intelligence, 14(1), 65-76.

Michael, R. G., & Johnson, D. S. (1979). *Computers and Intractability: A guide to the theory of NP-completeness*. WH Freeman & Co., San Francisco.

Miller, G. A., & Charles, W. G. (1991). *Contextual correlates of semantic similarity*. Language and cognitive processes, 6(1), 1-28.

Miller, G. A. (1995). *WordNet: a lexical database for English*. Communications of the ACM, 38(11), 39-41.

Mitchell, J., & Lapata, M. (2008, June). *Vector-based Models of Semantic Composition*. In ACL (pp. 236-244).

Morris, J., & Hirst, G. (1991). *Lexical cohesion computed by the saural relations as an indicator of the structure of text*. Computational linguistics, 17(1), 21-48.

Noah, S. A., Amruddin, A. Y., & Omar, N. (2007). *Semantic similarity measures for Malay sentences*. In Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers (pp. 117-126). Springer Berlin Heidelberg.

Och, F. J. (2005). *Statistical machine translation: Foundations and recent advances*. Tutorial at MT Summit.

Osathanunkul, K., O’Shea, J., Bandar, Z., & Crockett, K. (2011). *Semantic similarity measures for the development of Thai dialog system*. In Agent and Multi-Agent Systems: Technologies and Applications (pp. 544-552). Springer Berlin Heidelberg.

O’Shea, J., Bandar, Z., Crockett, K., & McLean, D. (2008). *A comparative study of two short text semantic similarity measures*. In Agent and Multi-Agent Systems: Technologies and Applications (pp. 172-181). Springer Berlin Heidelberg.

O'Shea, J., Bandar, Z., Crockett, K., & McLean, D. (2010). *Benchmarking short text semantic similarity*. International Journal of Intelligent Information and Database Systems, 4(2), 103-120.

O'Shea, J., Bandar, Z., & Crockett, K. (2013). *A new benchmark dataset with production methodology for short text semantic similarity algorithms*. ACM Transactions on Speech and Language Processing (TSLP), 10(4), 19.

O'Shea, K., Bandar, Z., & Crockett, K. (2009, November). *A semantic-based conversational agent framework*. In Internet Technology and Secured Transactions, 2009. ICITST 2009. International Conference for (pp. 1-8). IEEE.

Pedersen, T., Pakhomov, S. V., Patwardhan, S., & Chute, C. G. (2007). *Measures of semantic similarity and relatedness in the biomedical domain*. Journal of biomedical informatics, 40(3), 288-299.

Pirró, G. (2009). *A semantic similarity metric combining features and intrinsic information content*. Data & Knowledge Engineering, 68(11), 1289-1308.

Php.net (2001). *strip_tags*. Available: www.php.net/strip_tags. Last accessed 20th Feb 2013.

Quarteroni, S., & Manandhar, S. (2009). *Designing an interactive open-domain question answering system*. Natural Language Engineering, 15(1), 73-95.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). *Development and application of a metric on semantic nets*. Systems, Man and Cybernetics, IEEE Transactions on, 19(1), 17-30.

Resnik, P. (1995). *Using information content to evaluate semantic similarity in a taxonomy*. In Proceedings of IJCAI-95, pages 448–453, Montreal, Canada.

Ricci, F., Arslan, B., Mirzadeh, N., & Venturini, A. (2002). *ITR: a case-based travel advisory system*. In Advances in Case-Based Reasoning (pp. 613-627). Springer Berlin Heidelberg.

Rodríguez, M. A., & Egenhofer, M. J. (2003). *Determining semantic similarity among entity classes from different ontologies*. Knowledge and Data Engineering, IEEE Transactions on, 15(2), 442-456.

Rubenstein, H., & Goodenough, J. B. (1965). *Contextual correlates of synonymy*. Communications of the ACM, 8(10), 627-633.

Sahami, M., & Heilman, T. D. (2006). *A web-based kernel function for measuring the similarity of short text snippets*. In Proceedings of the 15th international conference on World Wide Web (pp. 377-386).ACM.

Sammut, C. (2001). *Managing context in a conversational agent*. Linkoping Electronic Articles in Computer & Information Science, 3(7).

Sammut, C. (2007) *Conversational Agent*. Available: http://www.cse.unsw.com/~claude/thesis_topics/ConversationalAgent.html. Last accessed 24th Jan 2013.

Seymour, T., Frantsvog, D., & Kumar, S. (2011). *History Of Search Engines*. International Journal of Management & Information Systems, 15(4).

Simahasan, E.et al. (2002) *ภาษาไทย ป.๖*. Bangkok: Aksorncharoen tat

Sinclair, J. M. (2001). *Collins COBUILD English dictionary for advanced learners*. HarperCollins.

Sornlertlamvanich, V., Potipiti, T., & Charoenporn, T. (2000, July). *Automatic corpus-based Thai word extraction with the C4.5 learning algorithm*. In Proceedings of the 18th conference on Computational linguistics-Volume 2 (pp. 802-807). Association for Computational Linguistics.

Sornlertlamvanich, V., Charoenporn, T., Mokarat, C., Isahara, H., Riza, H., & Jaimai, P. (2008, January). *Synset Assignment for Bi-lingual Dictionary with Limited Resource*. In IJCNLP (pp. 673-678).

Sornlertlamvanich, V., Charoenporn, T., Robkop, K., Mokarat, C., & Isahara, H. (2009). *Review on Development of Asian WordNet*. In JAPIO 2009 Year Book, Japan Patent Information Organization, Tokyo, Japan, 2009.

Steiger, J. H. (1980). *Tests for comparing elements of a correlation matrix*. Psychological Bulletin, 87, 245-251.

Thai-language.com (1999). *Thai Consonants and Their Transcription*. Available: http://www.thai-language.com/ref/consonants. Last accessed 16th Jul 2014.

The Royal Institute (2011). *พจนานุกรมฉบับราชบัณฑิตยสถาน.* Bangkok: The Royal Institute.

Technofreak (2012). *Get Google Search Results with PHP,* Available: www.php.net/strip_tags. Last accessed 24th Jan 2013.

Thoongsup, S., Robkop, K., Mokarat, C., Sinthurahat, T., Charoenporn, T., Sornlertlamvanich, V., & Isahara, H. (2009, August). *Thai wordnet construction*. In Proceedings of the 7th Workshop on Asian Language Resources (pp. 139-144). Association for Computational Linguistics.

Trakultaweekoon, K., Porkaew, P., & Supnithi, T. (2007, December). *LEXiTRON Vocabulary Suggestion System with Recommendation and Vote Mechanism*. In Proceedings of Conference of SNLP.

Turing, A. M. (1948). *Intelligent machinery. report for national physical laboratory*. reprinted in ince, dc (editor). 1992. mechanical intelligence: Collected works of am turing.

Turing, A. M. (1950). *Computing machinery and intelligence*. Mind, 433-460.

Turing, A. M. (1952). *Can Automatic Calculating Machines Be Said To Think?*, BBC 3rd programme.

Tversky, A. (1977). *Features of similarity*. Psychological review, 84(4), 327.

Wu, Z., & Palmer, M. (1994, June). *Verbs semantics and lexical selection*. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics (pp. 133-138). Association for Computational Linguistics.

Yeh, J. Y., Ke, H. R., & Yang, W. P. (2008). *iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network*. Expert Systems with Applications, 35(3), 1451-1462.

You, L., & Liu, K. (2005). *Building chinese framenet database*. In Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on (pp. 301-306). IEEE.

Appendices

Appendix 1.1 Ethics Statement

แบบสอบถามการทดลองความคล้ายคลึงกันของความหมายของคำ

เราอยากจะขอการมีส่วนร่วมในการศึกษาทางวิทยาศาสตร์ความหมายและคล้ายคลึงกันของคำด้วยเหตุผลทางจริยธรรม ที่เรา จะต้อง ขออนุญาต ล่วงหน้า และ แจ้งให้คุณทราบสิ่งที่คุณกำลัง เห็นพ้องที่จะ เราได้ให้ คำตอบของคำถาม ทางจริยธรรมที่สำคัญด้านล่าง

สิ่งที่คุณจะ ถามฉัน จะ ทำอย่างไรหากคุณเห็นด้วย

ถ้าคุณตกลง ขอให้กรอกแบบสอบถาม โดยให้ค่าความคล้ายคลึงกันของความหมาย จาก 65 คู่ ของคำ

คุณจะถูกถามคำถามบางอย่างเกี่ยวกับตัวเองคือ ชื่อ อายุ และ ระดับการศึกษาสูงสุดของคุณ คุณจะถูกขอให้ยืนยันว่าคุณเป็นเจ้าของภาษาไทย เราขอข้อมูลส่วนบุคคลบางส่วน เพราะการศึกษาทางวิทยาศาสตร์ บางครั้งผลที่ได้รับน่าแปลกใจ ซึ่งจะต้องมีการวิเคราะห์ข้อมูล

มีความเสี่ยงใด ๆ ไหม

ความเสี่ยงที่เกี่ยวข้องกับการทดลองนี้เทียบเท่ากับการหาค่าคำสามัญในพจนานุกรม

นานแค่ไหนที่ข้อมูลที่จะถูกเก็บไว้

สำหรับคำตอบสำหรับคำถามเกี่ยวกับตัวเองจะถูกเก็บไว้ ไม่เกินความจำเป็นที่ต้องตรวจสอบหาข้อผิดพลาด หรือคุณสมบัติ ที่น่าสนใจของ ข้อมูล นี้จะ ไม่เกิน 3เดือน หลังจากที่ผล ครั้งแรกที่มีการเผยแพร่คะแนนที่ของคุณ

คุณจะเผยแพร่ข้อมูลส่วนบุคคลของฉันไหม

เราจะไม่เปิดเผยข้อมูลส่วนบุคคลของคุณให้กับทุกคนที่อยู่นอกโครงการ บทสรุปของสถิติการจัดอันดับของความหมายและคล้ายคลึงกันของคำ จะได้รับการ ตีพิมพ์ ในระดับนานาชาติ

Appendix 1.2 Instructions

การสำรวจ: ความคล้ายคลึงกันของความหมายของคำ

ขอขอบคุณทุกท่านที่เป็นอาสาสมัคร ที่จะเข้าร่วมในการศึกษานี้ คุณสามารถจะถอนตัว ก่อนที่จะเริ่ม แบบสอบถามหรือ ที่จุดใด ๆ ในขณะทำแบบสอบถาม

จะมีแบบสอบถาม ชุดของบัตรให้กับคุณ และ แผ่นบันทึกที่จะเขียนการตัดสินของคุณ (โปรดอย่าเขียนอะไรบนบัตร) บัตรที่ได้รับจะลำดับแบบสุ่ม

บัตรแต่ละใบ จะมีสองคำเขียนไว้ กรุณาเริ่มต้นด้วยการอ่านบัตรแต่ละใบ และคิดเกี่ยวกับความคล้ายคลึงกันของความหมายของสองประโยค

กรุณาจัดเรียงบัตรแต่ละใบ ให้อยู่ในลำดับความคล้ายคลึงกันของความหมายของคำให้เป็นสี่กลุ่ม

หลังจากนั้น โปรดอัตราความคล้ายคลึงกันของความหมายของแต่ละคู่ของคำ โดยการเขียนตัวเลขระหว่าง 0.0 (ความคล้ายคลึงกันของความหมายต่ำสุดที่) และ 4.0 (ความคล้ายคลึงกันของความหมายสูงสุด) บนแผ่นบันทึก คุณสามารถใช้ทศนิยมสองตำแหน่ง (เช่น 2.2)

หากคุณมีปัญหาใด ๆ คำถามหรือความคิดเห็น โปรดพูดคุยกับนักวิจัย

โปรดทราบว่าการศึกษา ไม่ได้ประเมินคุณในทางใดทางหนึ่ง - ไม่มี "ถูก" หรือ "ผิด"

Appendix 1.3 Sample Card

คู่ที่ 10	
คำที่ 1	เด็กผู้ชาย
คำที่ 2	เด็กหนุ่ม

Appendix 1.4 Sample Rating Sheet (for TWS-30)

แผ่นบันทึกอัตราความคล้ายคลึงกันของความหมายของแต่ละคู่ของคำ

โปรดอัตราความคล้ายคลึงกันของความหมายของแต่ละคู่ของคำ โดยการเขียน ตัวเลขระหว่าง 0.0 (ความคล้ายคลึงกันของความหมายต่ำสุดที่) และ 4.0 (ความคล้ายคลึงกันของความหมายสูงสุด) บนแผ่นบันทึก คุณสามารถใช้ทศนิยมสองตำแหน่ง (เช่น 2.2)

คู่ที่ 01		คู่ที่ 41		คู่ที่ 56	
คู่ที่ 05		คู่ที่ 47		คู่ที่ 57	
คู่ที่ 09		คู่ที่ 48		คู่ที่ 58	
คู่ที่ 13		คู่ที่ 49		คู่ที่ 59	
คู่ที่ 17		คู่ที่ 50		คู่ที่ 60	
คู่ที่ 21		คู่ที่ 51		คู่ที่ 61	
คู่ที่ 25		คู่ที่ 52		คู่ที่ 62	
คู่ที่ 29		คู่ที่ 53		คู่ที่ 63	
คู่ที่ 33		คู่ที่ 54		คู่ที่ 64	
คู่ที่ 37		คู่ที่ 55		คู่ที่ 65	

Appendix 1.5 Personal Data Sheet

กรุณากรอกข้อมูลส่วนบุคคล

ชื่อ:

อายุ:

ระดับการศึกษาสูงสุด:

การยืนยันว่าคุณเป็นเจ้าของภาษาไทย กรุณาเซ็น:

Appendix 1.6 Semantic Anchors

คำแนะนำ

หากคุณมีปัญหาในการประเมินผลนี่เป็นคำอธิบายของระดับความคล้ายคลึงที่จะช่วยให้คุณ:

0.0 ประโยคที่มีความหมายไม่เกี่ยวข้องกัน

1.0 ประโยคที่มีความคลุมเครือในความหมาย

2.0 ประโยคที่พอจะมีความหมายเหมือนกัน

3.0 ประโยคที่เกี่ยวข้องอย่างมากในความหมาย

4.0 ประโยคที่มีความหมายเหมือนกัน

คุณสามารถใช้ทศนิยมหนึ่งตัว ตัวอย่างเช่นถ้าคุณคิดว่าความคล้ายคลึงของความหมายเป็นครึ่งหนึ่งระหว่าง 3.0 และ 4.0 คุณสามารถใช้ค่า 3.5

Appendix 1.7 Sample Rating Sheet (for TWS-65)

แผ่นบันทึกอัตราความคล้ายคลึงกันของความหมายของแต่ละคู่ของคำ

โปรดอัตราความคล้ายคลึงกันของความหมายของแต่ละคู่ของคำ โดยการเขียน ตัวเลขระหว่าง 0.0 (ความคล้ายคลึงกันของความหมายต่ำสุดที่) และ 4.0 (ความคล้ายคลึงกันของความหมายสูงสุด) บนแผ่นบันทึก คุณสามารถใช้ทศนิยมสองตำแหน่ง (เช่น 2.2)

คู่ที่ 01		คู่ที่ 32		คู่ที่ 63	
คู่ที่ 02		คู่ที่ 33		คู่ที่ 64	
คู่ที่ 03		คู่ที่ 34		คู่ที่ 65	
คู่ที่ 04		คู่ที่ 35		คู่ที่ 70	
คู่ที่ 05		คู่ที่ 36		คู่ที่ 71	
คู่ที่ 06		คู่ที่ 40		คู่ที่ 72	
คู่ที่ 10		คู่ที่ 41		คู่ที่ 73	
คู่ที่ 11		คู่ที่ 42		คู่ที่ 74	
คู่ที่ 12		คู่ที่ 43		คู่ที่ 75	
คู่ที่ 13		คู่ที่ 44		คู่ที่ 80	
คู่ที่ 14		คู่ที่ 45		คู่ที่ 81	
คู่ที่ 15		คู่ที่ 46		คู่ที่ 82	
คู่ที่ 16		คู่ที่ 50		คู่ที่ 83	
คู่ที่ 20		คู่ที่ 51		คู่ที่ 84	
คู่ที่ 21		คู่ที่ 52		คู่ที่ 85	
คู่ที่ 22		คู่ที่ 53		คู่ที่ 90	
คู่ที่ 23		คู่ที่ 54		คู่ที่ 91	
คู่ที่ 24		คู่ที่ 55		คู่ที่ 92	
คู่ที่ 25		คู่ที่ 56		คู่ที่ 93	
คู่ที่ 26		คู่ที่ 60		คู่ที่ 94	
คู่ที่ 30		คู่ที่ 61		คู่ที่ 95	
คู่ที่ 31		คู่ที่ 62			

Appendix 2.1 Instructions for High Similarity Word Pairs

การสำรวจ: ความคล้ายคลึงกันของความหมายของคำ

ขอขอบคุณทุกท่านที่เป็นอาสาสมัคร ที่จะเข้าร่วมในการศึกษานี้ คุณสามารถจะถอนตัว ก่อนที่จะเริ่ม แบบสอบถามหรือ ที่จุดใด ๆ ในขณะทำแบบสอบถาม

จะมีแบบสอบถาม ชุดของคำให้กับคุณ และ แผ่นบันทึกที่จะเขียนการตัดสินของคุณ

กรุณาเลือกคำจากสองกลุ่ม กลุ่มละคำที่คุณคิดมีความคล้ายคลึงกันของความหมายของคำมากที่สุด และกรอกลงในแผ่นบันทึก

หากคุณมีปัญหาใด ๆ คำถามหรือความคิดเห็น โปรดพูดคุยกับนักวิจัย

โปรดทราบว่าการศึกษา ไม่ได้ประเมินคุณในทางใดทางหนึ่ง - ไม่มี "ถูก" หรือ "ผิด"

Appendix 2.2 List of Theme Words

กลุ่มของคำ

กลุ่ม A

1. นักบวช
2. อัญมณี
3. โรงภาพยนต์
4. เด็กผู้ชาย
5. ผ้าฝ้าย
6. อาหาร
7. ลุง
8. ทาส
9. การเดินทาง
10. ลายมือชื่อ
11. นักมายากล
12. รถเก๋ง
13. สุนัข
14. เที่ยงวัน
15. วัด
16. ฝั่งทะเล
17. นิตยสาร
18. ครู
19. เนินเขา
20. อุปกรณ์
21. แก้ว
22. ป่าช้า
23. พืช
24. ป่าไม้

กลุ่ม B

1. ภูเขา
2. รถยนต์
3. หนังสือ
4. พ่อมด
5. ข้ารับใช้
6. อาจารย์
7. เครื่องมือ
8. เพชรพลอย
9. ชายฝั่ง
10. พระ
11. พงไพร
12. โรงละคร
13. ถ้วย
14. สุสาน
15. การท่องเที่ยว
16. ผ้าไหม
17. โบสถ์
18. ผลไม้
19. เด็กหนุ่ม
20. หมา
21. กลางวัน
22. ต้นไม้
23. ป่า
24. ลายเซ็น

Appendix 2.3 High Similarity Word Pairs Recording Sheet

แผ่นบันทึกคู่ของคำที่มีความคล้ายคลึงกันของความหมายของคำมากที่สุด

กรุณาเลือก 20 คู่ของคำจากสองกลุ่ม กลุ่มละคำที่คุณคิดมีความคล้ายคลึงกันของความหมายของคำมากที่สุด และกรอกลงในแผ่นบันทึก

กลุ่ม A	**กลุ่ม B**	**กลุ่ม A**	**กลุ่ม B**

Appendix 2.4 Instructions for Medium Similarity Word Pairs

การสำรวจ: ความคล้ายคลึงกันของความหมายของคำ

ขอขอบคุณทุกท่านที่เป็นอาสาสมัคร ที่จะเข้าร่วมในการศึกษานี้ คุณสามารถจะถอนตัว ก่อนที่จะเริ่ม แบบสอบถามหรือ ที่จุดใด ๆ ในขณะทำแบบสอบถาม

จะมีแบบสอบถาม ชุดของคำให้กับคุณ และ แผ่นบันทึกที่จะเขียนการตัดสินของคุณ

กรุณาเลือกคำจากสองกลุ่ม กลุ่มละคำที่คุณคิดมีความคล้ายคลึงกันของความหมายของคำมากที่สุดที่ไม่ซ้ำกับการทดลองที่แล้วและกรอกลงในแผ่นบันทึก

หากคุณมีปัญหาใด ๆ คำถามหรือความคิดเห็น โปรดพูดคุยกับนักวิจัย

โปรดทราบว่าการศึกษา ไม่ได้ประเมินคุณในทางใดทางหนึ่ง - ไม่มี "ถูก" หรือ "ผิด"

Appendix 2.5 Medium Similarity Word Pairs Recording Sheet

แผ่นบันทึกคู่ของคำที่มีความคล้ายคลึงกันของความหมายของคำมากที่สุดที่ไม่ซ้ำกับการทดลองที่แล้ว

กรุณาเลือก 21 คู่ของคำจากสองกลุ่ม กลุ่มละคำที่คุณคิดมีความคล้ายคลึงกันของความหมายของคำมากที่สุดที่ไม่ซ้ำกับการทดลองที่แล้วและกรอกลงในแผ่นบันทึก

กลุ่ม A	**กลุ่ม B**	**กลุ่ม A**	**กลุ่ม B**

Appendix 3.1 Instructions (for TSS-65)

การสำรวจ: การเลือกความหมายที่เหมาะสมของคำ

ขอขอบคุณทุกท่านที่เป็นอาสาสมัคร ที่จะเข้าร่วมในการศึกษานี้ คุณสามารถจะถอนตัว ก่อนที่จะเริ่ม แบบสอบถามหรือ ที่จุดใด ๆ ในขณะทำแบบสอบถาม

จะมีแบบสอบถาม ชุดของความหมายให้กับคุณ และ แผ่นบันทึกที่จะเขียนการตัดสินของคุณ

กรุณาเลือกการเลือกความหมายที่เหมาะสมของคำและกรอกลงในแผ่นบันทึก

หากคุณมีปัญหาใด ๆ คำถามหรือความคิดเห็น โปรดพูดคุยกับนักวิจัย

โปรดทราบว่าการศึกษา ไม่ได้ประเมินคุณในทางใดทางหนึ่ง - ไม่มี "ถูก" หรือ "ผิด"

Appendix 3.2 Sample Question Sheet

ความหมาย	**คำ *ต้นไม้***	**กรุณาเลือก**
1	ต้นไม้คือพืชคือสิ่งมีชิวิตสีเขียวชนิดนึง	
2	ต้นไม้คือคำรวมเรียกพืชทั่วไป โดยปกติชนิดมีลำต้น	
3	ต้นไม้คือไม้ยืนต้นขนาดใหญ่	
4	ต้นไม้คือพืชที่มีอายุยืนยาว	
5	ต้นไม้คือพืชชนิดที่มีลำต้นใหญ่มีกิ่งแยกออกไป	

Appendix 4.1 Ethics Statement (for TSS-65)

แบบสอบถามการทดลองความคล้ายคลึงกันของความหมายของประโยค

เราอยากจะขอการมีส่วนร่วมในการศึกษาทางวิทยาศาสตร์ความหมายและคล้ายคลึงกันของประโยค ด้วยเหตุผลทางจริยธรรม ที่เรา จะต้อง ขออนุญาต ล่วงหน้า และ แจ้งให้คุณทราบสิ่งที่คุณกำลัง เห็นพ้องที่จะ เราได้ให้ คำตอบของคำถาม ทางจริยธรรมที่สำคัญด้านล่าง

สิ่งที่คุณจะ ถามฉัน จะ ทำอย่างไรหากคุณเห็นด้วย

ถ้าคุณตกลง ขอให้กรอกแบบสอบถาม โดยให้ค่าความคล้ายคลึงกันของความหมาย จาก 65 คู่ ของประโยค

คุณจะถูกถามคำถามบางอย่างเกี่ยวกับตัวเองคือ ชื่อ อายุ และ ระดับการศึกษาสูงสุดของคุณ คุณจะถูกขอให้ยืนยันว่าคุณเป็นเจ้าของภาษาไทย เราขอข้อมูลส่วนบุคคลบางส่วน เพราะการศึกษาทางวิทยาศาสตร์ บางครั้งผลที่ได้รับน่าแปลกใจ ซึ่งจะต้องมีการวิเคราะห์ข้อมูล

มีความเสี่ยงใด ๆ ไหม

ความเสี่ยงที่เกี่ยวข้องกับการทดลองนี้เทียบเท่ากับการหาค่าคำสามัญในพจนานุกรม

นานแค่ไหนที่ข้อมูลที่จะถูกเก็บไว้

สำหรับคำตอบสำหรับคำถามเกี่ยวกับตัวเองจะถูกเก็บไว้ ไม่เกินความจำเป็นที่ต้องตรวจสอบหาข้อผิดพลาด หรือคุณสมบัติ ที่น่าสนใจของ ข้อมูล นี้จะ ไม่เกิน 3เดือน หลังจากที่ผล ครั้งแรกที่มีการเผยแพร่คะแนนที่ของคุณ

คุณจะเผยแพร่ข้อมูลส่วนบุคคลของฉันไหม

เราจะไม่เปิดเผยข้อมูลส่วนบุคคลของคุณให้กับทุกคนที่อยู่นอกโครงการ บทสรุปของสถิติการจัดอันดับของความหมายและคล้ายคลึงกันของประโยค จะได้รับการ ตีพิมพ์ ในระดับนานาชาติ

Appendix 4.2 Instructions (for TSS-65)

การสำรวจ: ความคล้ายคลึงกันของความหมายของประโยค

ขอขอบคุณทุกท่านที่เป็นอาสาสมัคร ที่จะเข้าร่วมในการศึกษานี้ คุณสามารถจะถอนตัว ก่อนที่จะเริ่ม แบบสอบถามหรือ ที่จุดใด ๆ ในขณะทำแบบสอบถาม

จะมีแบบสอบถาม ชุดของบัตรให้กับคุณ และ แผ่นบันทึกที่จะเขียนการตัดสินของคุณ (โปรดอย่าเขียนอะไรบนบัตร) บัตรที่ได้รับจะลำดับแบบสุ่ม

บัตรแต่ละใบ จะมีสองประโยค เขียนไว้ กรุณาเริ่มต้นด้วยการอ่านบัตรแต่ละใบ และคิดเกี่ยวกับความคล้ายคลึงกันของความหมายของสองประโยค

กรุณาจัดเรียงบัตรแต่ละใบ ให้อยู่ในลำดับความคล้ายคลึงกันของความหมายของประโยค ให้เป็นสี่กลุ่ม

หลังจากนั้น โปรดอัตราความคล้ายคลึงกันของความหมายของแต่ละคู่ของประโยค โดยการเขียน ตัวเลขระหว่าง 0.0 (ความคล้ายคลึงกันของความหมายต่ำสุดที่) และ 4.0 (ความคล้ายคลึงกันของความหมายสูงสุด) บนแผ่นบันทึก คุณสามารถใช้ทศนิยมสองตำแหน่ง (เช่น 2.2)

คุณอาจจะหรืออาจจะไม่เห็นด้วยกับ ความหมายของแต่ละประโยค กรุณาอย่าอนุญาตให้มีอิทธิพลต่อการตัดสินใจของคุณ

หากคุณมีปัญหาใด ๆ คำถามหรือความคิดเห็น โปรดพูดคุยกับนักวิจัย

โปรดทราบว่าการศึกษา ไม่ได้ประเมินคุณในทางใดทางหนึ่ง - ไม่มี "ถูก" หรือ "ผิด"

Appendix 4.3 Sample Card (for TSS-65)

คู่ที่ 63	
ประโยค ที่ 1	**วัดคือที่สถานที่พักพิงทางศาสนาพุทธ**
ประโยค ที่ 2	**โบสถ์คือสถานที่ทางศาสนาคริสต์**

Appendix 4.4 Sample Rating Sheet (for TSS-65)

แผ่นบันทึกอัตราความคล้ายคลึงกันของความหมายของแต่ละคู่ของประโยค

โปรดอัตราความคล้ายคลึงกันของความหมายของแต่ละคู่ของประโยค โดยการเขียน ตัวเลขระหว่าง 0.0 (ความคล้ายคลึงกันของความหมายต่ำสุดที่) และ 4.0 (ความคล้ายคลึงกันของความหมายสูงสุด) บนแผ่นบันทึก คุณสามารถใช้ทศนิยมสองตำแหน่ง (เช่น 2.2)

คู่ที่ 01		คู่ที่ 32		คู่ที่ 63	
คู่ที่ 02		คู่ที่ 33		คู่ที่ 64	
คู่ที่ 03		คู่ที่ 34		คู่ที่ 65	
คู่ที่ 04		คู่ที่ 35		คู่ที่ 70	
คู่ที่ 05		คู่ที่ 36		คู่ที่ 71	
คู่ที่ 06		คู่ที่ 40		คู่ที่ 72	
คู่ที่ 10		คู่ที่ 41		คู่ที่ 73	
คู่ที่ 11		คู่ที่ 42		คู่ที่ 74	
คู่ที่ 12		คู่ที่ 43		คู่ที่ 75	
คู่ที่ 13		คู่ที่ 44		คู่ที่ 80	
คู่ที่ 14		คู่ที่ 45		คู่ที่ 81	
คู่ที่ 15		คู่ที่ 46		คู่ที่ 82	
คู่ที่ 16		คู่ที่ 50		คู่ที่ 83	
คู่ที่ 20		คู่ที่ 51		คู่ที่ 84	
คู่ที่ 21		คู่ที่ 52		คู่ที่ 85	
คู่ที่ 22		คู่ที่ 53		คู่ที่ 90	
คู่ที่ 23		คู่ที่ 54		คู่ที่ 91	
คู่ที่ 24		คู่ที่ 55		คู่ที่ 92	
คู่ที่ 25		คู่ที่ 56		คู่ที่ 93	
คู่ที่ 26		คู่ที่ 60		คู่ที่ 94	
คู่ที่ 30		คู่ที่ 61		คู่ที่ 95	
คู่ที่ 31		คู่ที่ 62			