

國立臺灣大學文學院語言學研究所



碩士論文

Graduate Institute of Linguistics

College of Liberal Arts

National Taiwan University

Master Thesis

詞彙穩定的秘密—對各語言學面向的質性與量化分析

Secrets of Lexical Conventionalization:

A Quantitative and Qualitative Exploratory Analysis on Linguistic
Factors

王伯雅

Po-Ya Angela Wang

指導教授：謝舒凱 博士

Advisor: Shu-Kai Hsieh, Ph.D.

中華民國 104 年 7 月

July, 2015

國立臺灣大學碩士學位論文
口試委員會審定書

詞彙穩定的秘密—對各語言學面向的質性與量化分析
Secrets of Lexical Conventionalization: A Quantitative
and Qualitative Exploratory Analysis on Linguistic Factors

本論文係王伯雅君 (r01142009) 在國立臺灣大學語言學學系、所
完成之碩士學位論文，於民國 104 年 07 月 22 日承下列考試委員審查
通過及口試及格，特此證明

口試委員：

謝麗凱

(簽名)

(指導教授)

劉德馨

高照明

Acknowledgements



真的很難想像，三年...一晃眼就過去了...

三年的時間修了許多課，認識了許多人，不管是知識上的充實，或是心靈上的溫暖，都讓我十分感激。很謝謝指導我的謝舒凱老師，總是在許多時刻鼓勵我，給予我最真實的溫暖，讓我可以勇敢地跨越每一個關卡，能夠受到謝老師指導真的是我極大的幸運。感謝口試委員劉德馨老師以及高照明老師在質性與量化研究都給予了我許多意見和啟發，更不吝地給予我肯定。修課的過程中很感謝呂佳蓉老師課堂上針對我論文雛型所給予的討論，李佳霖老師碩一以來對我的照顧及腦心研究面向的啟發，蘇以文老師、宋麗梅老師、劉德馨老師、馮怡蓁老師、周泰立老師課堂上知識的傳授，讓我能在論文中思量語言學各角度切入的可能性。每位老師還有溫暖的美玲助教、嘉蘭姊、白姊與黃宣範老師課餘的關心和問候更是點滴在心頭。除了研究所的良師，大學時期英語系、國文系以及學程的老師們在做學問以及待人接物的指導我都深懷感激。

所上學長姊、學弟妹與同儕更是心靈上的支柱。謝謝總是跟我分享快樂與不快樂的神之妍，總是樂於跟我討論的喬神，耐心關心我程式的 Simon、Mars、阿吉、Taco，找到相關研究馬上分享給我的 Mike 底迪以及天才 Nate，還有一起咬緊牙關，相互打氣走到現在的論文三妹 Emily、Yvonne 與靜琛。謝謝瑜珈好朋友 Sally 學姊，謝謝關心鼓勵我的玥彤、婉如、盛傑、Sara 學姊、乃欣學姊、安婷、聖富、雷神、南西、Dave、高高、Debbie、于萱、Rita、Winnie、Iju 等等，不管我們同屆不同屆，每次相遇的關懷讓我時刻充滿了動力。感謝即便我很難約，也還是愛著我，用不同方式為我加油打氣的故人們：大星、線線、May、葛雷斯、Lucy、小樹、Annie、Cabi、Kimberly、Fanny、Jane、Wendy、Mavis、Corrinne、Gracie、Joyce、Ocean。時不時拉我充電陪我解憂的兔兔、王子、菇菇、腹部、左左、雨滴、姥姥、Jesse、比斯吉、蝦子、Derek、總總、彥廷、洋蔥、西批、KK、來福。

還有，我最愛最愛的媽媽、爸爸、哥哥、阿姨舅舅們與表姊妹們，因為有你們...我，才能是我。不管有沒有一一列出，這段時間的點點滴滴我想一輩子留在心頭，謝謝出現在我生活中的人們。

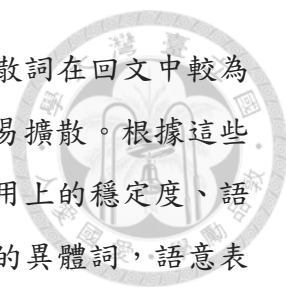


摘要

前人的在語詞上的研究有許多見解，主要可分為兩部分：語言理論上的分析和語言處理的應用。理論上的分析主要包含三個角度：研究語言現象歷史發展的歷史語言學，新詞共時表現的詞彙語義學，預測詞彙存留的計算語言學。他們都可以運用於字典學，設計語言教材，構建自然語言處理所需的資源。然而，在相關研究中少有同時採用量化和質性角度的探討。其次，前人研究中所選取的目標詞彙有其侷限性。同時，時間訊息以及各類語言學變相都應納入討論以及更深刻的了解詞彙穩定的肇因。詞彙以概念連結的組構模式以及隨著時間積累的心理詞庫都應在探討本議題時納入考量。因此，本文欲以量化和質性觀點切入研究，提出詞彙可能有的三種生命形態（擴散、穩定、失去活性），透過時間資訊以及六種語言學面向（聲韻、構詞、語意、句法、語用、社會語言學）來探討本議題，並期能將結果運用於詞彙預測以及資源建構。

量化分析的角度來看，線性回歸模型用以研究區分不同時間點詞彙的語言學特色。語用學顯著地解釋了1950年以前存在的詞彙期使用穩定度的高低，而1950年以後所造的詞是否在語言中穩定使用則有賴語法面向的因素來解釋。這樣的結果暗示詞彙活得越久越與經驗性和語用性知識相關，但對於近期新生的詞彙句法結構的結合性對於其是否會被穩定使用有著決定性的意義。新起的擴散詞以及存在數世紀的詞彙在使用穩定度上十分相似，但藉由邏輯回歸模型可以發現數音節、近義詞數、同義詞數目、在回文中使用的活躍度、是否為外來語成功區別擴散詞以及存在數世紀的詞彙。另一方面，語言學特質的角度而言1950年後新生的詞彙與近來新起的擴散詞有相似的語言學特徵。所以將1950年以後新生的詞作為訓練資料建構預測模型來理解現下擴散的詞未來發展的趨勢。結果顯示目標詞前後共現的不同詞彙數有顯著的預測能力，達到0.6335的準確度。

質性分析的面向從同義詞間的競爭來探討，句法上的兼容性和該詞概念關係的豐富度應為是否能贏過其他同義詞而被大量使用的關鍵。此外，不同時間點生



成的詞在貼文與回文中有不同的使用活性。不同於其他兩者擴散詞在回文中較為活躍，這暗示他們在類似回覆導向的口語風格中以及互動中較易擴散。根據這些研究發現，我們可以進一步應用於增補詞彙於語言資源中。語用上的穩定度、語法上的結合性，以及語意可作為增補詞彙的標準，較廣泛使用的具體詞，語意表達中較穩定使用的詞彙，以及來自同一概念經歷詞彙化的詞項皆收錄於增補後的詞，由此可知所提標準的涵蓋性。

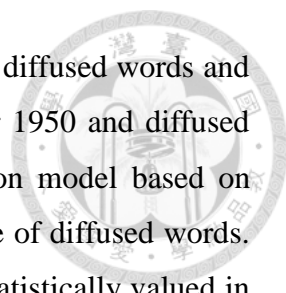
關鍵詞：詞彙穩定、詞彙生命、新詞、詞彙擴散、網路語言、語言改變、量化語言學、語料庫、字典學

Abstract



Previous studies have many insights in understanding lexical items. They can be generally captured into two parts: linguistic analysis and application. Linguistic analysis mainly includes three angles: studies on historical development of linguistic phenomenon from Historical Linguistics, probes on synchronic emergence of neologisms from Lexical Semantics, and prediction models built for understanding survival of words from Computational Linguistics. They can all be applied on including words for Lexicology, designing language teaching materials, and constructing resources for Natural Language Processing. However, there is rarely a single work include quantitative and qualitative methods simultaneously. Second, the generality of included target words in previous studies needs reconsideration. Meanwhile, temporal information of lexical items and various linguistic aspects should be invited to probe deeper for understanding factors contributing to conventionalization of a word. The conceptual associations of organization in mental lexicon and temporal accumulation for mental lexicon should all be considered when facing this issue. Thus, this thesis is aimed to conduct quantitative profiling and qualitative analysis as well as to apply them in constructing lexical resources with proposing three life stages of lexical items (diffusion, conventionalization, and inactivation), including target words from different temporal points, and adopting linguistic variables from six linguistic aspects (phonology, morphology, semantics, syntax, pragmatics, and sociolinguistics).

In quantitative profiling, the linear regression model has built to distinguish words from different temporal points. The result shows that pragmatics can best account behavioral performance of words before 1950 and syntax can best capture words after 1950, which implies that words live longer may correlated with rich experiential and pragmatic using knowledge, but for those who are born recently their structurally syntactic compatibility plays important role in deciding their fluctuation in use. Diffused words are similar to words existing over centuries in their Revised Constant U. From logistic regression model it is found that number of syllable, number of near-synonym, number of synonym, activeness in used in comments, and borrowing from other



language or not are statistically significant variables that distinguish diffused words and words existing over centuries. On the other hand, words born after 1950 and diffused words are quite similar in their linguistic characteristics. Prediction model based on training data from words after 1950 are built to foretell potential life of diffused words. It shows that number of types co-occurring before target words is statistically valued in prediction. With words before 1950 and recent diffused words as test data the accuracy of model reaches 0.6335.

Qualitative analysis on competitions among words from the same synset indicates that structural compatibility and involved conceptual relations may be the key for one lexical item to winning over the other synonymous member. Besides, words coming from different temporal points show differences in their activeness in being used in comments and posts on PTT. Diffused words are more active in comments, which implies they are more correlated with feedback oriented oral style and diffused in interaction. With these findings we can further apply them on proposing suggestions for lexicology. Pragmatically stable in use, syntactic compatibility, and semantically number of senses are taken as standard to expanding inclusion of words. The updated inclusion of popularly used variants, more stable semantic representation, and words lexicalized from the same conceptual experiences indicates the inclusiveness of proposed standards.

Key words: conventionalization, life cycle of words, neologism, diffusion, internet language, language change, quantitative linguistics, corpus, lexicology

Table of Contents



Acknowledgements	ii
摘要	iii
Abstract.....	v
Table of Contents.....	vii
List of Figures.....	x
List of Tables	xiii
List of Appendices	xv
Chapter 1. Introduction.....	1
1.1. Background.....	1
1.2. Purpose	3
1.3. Organization	4
Chapter 2. Literature review	6
2.1. Qualitative Discussion from Historical linguistics and Lexical Semantics.....	7
2.1.1. Historical Linguistics: Grammaticalization, Degrammaticalization, Lexicalization, and Exaptation	7
2.1.2. Lexical Semantics on Neology	12
2.2. Quantitative Analysis on “Life Cycle” of Lexical Items.....	18
2.2.1. Analysis on “Life Cycle” of Different Lexical Items	18
2.2.2. Quantitative Profiling on “Life” of Lexical Items.....	22
2.3. Applications	26
Chapter 3. Methodology	29
3.1. Scope of Study.....	29
3.1.1. Unit of Observation	29
3.1.2. Types of Target Words	31
3.1.3. Potential Limitation and Corresponding Compensation in Current Study	34
3.1.4. Proposed Life Stages	35
3.1.5. Operational Definitions on Predicted Value	39

3.2. Resource for Collecting Target words	42
3.2.1. Kim (2006) and Chang and Ahrens (2008)	43
3.2.2. Google Books Ngram Corpus (GBNC).....	43
3.2.3. Web	45
3.2.4. Newspaper	46
3.2.5. Chinese Wordnet.....	46
3.3. Categorization on Target Lexical Items.....	48
3.3.1. Target Lexical Items for Understanding Diffusion.....	49
3.3.2. Target Verbs for Understanding Conventionalization	50
3.4. Proposed Linguistic Predictors for Understanding Stabilization	52
3.4.1. Phonology.....	56
3.4.2. Morphology	57
3.4.3. Syntax	66
3.4.4. Semantics.....	70
3.4.5. Sociolinguistics.....	73
3.4.6. Pragmatics	75
Chapter 4. Exploratory Analysis and Modeling	81
4.1. Revised Constant U in Three Types of Targets	81
4.2. Performance of Linguistic Factors in Target Words	88
4.3. Linguistic Regression Models for Three Sets of Words.....	99
4.3.1. Revised Constant U and Phonology	102
4.3.2. Revised Constant U and Morphology	103
4.3.3. Revised Constant U and Semantics.....	104
4.3.4. Revised Constant U and Syntax	105
4.3.5. Revised Constant U and Pragmatics.....	105
4.3.6. Revised Constant U and Sociolinguistics	107
4.3.7. Logistic Regression Model.....	109
4.4. Qualitative Analysis on Members of Synset	114
4.5. Application: Inclusion of Lexical Items for Lexicology	121
Chapter 5. General discussion and conclusion	128
5.1. Conclusion.....	128
5.2. Implication and future study.....	129

References 131
Appendices 141



List of Figures



Figure 1 Model of the Invited Inferencing Theory of Semantic Change (Traugott and Dasher, 2004).....	9
Figure 2 Distribution of POS in GBNC	32
Figure 3 "Lifespan" of all Parts of Speech in GBNC	32
Figure 4 Comparison in English VerbNet, FrameNet, Levin's classification, Roget's thesaurus, and WordNet (Baker, 2008).....	47
Figure 5 Quantitative Information About CWN.....	47
Figure 6 Number of Three Types of Target Words Observed in Current Study	52
Figure 7 Schematic Representations About Classification on Borrowing by Duckworth (1977)	73
Figure 8 Cluster of Concepts Adopted from Speer and Havasi (2012).....	76
Figure 9 Distribution of Revised Constant U for all Target words.....	82
Figure 10 Distribution of by Revised Constant U for Target words Before 1950.....	82
Figure 11 Cross Month Frequency Distribution for Top 10 Target Words Before 1950	83
Figure 12 Cross Month Frequency Distribution for Tail 10 Target Words Before 1950	84
Figure 13 Distribution of by Month Constant U for Target Words After 1950	84
Figure 14 Cross Month Frequency Distribution for Top 10 Target Words After 1950 ..	85
Figure 15 Cross Month Frequency Distribution for Tail 10 Target Words After 1950 ..	86
Figure 16 Distribution of by Month Constant U for Diffused words	86
Figure 17 Cross Month Frequency Distribution for Top 10 Diffused Words.....	87
Figure 18 Cross Month Frequency Distribution for Tail 10 Diffused Words.....	87
Figure 19 Distribution of Revised Constant U of Three Sets of Target Words	88
Figure 20 Number of Syllables for Three Sets of Target Words	89
Figure 21 With Mixed Originated Morphemes or not.....	89
Figure 22 Encoded in Chinese or not	90
Figure 23 Component Richness: Realized Productivity.....	91
Figure 24 Component Richness: Type- Token Ratio.....	91
Figure 25 Distribution of Variants	92

Figure 26 Distribution of Parts of Speech	92
Figure 27 Distribution of Co-occurring Types (Before Target Words)	93
Figure 28 Distribution of Co-occurring Types (After Target Words)	93
Figure 29 Distribution of Number of Senses	95
Figure 30 Distribution of Number of Involved Synonymous Relation	96
Figure 31 Upper Panel: Distribution of Involved Conceptual Relationships, Lower Panel: Distribution of Related Conceptual Words	96
Figure 32 Actively used in Posts or not	97
Figure 33 Actively used in Comments or not	98
Figure 34 Distribution of Loan Words in Each Target Word Set	98
Figure 35 Dissemination in Each Target Set of Words	99
Figure 36 Density Plots for Revised Constant U of all Target Words: from top to the bottom shows separately the distribution of original data, of log transformation, of square root, and of inverse in transforming data	100
Figure 37 Density Plots for Constant U of 3 Sets of Words: from top to the bottom shows separately the distribution of original data in Words Before 1950, Words after 1950, and Diffused Words	101
Figure 38 Density Plots for Constant U of 3 Sets of Words: from top to the bottom shows separately the distribution of log transformation of Revised Constant U in Words Before 1950, Words after 1950, and Diffused Words	102
Figure 39 Residual Plots for Pragmatic Model for Words Before 1950	106
Figure 40 Residual Plots for Multiple Linear Regression model for Diffused Words ..	112
Figure 41 Cross Month Frequency of Synset Members	114
Figure 42 Distribution of Total frequency and Revised Constant U	115
Figure 43 Synset Members Separately Ordered by Frequency (Upper panel) and Revised Constant U Value (Lower panel) Decreasingly	116
Figure 44 Number of Co-occurring Words (After Target) with Target Words Ordered by Constant U Value Decreasingly	117
Figure 45 Number of Co-occurring Words (Before Target) with Target Words Ordered by Constant U Value Decreasingly	117
Figure 46 Number of Related Conceptual Words with Target Words Ordered by Revised Constant U Value Decreasingly	118

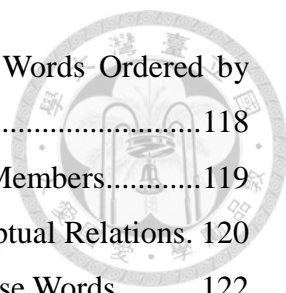


Figure 47 Number of Involved Conceptual Relations with Target Words Ordered by Revised Constant U Value Decreasingly	118
Figure 48 Distribution of Involved Conceptual Relation and Synset Members.....	119
Figure 49 Synset Members and their Corresponding Involved Conceptual Relations.	120
Figure 50 Revised Constant U Values of Target Words in 8000 Chinese Words	122
Figure 51 Words Suggested to be Included in 8000 Chinese Words from Words after 1950	123
Figure 52 Variants and Synset of “吸煙” Ordered by Revised Constant U	125
Figure 53 Involved Conceptual Relations for Variants and Synset of “吸煙”	125
Figure 54 Words from Embodiment of Experiences about Temperature	127

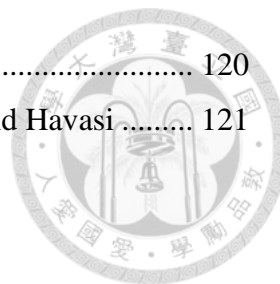
List of Tables



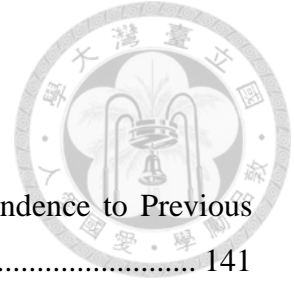
Table 1 Proposed Stages of Words (Schmid, 2008).....	16
Table 2 Threshold for "Survival" or "Failure" of Words (Chang & Ahrens, 2008).....	23
Table 3 Proposed Hypotheses of Influencing Life of Words by Kerremans (2015).....	25
Table 4 Types of Diffusion Words Included	50
Table 5 Factors from Kjellmer (2000), Summarized by Chang and Ahrens (2008)	53
Table 6 FUDGE scale from Metcalf (2002),	55
Table 7 Factors used in Hybrid Model from Chang and Ahrens (2008)	55
Table 8 Summarized Chart for Phonological Predictor.....	57
Table 9 Comparisons on Methods used in Calculating Morphological Productivity.....	61
Table 10 Summarized Chart for Morphological Predictors.....	66
Table 11 Summarized Chart for Syntactic Predictors.....	69
Table 12 Summarized Chart for Semantic Predictors	72
Table 13 Summarized Chart for Sociolinguistics Predictors.....	75
Table 14 Interlingual relations in ConceptNet Adopted from Speer and Havasi (2012) 77	
Table 15 Summarized Chart for Pragmatic Predictors	80
Table 16 Summary Statistics for Before and After Co-occurring Word Types for Each Set.....	94
Table 17 Revised Constant U and Phonology	103
Table 18 Revised Constant U and Morphology.....	104
Table 19 Revised Constant U and Semantics	104
Table 20 Revised Constant U and Syntax	105
Table 21 Revised Constant U and Pragmatics.....	106
Table 22 Revised Constant U and Sociolinguistics	107
Table 23 Parametric Statistic Wald test for Logistic Model of Conventionalized and Unconventionalized Words.....	110
Table 24 Statistic Information for Logistic Model of Conventionalized and Unconventionalized Words.....	110
Table 25 Formula of Multiple Linear Regression Model of Words Born After 1950 ...	113
Table 26 Number of Conceptual Relations and Revised Constant U Value of Synset	

Members 120

Table 27 Interlingual relations in ConceptNet Adopted from Speer and Havasi 121



List of Appendices



Appendix 1 Predictors Used in Current Study and Their Correspondence to Previous Models	141
Appendix 2 Brief Summarization on Boards Used in Current Study	150
Appendix 3 Constant U value for Lexical Items Before 1950	160
Appendix 4 Constant U value for Lexical Items After 1950	165
Appendix 5 Constant U value for Diffused Lexical Items	172

Chapter 1.

Introduction



1.1. Background

Current study assumes that words are important targets in studies. A word uttered by a person may reveal his or her gender, belief, age, and cognition operation. Words are more than conveying literal meaning, but also bearing important pragmatic, syntactic, sociolinguistic, phonological and morphological issues. Lexicology is a branch of linguistics which concerns words from various perspectives, for example how to include words that can be used generation by generation, how to display synonyms, or how to design efficiently for language users to consult. Its importance has been increasingly recognized in recent years, due to the fact that many lexical resources have been successfully applied to Natural Language Processing (NLP), socio-cultural understanding as well as language teaching and learning. However, given that there are rich discussions on computational lexicology, semantic relationship in lexicology, user-oriented lexicology, or digital revolution on lexicology (Moon, 2013; Murphy, 2013; Chiara, 2013; Marie-Claude, 2014; Fellbaum, 2014; Polguere, 2014), there is few discussion on how words enter and live their life in our lexicon as well as the linguistic factors driven behind. It is even harder to tackle with the issue in Chinese since the notion of wordhood is still in great controversy. Linguistically we may propose that frequency-effect takes the lead and entrenchment drives the effects; nevertheless, how frequent the occurrence should be could we claim it is entrenched is a question. On the other hand, in the society with rich language contact, code-switching and loan words are common phenomenon, at what degree can we claim that the word is already loaned into

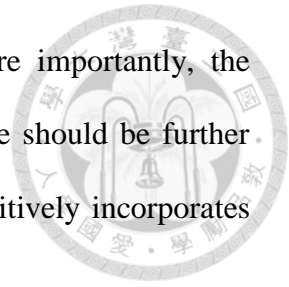
a language to be used so naturally as in the case of “葡萄”, or it is just code-switching for communicative need? A quantitative evaluation should be proposed to answer these questions.

Additionally, though the motives for words to emergent have been proposed from various points of view (Chao, 1976; Keller, 1994; Hudson, 1996; Aitchison, 2001 ; Hickey, 2003; Love, 2006; Halliday, 2007; Milroy, 2008), but they are not investigated in depth, and cognitive factors that are involved to influence the words' being used as well as comprehended should also be probed. Thus, current study is aimed to adopt both qualitative theoretical insights and quantitative experiential evidences to provide suggestions in constructing lexical resources that can reflect mental lexicon. I assume that lexicon should include comprehensible aspect and performance aspect. Senses and forms bearing the senses, the lemma, can be best observed in comprehension test; nevertheless, the real natural performance can only be observed in real language usage, so in this study, I will highlight this aspect. Though the lexicon utilized in comprehension may be larger than those used in performance, the lexical items used in language performance should be fully comprehensible for they are the resources for initiating daily communication in life. In order to tackle this issue, the key step is to understand the quantitative index for words that are newly diffused, words that have been already conventionalized, and words that are inactivated. Thus, instead of collecting only certain type of words, present study observes linguistic items that have existed for over 50 years, that have coined for around 50 years, and that have newly diffused for about 10 years in order to realize the driven factors for a lexical item to enter into our lexicon, and to be used in real life communication.

1.2. Purpose

There can be found many linguistic insights in the literature of neologisms studies (Fischer, 1998; Hsu, 1999; Kjellmer, 2000; Metcalf, 2002). Though linguistic insights are rich, there has been less experimental or empirical evidence of the arguments. There are some quantitative observations or formula proposed (Chang & Ahrens, 2008; Wang, 2010; Altmann, 2011, 2013; Antoinette, 2013; Kerremans, 2015) to delineate life stages and to predict whether a word may be survived after being coined. However, their definitions on survival are inappropriate in that, as I will argue in this study, current study supposes that once a lexical item is coined it is existed, the only difference locates on whether it will be passed down to be used in the communication of next generation. Meanwhile, even a lexical item is less stabilized in use of contemporary generation it is still with the potential to be revived in the use of future generation. In addition, targets perceived in these studies, except Chang and Ahrens (2008), are mainly easily fluctuated nouns. Besides, the word included are without generality but biased in either only words on Internet or words in textbooks or dictionaries. Among them, the authors who have approached to propose quantitative index in defining words are Chang and Ahrens (2008) as well as Wang (2010). Chang and Ahrens (2008) have proposed to use normalized frequency within a year to judge whether a once diffused new word is conventionalized in using or is failed to be captured. However, current work argues that normalized frequency within a year cannot really reflect conventionalization, for it may be result of temporary burst. Being conventionalized or not should be viewed from more longitudinal temporal information and cross-timing points' stabilization. Moreover, the Constant U proposed by Wang (2010) in evaluating textbook words may not only be used in defining whether a word is activated or not, but should also be used

to observe the stabilization developmental trend of a word. More importantly, the linguistic factors driven behind this surface behavioral constant use should be further explored for the deeper understanding of how mental lexicon cognitively incorporates new words.



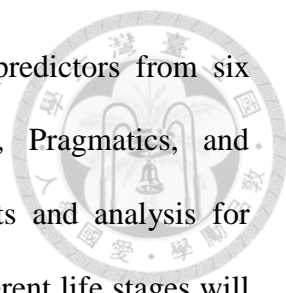
Overall, based on previous proposals and limitations, the purposes of present work are set to explore following issues:

1. To sketch linguistic characteristics for words from different temporal points
2. To sketch linguistic characteristics distinguished conventionalized and diffused words
3. To build prediction models for foretelling possible future life of diffused words
4. To qualitatively analyze competitions among words from the same synset
5. To propose suggestions on including stabilized words into lexical resources

In order to solve these proposed issues quantitative and qualitative methods are both employed. The generality and temporal information of included target words are tried best to be manipulated based on available resources. Various linguistic variables are also proposed in order to find out linguistic factors driven behind behavioral quantitative performance of lexical items.

1.3. Organization

In order to achieve the set aims chapter 2 will discuss related studies on understanding life of lexical items. Insights from qualitative studies in Historical Linguistics and Lexical Semantics to quantitative profiling with Corpus Linguistic investigations and Computational Linguistic prediction models are all reviewed in order to assist in constructing hypothesis and research method in this study. Chapter 3 will introduce scope of current work, resources in use, and proposed predictors in detail. The



selections on 384 lexical items and considerations on proposed predictors from six linguistic aspects: Phonology, Morphology, Syntax, Semantics, Pragmatics, and Sociolinguistics are discussed. Chapter 4 will illustrate the results and analysis for answering the five proposed issues. Characteristics of words in different life stages will be probed qualitatively and quantitatively. These findings will be applied in building models for predicting conventionalization. In addition, qualitative analysis in between words competitions from the same synset will be discussed. General results will be applied in proposing new updates for including stabilized words into lexical resources. Chapter 5 will summarize findings in present study and propose possible future direction in follow-up investigations.



Chapter 2.

Literature review

A lexical item has so many aspects that can be evaluated from many perspectives. It can be analyzed on its historical origin, semantic extension, morphological formation, syntactically stigmatic relation, phonological interface, social connotation, or pragmatic knowledge. These aspects may influence its inclusion in lexicography, its teaching design in language learning, or its weighting in natural language processing. The approaches to understand a lexical unit can be quantitative and qualitative. However, there are few investigations incorporating multiple aspects to understand factors influencing stabilization of general words holistically, namely, the living continuum a word owns. This chapter reviews related studies from different domains and illustrates the corresponding findings among them in order to sew together these insights for further understanding of factors influencing the fluctuation of lexical items.

Given the fact that what we concern is around the occurrences of a lexical item along with temporal fluctuation, how Historical linguistics tackling the diachronic development of lexical items and how Lexical Semantics investigate the synchronic emergence of new expression are reviewed. Following that, studies paying attention on “life cycle” of lexical items as well as proposed features for words to be “survival” are introduced, whose insights are further utilized to guide the design of current study. In the third part of this section, the application on lexicography, natural language processing, and language teaching from understanding of mental lexicon will be discussed.



2.1. Qualitative Discussion from Historical linguistics and Lexical Semantics

In the field of Lexical Semantics and Corpus Linguistics, neology is a popular issue. Neology can be probed from three aspects (Antoinette, 2013): semantic neology, lexical neology, and grammatical neology. Semantic neology studies new emergent senses. The identification of it can be probed with collocational environment because the new sense of an existed word would collocate with different words from its original sense. Lexical neology is about the formation of new words, which can be identified in diachronic corpus. Grammatical neology probes issues like conversion, so it can be studied by post-processing with parts of speech tagging. In addition to the birth of a word, the death of a word, the reason why certain words can survive over decades and why certain words die so early are all important issues which should be paid attention, too. Namely, the life cycle of a word, its birth, its settling-down, its death, or its re-birth, should shed lights on the secrets of words, and the cognition of human. To approach these issues we cannot just focus on newly coined words, but have to bring our attention to the words stored in our lexicon over generations. Thus, the discussions from Historical Linguistics on how words sustained and expanded their meanings and functions as well as how words fluctuate are also introduced with the attempt to explore the life journey of lexical units.

2.1.1. Historical Linguistics: Grammaticalization, Degrammaticalization, Lexicalization, and Exaptation

Grammaticalization is a popular issue on verbs from Historical Linguistics' view.



It captures the information from both diachronic and synchronic aspects. Context also plays important role in language development (Fischer, 2000; Jucker, 2010). For diachronic parts, instead of proposing rules, the focus on grammaticalization path gives insights into changing tendency. For synchronic parts, it focuses on emergent properties of languages. Metaphorical Extension Approach and Invited Inferencing Approach are two approaches commonly adopted in studying grammaticalization. They are proposed separately by Heine and Traugott, and are termed as this by Evans and Green (2006). Metaphorical Extension Approach is concerned about the metaphorical extension of human analogy ability which contributes to the development of new grammatical concepts for an expression. Heine, Ulrike and Friederike (1991) propose a metaphorical source domain hierarchy:

PERSON > OBJECT > ACTIVITY > SPACE > TIME > QUALITY

Invited Inferencing Approach, on the other hand, is from the perspective on conventionalization of pragmatic inference, and on subjectification increasing to account grammaticalization as pragmatic enrichment with constraints from previous meanings rather than bleaching of old meanings.

Since the new meaning is pragmatically inferred from old meaning, the inferred meaning can be further distinguished into three levels: invited inference (IIN), generalized invited inference (GIIN), and coded meaning (Traugott and Dasher, 2004). As long as the old meaning is accessible, the new meaning is just the invited inference that is derived in combination with the discourse context, so in this stage the new

meaning has not yet been coded. However, if only the new meaning is accessible in certain context, the earlier old meaning(s) of the item would not make sense in this context, then GIIN can be considered to have become semanticized as a new coded meaning, as shown in Figure 1. This actually shows some insights on how semantic neology is settled down.

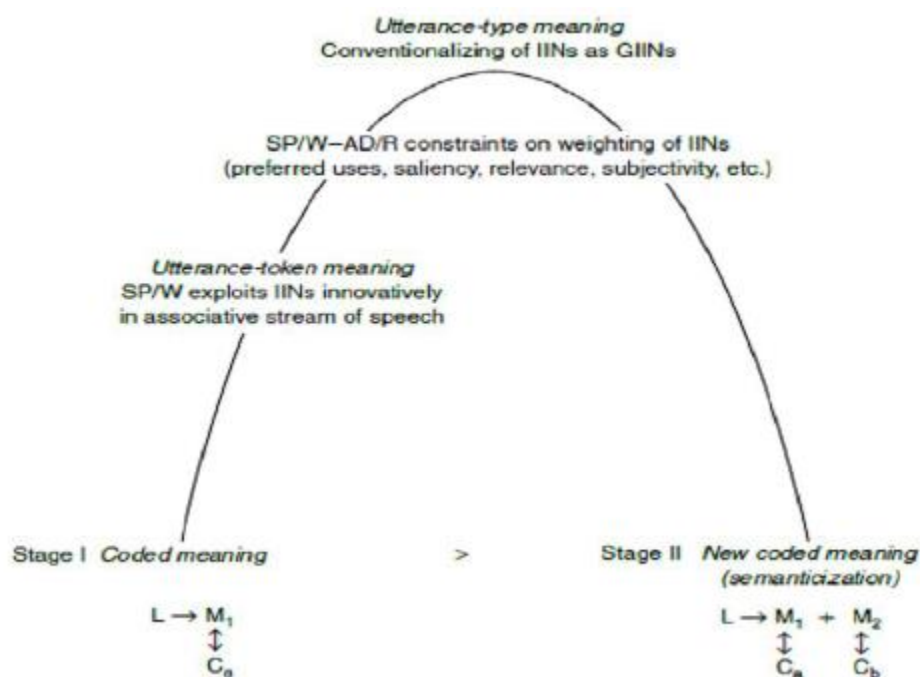
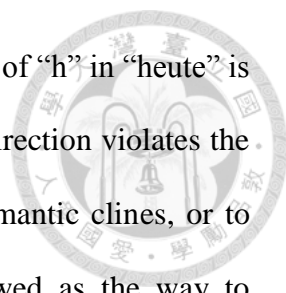


Figure 1 Model of the Invited Inferencing Theory of Semantic Change (Traugott and Dasher, 2004)

These two approaches do not contradict to, but complement with each other because the former one indicates the direction of grammaticalization, and the latter one illustrates the process of grammaticalization (Traugott, 2003).

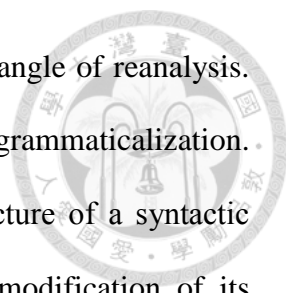
Some counter-examples to grammaticalization are: degrammaticalization, lexicalization, and exaptation. **Degrammaticalization** stands for two different situations. The first refers to prototypical cases of end-stage in grammaticalization. For



example, the development into unanalyzable segments as in the case of “h” in “heute” is one of such case. The other situation is the case that the changing direction violates the unidirectionality of grammaticalization, namely the direction of semantic clines, or to upgrade inflectional or derivational forms. **Lexicalization** is viewed as the way to enrich the lexicon. Lipka (1990) has defined it as “...the phenomenon that a complex lexeme once coined tends to become a single complete lexical unit, a simple lexeme.” It may employ “conversion” as a strategy to use grammatical items as other parts of speech as in “to up the ante,” ”F-words,” ”calendar,” “forget-me-not,” or ”laser.” This process is also called as “univerbation,” for it loses the character of a syntagma to a greater or lesser degree. For example, “arise” from ‘on’ + ’rise’ now functions as monomorphemic non-compositional elements, and belongs to major class (noun and verb), or “already” (‘all’+’ready’) belongs to minor class. Lehmann (1995) or Wischer (2002) indicate that there is intersection of grammaticalization and lexicalization for lexical phrases must be first lexicalized (frozen) before they go into grammaticalization.

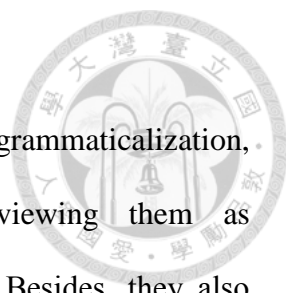
Exaptation refers to the situation that a form is given with a new function. It is widely discussed in studies on language evolution (Hurford, Studdert-Kennedy and Knight, 1998), and in historical morphosyntax (Lass, 1990, 1997; Vincent, 1995; Norde, 2002). Traugott (2004) has pointed out that some terms have also been used with reference to similar phenomena, such as “regrammaticalization” (Greenberg, 1991), “functional renewal” (Brinton and Stein, 1995), “degrammaticalization” (Norde, 2002; Heine, 2003), and “hypoanalysis” (Croft, 2000). Lass (1990) earliest identified the phenomenon of exaptation by describing the three possibilities of a form that loses its function, or is marginal within a system:

- a) be lost, b) be kept as marginal, c) be reused for something else (= exaptation).



We can compare exaptation with grammaticalization with the angle of reanalysis. Reanalysis has been viewed as significant mechanism in triggering grammaticalization. It is defined as “a mechanism which changes the underlying structure of a syntactic pattern and which does not involve any immediate or intrinsic modification of its surface manifestation” in Harris and Campbell (1995). Vincent (1995) highlights that grammaticalization is to give a lexical item a new form and a new function relative to the original system, whereas in exaptation the form, is still kept with new function given. In this way it seems that grammaticalization is the co-variation of form and meaning, and exaptation is to give the old form new functions. Brinton and Stein (1995) propose that exaptation can also be found in syntactic level. From the discussion on the “conclusive perfect” HAVE + PP + object construction, e.g. “I have a letter written,” and the development of the perfect, e.g. “I have written a letter.” They propose that in Old English the two constructions were in competition. The perfect became regularized in the sixteenth century, while the conclusive perfect was in marginal. The latter then reemerged in the seventeenth century with new constraints. This is called by Brinton and Stein as “functional renewal”. This is different from Meillet’s “**renouvellement**” as reviewed in Traugott (2004). “Renouvellement” refers to two forms compete for the same function, and the older one is replaced by the newer form, e.g. in English the replacement of negative “ne” by “not.” The “functional renewal” is to use an old construction in a different new way, as quoted here:

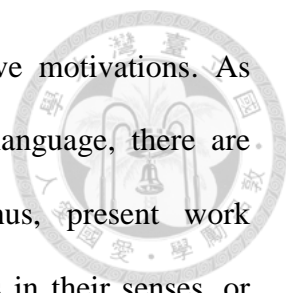
“It is the retention or revival of an existing syntactic form with a new or renewed function ... In functional renewal, an older form makes a resurgence with a meaning which is new, has been lost, or was on the decline.” (Brinton and Stein 1995: 34)



From above discussion on grammaticalization, degrammaticalization, lexicalization, and exaptation I propose that instead of viewing them as counter-examples, they reflect different stages of words' change. Besides, they also imply the relation between form and meaning. Grammaticalization is the long-living secrets for lemma. It derives pragmatically enriched functions from existing senses. Degrammaticalization and Lexicalization can be perceived as the birth of new sense or function. Exaptation illustrates the reviving of old forms. The reviving mechanisms include: an old lemma +brand new sense (dissociated with original sense), an old lemma + its original sense, or the reviving of one of less significant usage that has once appeared in past but became marginal for competition with another usage. Meillet's "renouvellement" illustrates another angle in understanding the competition of lemmas in the same sense class. Examples for "life cycle" of lexical items are going to be deeper exemplified in 3.1.4 with uniting other views from different linguistic branches on life of words.

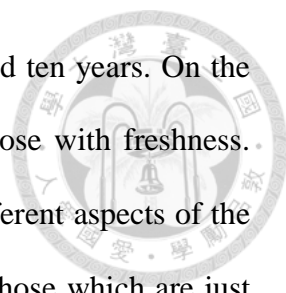
2.1.2. Lexical Semantics on Neology

Historical Linguistics provides insights into the semantic development and reviving of stabilized senses and lemmas. It describes how a lexicalized frozen word, a coded meaning, leads its life after settling down. However, how these expressions are born has not yet been fully answered. Neologisms, which may be newly-coined word forms or new senses of an existing word form, have been constantly appeared (Algeo, 1980; Lehrer, 2003). Thus, Lexical Semantics' study on Neology provides insights on how words are born. However, current study assumes that compared with contemporarily recent emergent expressions, early words that have started their lives



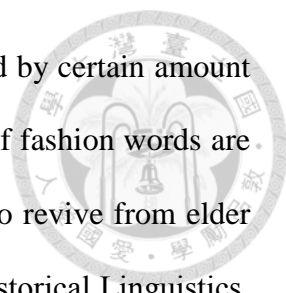
since emergence of language should own different communicative motivations. As indicated in Wang and Minett (2004) with the development of language, there are increasing needs to express more complicated meanings. Thus, present work hypothesizes that later new born words bear richer presuppositions in their senses, or they bear more unique senses to compensate or balance semantic network. For example, “打臉” in Chinese uses two syllables to concisely lexicalize the complicated concept that “A person says or does something that contradicts to what he or she has said.” In this case, this verb lexicalize an experience in our flux of life, and it owns unique status in this lexicalization. Wang and Minett (2004) adopt computational linguistics method in simulating how first word is emerged. They describe stages of emergent words as: individual innovation→diffusion→language acquisition. Present work proposes that the stage of “language acquisition” means that expressions are recognized to enter into the lexicon, for they are conventionalized and can be passed from one generation to the next generation.

Recent studies on Chinese neology have rich insights in probing issues about the paths new words are born, the rules to coin new words, the characteristics of new words, the reviving of old usages, or comparisons on new born words; however, as indicated in Chou (1995) the reasons and conditions for new words to be born, the reasons for lexicalization and for not being lexicalized, and the diffusion or step down of expressions should all be deeper conceived. Besides, the definitions on new words in related Chinese neology studies remain unclear. From the way previous studies illustrate their analysis, it seems that they focus more on those who are diffused ones. For example, Wang (1992), Yao (1996), and Xu (1999) stress that neologies should be those who have been used for a long time, used widely, and used across registers. Xu (1999)



has proposed that we should observe words that have existed around ten years. On the contrary, Guo (1996) has highlighted that new words should be those with freshness. Present study supposes that each of these studies is highlighting different aspects of the life stages of new coming expressions. The fresh words should be those which are just newly born, namely, the individual innovation named in Wang and Minett (2004). What Wang (1992), Yao (1996), and Xu (1999) have identified should be those more diffused. The diffused ones are actually the focus in studies of neologisms. Neologisms are not just about new words but those who have lost their nonce status in formation and is becoming established in language and used by most members in speech community (Fischer, 1998; Hohenhaus, 2005).

In addition, the words that have been included in analysis are more about synchronic observations. Xu (1999) and Kim (2006) both include new word lists that Ministry of Education R.O.C has updated in online dictionary in 1997. Nevertheless, evaluating these words from using phenomenon in contemporary and from linguistic angle the words included in the lists are not really conventionalized and being used till recent years. Hence, it is valuable to deeper analyze the characteristics of these words that have once been risen as new comers, but cannot be conventionalized into our lexicon over years. On the other hand, it should be noted that fashion words have broader range than the new words we discussed here. Xu (1999) proposes that fashion words are those who may or may not be diffused into the whole society, and they may sporadically come into fashion in certain external events. For example, "力挺" may come and go follow the start and end of elections. Thus, the concepts and features of fashion words should be further delineated in present work. Fashion words should not be considered as the source of emerging new words. Fashion words should signify the

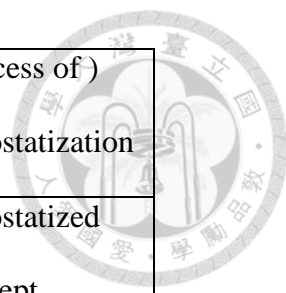


diffusion of words because it is due to the fact that they are accepted by certain amount of people, then we can claim them as “fashion”. But, the members of fashion words are not limited to those who are newly born, but also include those who revive from elder usages, which is similar to situation of “Exaptation” discussed in Historical Linguistics. Discussion on reviving has been touched in Kim (2006). For example, “倒閉” reappears in Mainland China after Mainland China adopts the reform and opening-up policy. Additionally, reviving may come from lemma with brand new sense as in the cases of “跳槽,” or “兵變” (Kim, 2007). However, present study further claims that words come back with original meanings may not just because of external social events, but may come back for its preciseness of internal linguistics meanings as in the case that “中肯.” It is popularly used in comments in PTT. There may also be cases that the reviving meanings have once existed in the past, but then become marginal and revived later. Hence, though diffusion can be illustrated by the fashion words, the fashion words are consisted of members representing different aspects.

In addition to focusing on characteristics of neologisms, general picture and driven factors on life stages of words have also been proposed in some studies. Among them Schmid (2008) and Kerremans (2011) take consideration on structural, socio-pragmatic, and cognitive perspectives as well as bring out corresponding three stages (creation, consolidation, and establishing) in these three perspectives as shown in

Table 1.

Perspectives / Stages	Structural perspective	Socio-pragmatic perspective	Cognitive perspective
creation	(product of) nonce-formation	(product of) nonce-formation	pseudo-concept

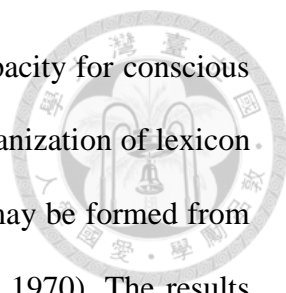


consolidation	stabilization	spreading	(process of) hypostatization
establishing	lexicalized lexeme	institutionalized lexeme	hypostatized concept

Table 1 Proposed Stages of Words (Schmid, 2008)

Kerremans (2011) delineated stages of conventionalization with reference to observation on patterns of new words per month. The proposed four stages of conventionalization are: (1) Non-conventionalization (2) Transitional conventionalization (3) Recurrent semi-conventionalization (4) Advanced conventionalization. The details of her hypothesis will be further illustrated in section 2.2.2 and section 3.3 of methodology. Schmid (2008) delicately proposed the driven reason of lexicalizing new words from conceptual perspective. The establishment of new words in mental lexicon can be probed from two frameworks (Schmid, 2005). The first one is “hypostatization,” which notifies that the coin of a word means the existence of referent signified by the word (Lipka, 1977). In addition to this philosophical and semantic aspect, the second framework roots in neuro-psycholinguistic views the organization of mental lexicon as network and requires entrenchment for activation.

“**Hypostatization**” is considered as “concept-forming power of the word” by Leech (1981). Nouns are considered to be more powerful in this aspect because they can profile concept of concrete things (Langacker, 1987). But, those who really create concept are not those concrete nouns that denote already bounded objects but the event nouns or abstract nouns carve segment form the flux of ongoing events (Schmid, 2005).



Mental lexicon is “the cognitive system that constitutes the capacity for conscious and unconscious lexical activity” (Jarema & Libben, 2007). The organization of lexicon is proposed to be a network (Aitchison, 2003). Meaning networks may be formed from experiences as shown from word association experiments (Jenkins, 1970). The results from association experiments show that the most common linking relation in network is coordination, collocation, superordination, and synonymy.

The process of activation and organization of storage in mental lexicon have several different proposals (Caramazza, 1997; Starreveld, 2004; Harley, 2005; Warren, 2012). One of widely discussed psycholinguistic model among them proposes that there is two-stage in lexicalization. The first stage produces lemma with semantic and syntactic information, and the second stage goes with adding in the phonological information to produce lexeme (Fromkin, 1971). In addition to overall organization, there is also categorization based on characteristics of words. The categorization on word classes has been evidenced in speech errors and aphasia diagnosis. Nouns are special in their inclusion of levels (G.A. Miller, 1990). Meronymy or partonymy is an important relationship for them. Adjectives may be categorizing into ascriptive and pertainyms (Gross et al., 1989). Verbs, on the other hand, own layers in hyponym and superordinate that are different from nouns’ layering (Aitchison, 2003).

Based on such assumptions on representation of mental lexicon, it becomes clear that how the new born words incorporated into it is an issue. Aitchison (2003) has pointed out that coining new words is just similar to have a tool box inside our mind, and we choose to use which tool to lexicalize our thought may depend on: (1) frequency of usage, (2) sound structure, (3) extent of modifying the existed one to produce new words, (4) tend to adopt suffix bearing consistent meanings. Thus, it is not surprising

that most of new words additions or re-combinations of existed lexical items.



2.2. Quantitative Analysis on “Life Cycle” of Lexical Items

In addition to the birth and recycle of lexical items from the results of qualitative analysis, there are some quantitative investigations around the “Life Cycle” of lexical items. Approaches in corpus-based studies in delineating stages of lexical items from either Newspaper or Web as well as computational methods in modeling fluctuation of words by external social indexes of internet community all shed lights in possible living appearances of lexical items, and illustrate external influences on words. Experimental examination on proposed theoretical predictors is also reviewed in accordance with quantified formula in evaluating constant use to be reference in setting scope in current study.

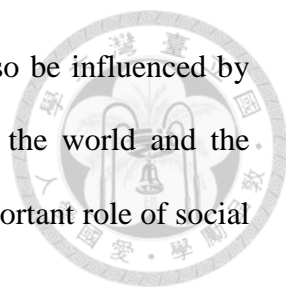
2.2.1. Analysis on “Life Cycle” of Different Lexical Items

In the field of Lexical Semantics and Corpus Linguistics, tracking the changing status of novel words across decades has become a research topic (Renouf, 2013). It is believed that the deeper understanding of the mechanism underlying the *life-cycle* of a word, including its birth, settling-down, death, even its re-birth, etc., will shed new lights on the coevolution of language and culture, as well as development of human cognition.

Previous research touched upon this issue mainly focused on the frequency aspects in diachronic dimension. Altmann et al. (2011) hold interest in the mystery on why some words live long life, while the others “died” soon. They propose that along the historical time scale, word frequency is the factor of word success; however, in short time scale the frequency of a word is determined by the amount of being used by

different individuals (**indexicality**) and the range of being used in different topics (**topicality**).


In addition to revealing the secret of being survived, the secret of brewing the birth of a word was also probed by analyzing the rising words. As defined in Altmann et al. (2011), the rising words are words that are not used during the first year of the group, but are consistently used for at least some years thereafter. The discussion on birth of rising word is similar to the topic used to be covered in the discussion of neology. However, Altmann et al. (2011) aimed to probe deeper into the understanding of what brews the birth of a word. They tackled this issue by analyzing the rising words: **product words (P-words) and slang words (S-words)**. The difference between product words and slang words is that the rise of P-words (e.g. Iraq) is driven exogenously by events that are external to the group, such as product releases or political policies, but the use of S-words (e.g. lol) is more endogenously influenced by the social values and language patterns of the communication group. Slang words are different from other words for being used to “establish or reinforce social identity or cohesiveness within a group, or with a trend or fashion in society at large” (Eble, 1996). The result shows that for general words, if they are less frequently used by different individuals (indexicality) and in different topics (topicality), then they will decline in frequency, but, interestingly, different from the fate of general words, even with low indexicality and low topicality the frequencies of P-words and S-words still rise. The rise of P-words indicates that exogenous forcing, social event, is efficient, and the rise of S-words shows that the endogenous forcing, social value, is also efficient comparing with the fate of words in general, which is predicted to be dead if they are less used by different individuals and in different topics. On the other hand, it also shows that it is



just similar to the life of human beings the life of words should also be influenced by the events in the world. The strong influence from the events in the world and the endogenous social value are identified. This result highlights the important role of social events and the social value in influencing the rise of words.

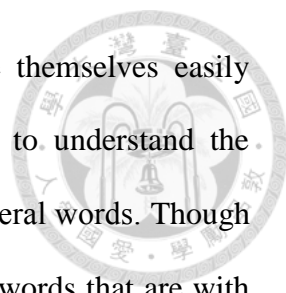
Though their study revealed some of factors influencing the life of new words, the whole life stages of words are still mysterious. Renouf (2013) uses 1.2 billion words from UK mainstream newspaper texts spanning from 1989 to 2011 to understand the life-cycle of proper nouns (e.g. *Arab Spring*), which is similar to the product words in the study of Altmann et al. (2011) that are also easily influenced by the events occurred in the external world. The result indicates that the life cycles of words should include: *Birth, Increase in Frequency, Orthographic Adjustment, Lexical Productivity, Creativity, Settling Down, Obsolescence, Death, Semantic neology, and Re-birth or Revival*. However, present work argues that Renouf (2013) has mistaken the *factors* influencing birth, settling-down, obsolescence, death, and re-birth as the *stages* of life cycle of words. For example, increasing in frequency may be the indicator of being born and settling-down. The threshold of what kind of increase in frequency should be called settling-down needs further discussion. Besides, just like the situation of rising words discussed, the increase in frequency may be influenced by different factors. Meanwhile, orthographic adjustment may be one of the reasons for the fade-away use being re-born with new appearance. In addition, Lexical Productivity and Creativity should be viewed as the characteristics of the words that influence their birth, death, and length of life span. Finally, the semantic neology should be one of the paths for reviving of lemma.

On the other hand, in addition to clear-cut stages of lexical items, Kerremans (2015) with her construction in automatically crawler for detecting non-sense new words has



highlighted the part in conventionalization. The proposed four stages of conventionalization are: (1) *Non-conventionalization* (2) *Transitional conventionalization* (3) *Recurrent semi-conventionalization* (4) *Advanced conventionalization*. Kerremans illustrates these 4 stages by using 44 neologisms retrieved from the Internet between October 2009 and January 2011. The **transitional conventionalization**, which is also referred as “Topicality” in her study is different from what Fischer (1998) has defined, “a word is used in connection with current affairs for a short period of time.” Kerremans (2015) defined it as momentary conventionalization of lexemes contributed from extralinguistic events, as in with a sudden burst in overall frequency. In her definition fashion or vogue words are cases of transitional conventionalization. This is similar to the diffusion stage proposed in present thesis. This phenomenon highlights the necessity of long-term observation on the frequency cycle of words. As stated in Kerremans (2015), “It depends on the degree and duration of an item’s topicality transitional conventionalization may segue into advanced conventionalization. Thus, it requires a high frequency of occurrence within a longer time span.” The emphasis on conventionalization and in recognition on reviving corresponds to our assumptions; however, her investigation is limited in the lack of quantifying index in capturing the differences among the stages as well as the type of new words being investigated.

Meanwhile, in addition to the concerns mentioned above, the targets chosen in these studies also need rethinking. Previous studies that have probed life of words only focus on the words that are easily fluctuated with the changes of the world. From these studies we can gain the insight that many words we used in daily life are born with the occurrence of certain event, so their life would be influenced by such social background.



Nevertheless, the words intentionally chosen in these studies are themselves easily changed with the momentary events on the world, so it is hard to understand the complete picture of words' life cycles, which is applicable to all general words. Though it is inspiring to realize that the increasing frequency goes with the words that are with increasing indexicality and topicality, but why these words can have increasing indexicality and topicality, and other factors involving in influencing the lifespan of words as well as in stabilization remains unknown.

2.2.2. Quantitative Profiling on “Life” of Lexical Items

In profiling “Life” of lexical items quantitatively, Chang and Ahrens (2008) proposed threshold for deciding whether a word is “survival” or “fail” with reference to developmental trend in slope as well as normalized frequency. In his study on non-sense-neologism fashion verbs, Chang and Ahrens (2008) collected the year-by-year frequencies in UDN (United Daily News) from 1996 to 2006. They focused on deciding a word is survived or is failed, whose idea is similar to the process from diffusion to stabilization or from diffusion to the flash in pan type inactivation. The threshold to decide a word is conventionalized or being lost from use as a flash in pan includes normalized ratio in 2006 and slope of the normalized ratios throughout the years. The retrieved frequency was normalized to the frequencies per 10,000 characters. Words' actual survival or failure is based on normalized ratios in 2006. A word is failed to survive when its normalized ratio is less than or equal to 0.3 in 2006 (e.g., 哈草, *ha1 cao3*, ‘to smoke,’ normalized ratio=0.11). A word is considered to be “success” if it is with normalized ratio greater than 3 (e.g., 抓包 *zhuo1 bao1*, ‘(to be) caught doing something,’ normalized ratio=4.24). In addition to normalized ratio developmental tendency is considered, so only words with a slope less than -0.06 would be counted as

failures. As shown in following table.



Failure		Survival	
$NR \leq 0.3$	$0.3 < NR \leq 3$		$NR > 3$
	$Slope \leq -0.06$	$Slope > -0.06$	
23	1	29	24

Table 2 Threshold for "Survival" or "Failure" of Words (Chang & Ahrens, 2008)

However, frequency can only highlight the activation aspect of a lexical item. A lexical item may be high in total frequency due to the fact that it has been used frequently within a short period, which does not signify its stabilization in use for it may be a flush in the pan. Hence, when understanding conventionalization we should not just focus on the activation aspect, but also take diachronic temporal information into consideration, which can be captured in the formula proposed by Wang (2010). With more temporal points included the measuring is more stringent. By calculating seasonal mean frequency divided by standard deviation of frequency, which is called as **Constant U** by Wang (2010), sudden burst in frequency would not be viewed as stabilization for with more fluctuation in the passage of time the value of Constant U would become smaller.

Though Wang (2010) has proposed the way to understand stabilization in text book selected Chinese Words with reference to Newspaper Corpus, the target words are the easily fluctuated nouns¹ and there is lack of closer analyzing on linguistic factors driven behind. Chesley and Baayen (2010) has conducted a study to propose prediction model for entrenchment of borrowings by predicting loan words' 10-year later

¹ Wang has also proposed exploratory study on verbs. However, there is no response from my written e-mail for permission in taking reference on this related study.

frequency with their 10-year ago frequency in French newspaper corpus. This is similar to the aim of this thesis; however, its scope is limited in loan words and the features proposed (*frequency, dispersion, sense pattern, cultural context, donor language*) are hard to be reduplicated, for there is lack of ideal available diachronic newspaper corpus for Chinese. For exploring not just loan words Kjellmer (2000) and Metcalf (2002) separately proposed theoretical hypothesis on conditions influencing words' stabilization on being passed down over generation. Kjellmer (2000) presented thirteen conditions in assessing potential words. These conditions have been reselected and divided into five categories: semantic, phonological, morphological, and graphematic conditions, and others, such as prestige. For Semantic conditions Kjellmer (2000) emphasizes on "Pre-existing semantic pattern" such as the suffix "–able: capable of being V-ed." However, factors like this may not be suitable for linguistic context in Chinese.

Metcalf (2002) proposed FUDGE scale with its assessing probability is to rank new words from level 0 to level 2 in each factor and sum up the total scores in the end. With the higher the scores are, the more likely the new words are to survive over time. The FUDGE is acronym of its five conditions: (1) *frequency of the words* (2) *unobtrusiveness: a successful word should not be exotic or too cleverly coined* (3) *diversity of users and situations* (4) *generation of other forms and meanings, namely the productivity of the word* (5) *endurance of the concept, related to the concept's reference to a historical event*. These factors have also been reviewed in Kerremans (2015) and being recombination as well as refined into six hypotheses in her qualitative analysis as shown in

Table 3:

H1: Semantic ambiguity.

H2: Dominant or disproportionately high use in metalinguistic mode inhibits conventionalization.

H3: A first or frequent use in more formal types of source.

H4: The authority or prominence of the coiner and first users promotes conventionalization.

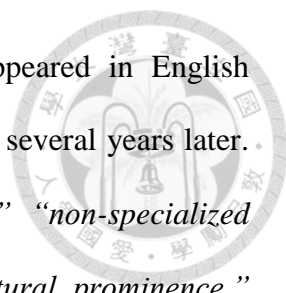
H5: The newsworthiness of the represented concept or its salience in society promotes conventionalization.

H6: The early development of syntagmatic lexical networks promotes conventionalization.

Table 3 Proposed Hypotheses of Influencing Life of Words by Kerremans (2015)

What Kerremans (2015) has proposed have some correspondences to previous two proposed models, but some hypothesis are hard to be measured and some may be lack of concrete stands, so in the methodology part the proposed hypothesis from Kerremans(2015) would be re-evaluated and refined. Meanwhile, although Chang and Ahrens (2008) have evaluated predictors from Kjellmer (2000) and Metcalf (2002) empirically in predicting Chinese novel verbs, here needs some reconsideration on experimental design and on selection on appropriate corresponding linguistic behaviors as well as addition on observing angle from linguistic view. Barnhart (2007) has used multiplication on factors proposed in investigations (Sheidlower, 1995; Barnhart, 2007; Hargraves, 2007; Metcalf, 2007) as indicator for understanding importance of a new word for including in a dictionary. The factors include: “*number of forms of target words,*” “*frequency,*” “*number of sources the target word occurs,*” “*number of genres a target word occurs,*” “*timespan a word has been observed.*”

Boulanger (1997) proposed 8 factors in her studies on comparing survived words

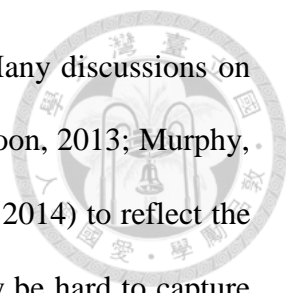


and fade-away words, which is defined by observing words appeared in English new-word dictionary in 1990 and its inclusion in general dictionary several years later. The proposed factors include: “*frequency*,” “*popular referent*,” “*non-specialized register*,” “*particular notional fields*,” “*variety of genre*,” “*cultural prominence*,” “*synonymous competition*,” and “*Taboo association*.” However, the decision on inclusion is solely determined by lexicographers’ looking-up in recording of dictionary, which is a relatively indirect method. Details on proposed features used in current study in understanding stabilization would be covered in Chapter 3.

2.3. Applications

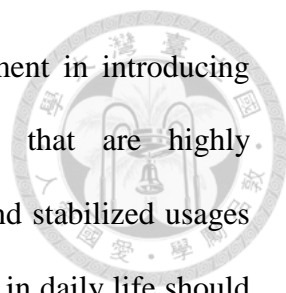
An adult may own a lexicon with around 50,000 actively used lexical items in it (Aitchison, 2003). Barnhart (1978) has claimed that there are nearly 500 new words recorded in dictionary every year. Metcalf (2002) proposed that in English there are 10,000 words coined in each day. The development in language acquisition include process like labeling task for symbolization, underextension, or overextension in broadening or narrowing meaning carried by form. This process may be similar to the way adults incorporate new words into mental lexicon. Mental lexicon is dynamic in its continually giving birth of new words and forming new connections, which is largely different from a fixed dictionary, so how to capture this dynamics and to reflect the collective mental lexicon in a speech community should be the goal lexicography aims at for further application in language teaching as well as resources for natural language processing.

Mental lexicon and dictionary may differ in the layout organization and stored contents. The layout for dictionary may be alphabetic instead of semantic relation based, though it is convenient for user consulting and it can be compensated by computerized



databases as well as by providing semantic related lexical items. Many discussions on lexicography also invite the points of view of semantic relations (Moon, 2013; Murphy, 2013; Chiara, 2013; Marie-Claude, 2014; Fellbaum, 2014; Polguere 2014) to reflect the organization of mental lexicon. The fixed number of dictionary may be hard to capture the real dynamic changes of mental lexicon because dictionaries may only record words fitting to requirements on frequency, use range, timespan, cruciality, and need in use (Sheidlower, 1995). Lexicographers capture neologisms by manually examining huge amounts of materials, but words included may not really useful by generations (O'Donovan and O'Neil, 2008; Cook, 2010). As what Samuel Johnson (1755) has said, "No dictionary of a living tongue can ever be perfect, since while it is hastening to publication, some words are budding, and some fading away." Thus, which words should be included and updated is immediately is an issue.

Natural language processing (NLP) studies have also turned their focus on neologisms detections (Cook, 2010; Kerremans, 2011), for it requires reliable lexicon for making judgment on unknown words (Cook, 2010). It is mainly developed in detecting morphological neology. For semantic neology it still relies on semi-automatic detection or more qualitative description. In identifying lexical items the combinations of words plays an important role (Giuliano, 1965; Choueka et al., 1983). These combinations include lexical expressions from compounds (black box), idiomatic expressions (kick the bucket), to compositional combinations with lexical restriction (handsome man vs. beautiful woman) (Stefan, 2004). Namely, it may include unit as multi-word expressions (MWE), multi-word units (MWU), bigrams and idioms. Though identification is important, which identified is meaningful to be learned by the machines for further application should be considered.



In the field of language teaching providing suitable arrangement in introducing useful words for communications is important, too. Words that are highly conventionalized should be included for its rootedness in lexicon and stabilized usages in daily life communication. Words that are instantly communicable in daily life should also be concerned by including those newly coined. These newly coined may fade away in a couple of months, but some of them may hold its significant status in lexicon to be used widely and spontaneously across generation. It will be too late to include them into teaching when next generation comes, so it is with necessity to have some standards facilitating in selecting contemporary newly born words into language teaching and learning.

Thus, lexicography, natural language processing resources, and wordlists for language teaching should own the capability to reflect contemporary communication. Current Study is aimed to provide reliable quantitative and qualitative references for selecting and arranging words that can reflect mental lexicon of language users. Differing from dictionary mental lexicon dynamically varies in progress. The easiness to acquire new word or new meaning from context has been testified (Clark and Gerrig, 1983). Besides, mental lexicon includes more information than denotative meaning, namely, the knowledge and experiences of a word is less captured in dictionary. Thus, it may take time for dictionary to catch up such information for all words, but it would be good to model the underlying mechanisms that how mental lexicon incorporates new words as the selecting conditions to incorporate words into dictionary, teaching plan, and resources for natural language processing.

Chapter 3.

Methodology



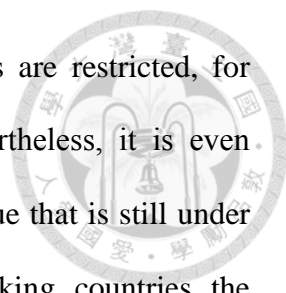
In this chapter scope of study, selected target words, resources for investigation, and proposed linguistic predictors will be illustrated in detail in each section.

3.1. Scope of Study

In this section the observation unit of target words in current study will be first illustrated. The highlighted type of words observed will be introduced in section 3.1.2. Section 3.1.3 discusses the observed source for understanding conventionalization of target words and underlying assumptions. Section 3.1.4 proposes life situations of words hypothesized in current study. Then, in section 3.1.5 operational definition on conventionalization and focuses on qualitative analysis will be introduced.

3.1.1. *Unit of Observation*

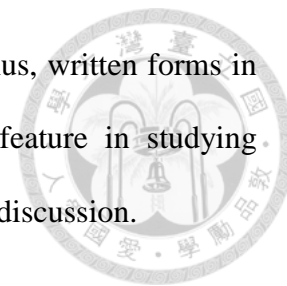
Word is the observation target in this study. It is difficult to give satisfactory definition of words for all languages (Cruse, 2001). It has been argued that most of lexical items in contemporary Chinese are compounds. Compounding is different from simple sequence of words in its formal characteristics and semantic property (Payne, 1997). The meaning of a compound is either more specific or entirely different. To verb compounds it could be noun incorporation or verb-verb incorporation. Different grammatical elements can be incorporated into a verb to adjust the verb's meaning. In noun incorporation it is more often to be object incorporation, where the object ceases to function as an independent argument. For verb-verb incorporation it may lose verbal characteristics or become derivational affixes. Incorporation is a way to be lexicalized



as in the case of “babysit,” but sometimes its syntactic behaviors are restricted, for example, we can say “fox-hunting,” but not “fox-hunted.” Nevertheless, it is even controversial in defining the unit of a word. Segmentation is an issue that is still under debate. Stefan (2004) has illustrated that even in English-speaking countries the “Knowledge-free” approaches that segmentize words based on white-space and punctuation can have inconsistency such as in variants like “whitespace,” “white-space,” or white space.“ Chinese owns unique properties in syntax and in lexicon (Tang, 1989; Yip, 2000).It may be objective and direct in measuring Chinese morphemes from the angle of syllables; however, single focus on phonetic form may lose the semantic importance in defining words. Thus, I would prefer not to limit in objective length on syllables of lexical items, but to take functional oriented angle in defining study target. Namely, this study is going to accept every lexical item that have been encoded as one single semantic bearing unit as the scope of study targets. Given upon the fact that expressions used in real language are diverse, instead of pre-defining the unit of observations, current study adopts functional angle in including expressions if only the expression can independently conduct functions in conveying meaning. Namely, adopting the definition of Stefan (2004) in defining an umbrella term for multi-word expressions (MWE), multi-word units (MWU), bigrams and idioms the targets included should be those “...whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon.”

Namely, a “word” is defined as a generic term for any minimalist independent construction that is used to convey communicative information, which is similar to what is proposed in Huang (2005). On the other hand, words encoded in written characters are important in natural language processing. Chinese characters own its unique status

in meaning and sound bearing from its diachronic development. Thus, written forms in Chinese should not be excluded from being part of important feature in studying Chinese language. Variants of written forms would be included into discussion.



3.1.2. Types of Target Words

In addition to observation unit, different from the focus of subjects in studying product words (P-words) and slang words (S-words) (Altmann et al., 2011), current study pays more attention on issues around predicates. The chosen target words are not nouns that are relatively easily fluctuated by external social factors. It is proposed that every language has two major parts of speech: verb and noun, but they are untidy at their boundary (Payne, 1997). Verbs and nouns have been proposed to be classified by anchoring their distributional syntactic behaviors or semantic classification. Givón (1984) has proposed that nouns like “rock, tree, house” are “most time-stable concepts,” and verbs like “die, run break” are “least time-stable concepts.” However, current study supposes that though core nouns or basic terms like kinship terms or body parts are long life, most of nouns may be easily affected by external factors as shown in the potential influence from indexicality and topicality delineated in Altmann et al. (2011), and the studies in Antoinette (2013) as well as Kerremans (2015) (P-words, S-words, proper nouns, or fashion words instigated from popularity). The target words they studied are more easily fluctuated by external factors. Hence, in current study the scope would like to focus on non-nouns.

Among these non-nouns, the target words included are mainly verbs. Verbs are relatively hard to be automatically detected for its significance in present core information in sentences (Aitchison, 2003; Cook, 2010). Though verbs are with significance in building sentences, verbs are three times less than nouns in amount (G.A.

Miller and Fellbaum, 1991). This gap in amount in Chinese has also been shown in current study's exploratory on data from Google Books Ngram Corpus (GBNC).

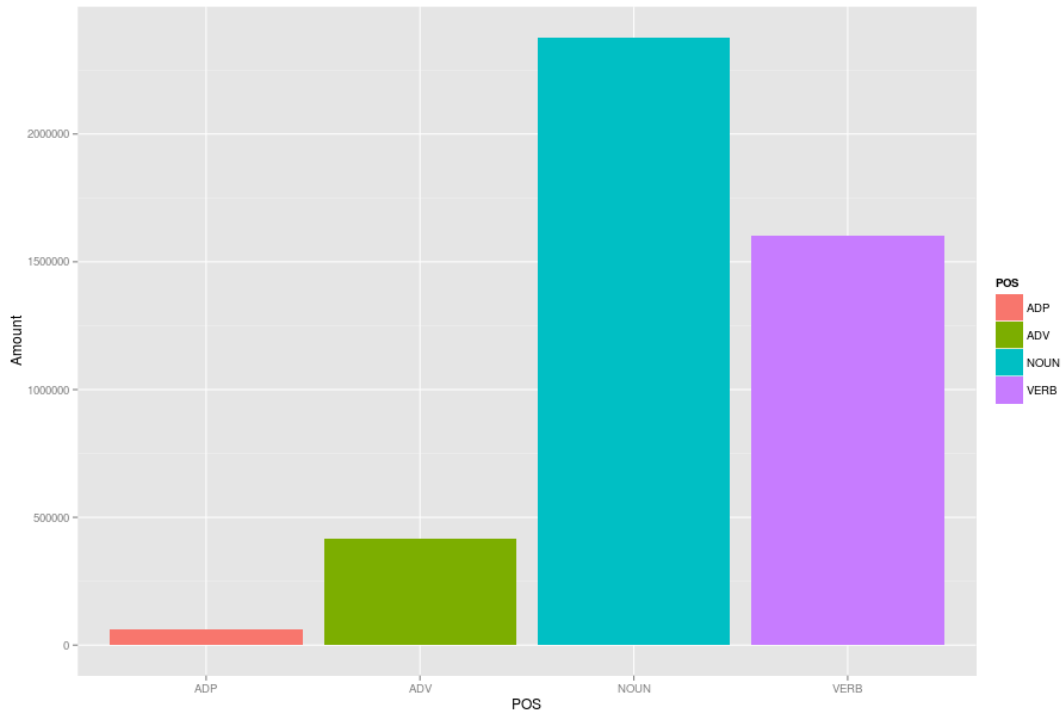
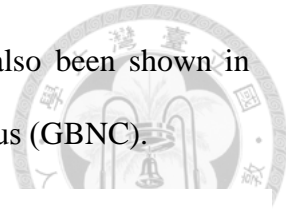


Figure 2 Distribution of POS in GBNC

From the exploratory of data it has further shown that the average “living span” for verbs is right skewed with extreme high outliers stocked around 400 to 500 years. This may imply the emergence of newly created verbs is rare, and the once created “elder” verbs are reliable in uses over century.

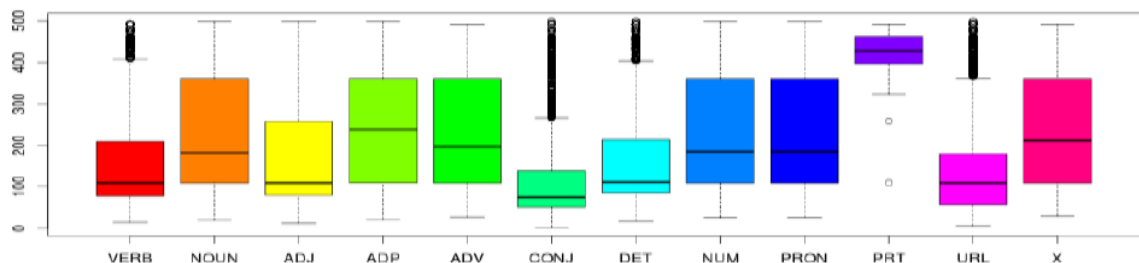


Figure 3 "Lifespan" of all Parts of Speech in GBNC

Additionally, verbs also hold special status conceptually. As proposed in Schmid

(2005) those who really create concept are not those concrete nouns that denote already bounded objects but the event nouns or abstract nouns carve segment form the flux of ongoing events. Thus, we assume that verbs should also hold special function in capture experiential presuppositions from daily life.

Meanwhile, borrowing constitutes a non-negligible portion of a language's lexicon (Masini, 1997). Abeille et al. (2003) estimated that .082% of all tokens in Le Monde treebank corpus from 1989 to 1992 come from borrowing. Borrowing in the definition of Thomason and Kaufman (1988) refers to lexical items that can be fluently adopted by speakers of recipient language. The donor languages for targeted loan words in current study include language system with written forms (English) and without written forms (SouthernMin). Thus, transliteration bearing by Chinese Characters and translation for borrowings are both included. Besides, lexical borrowings include idiomatic and multi-word expressions are also not excluded.

In addition to being aware of POS and borrowings, words originated in different time points are all included and compared equally in the contemporary synchronic corpus. Words with different number of senses and number of syllables are also included. Words with richer senses or being homophonic may mostly be those monosyllabic ones, and words with precise sense may mostly for those multi-syllabic because the increasing in component number in constituting the lexical item increases the anchoring on particular sense. For example, “打電話” with telephone in composing this lexical unit the sense of “打” can be precisely anchored at the sense of “calling.” The details in collecting different target words from different resources are discussed in section 3.2.

3.1.3. Potential Limitation and Corresponding Compensation in Current Study

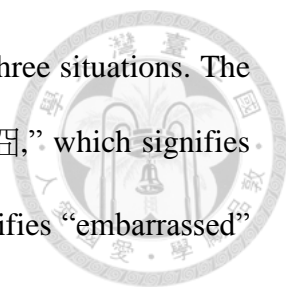
Current study does assume that lexicon should include comprehensible aspect and performance aspect. The lexicon has been widely explored from psycholinguistic angle. What is evaluated in psycholinguistics should be more comprehensible orientation. The lexicon utilized in comprehension may be larger than those used in performance, but the lexical items used in language performance should be fully comprehended by language users, or they cannot be the resources for initiating daily communication in life. Besides, the shortcoming for comprehension check is that there may be individual preferences and it is hard to be told how the lexicon is organized (Aitchison, 2003). With web corpora data and Chinese Wordnet data, present observational investigation has the chance to obtain spontaneous language performance that reflects collective lexicon shared by contemporary language users. The corpus used for equally compared target words is PTT corpus constructed by Liu (2014). Though it can only reflect synchronic usages around 10 years, it provides the opportunity to understand the diffused situation of words as well as is the epitome of living situations for those supposedly conventionalized words.

The other problem may meet in conducting current study is about sense and lemma. Both experimental comprehension check and data-crawling observational studies may meet “big dictionary effect,” which can be illustrated in homonyms like “must,” the elicited activation may be hard to anchor the exact activated meaning (Aitchison, 2003). Meanwhile, it is hard to anchor the activation level of knowledge in the word (Aitchison, 2003). When quantifying linguistic behaviors from big data of language in use the activated sense of retrieved lexical item may hard to be detected. Besides, the longer a word lives may accompany with richer senses. In order not to bias

on only lexical neology (newly formation of words), lemmas carrying richer senses or being homophonic are included in order to capture life in general words. Such lemma is termed as weighted lemma because its frequency may be enlarged by its rich sense or homophonic forms. The compensation made is the number of senses and number of variants would be explored as linguistic features in study.

3.1.4. Proposed Life Stages

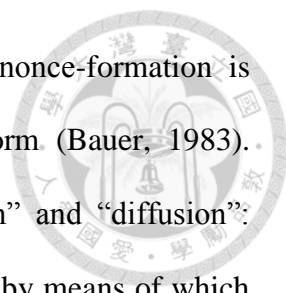
As reviewed in previous chapter there are diverse angles in understanding lexical items. For its fluctuations in life Historical Linguistics may be a good way to understand its grammaticalization or reviving. For its emergence in birth, neology can be probed from three aspects (Antoinette, 2013): semantic neology, lexical neology, and grammatical neology. The types of neology may be identified, but the stages of words before them are accepted into lexicon that pass from one generation to the next generation is still unclear. The potential stages words may experience will be proposed to be explored in this work. The description on “Life Cycle” of words described from previous studies may be overlapped in the stages or biased only to certain aspect as reviewed in Chapter 2. Some may focus only on words existing over century as the first emergent words (Wang and Minett, 2014), some may pay attention on types of conventionalization (Kerremans, 2015), and some may taking separate perspectives in discussion (Schmid, 2008). In this thesis the complicated delineation on a words’ life proposed from previous studies is abstracted as a continuum including stages of “Birth,” “Diffusion,” “Conventionalization,” and “Inactivation” in present thesis. It is supposed that life stages of words are more like a cycle, so those who are inactivated may be activated once the speech community decide to adopt it. Though this study has not yet touched this issue in depth, the path for inactivated words to be activated again should



be called as “reviving.” This study proposes that reviving includes three situations. The first one is old lemma with brand new meaning as in the case of “囧,” which signifies “bright” in its creation, but in the reviving of recent years “囧” signifies “embarrassed” for it looks like an embarrassed face. This reviving type should be corresponded to “exaptation” captured in historical linguistics reviewed in 2.1.1. The second type of reviving should be old lemma with original meaning, but is revived for the need of speech community. For example, “中肯” represents the meaning to recognize “precise saying,” so it gets its popularity in PTT, where focuses on opinion exchanging to express opinions in posts or comments, so it is common to see such recognition on opinions in response. This reviving is different from the first one, for the reviving of “中肯” is due to the preciseness of meaning this lexical item born with, so for its useful function in response it is chosen to be popularly used in contemporary. The third type of reviving is the reviving of once existed, but that function has not been in use over century.

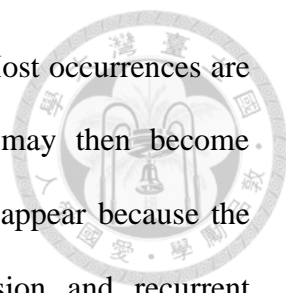
Words belong to the stage of “**Birth**” should include lexical neology, morphological neology, or reviving old lemma, but though homophonic in form it is with totally different senses (the “exaptation” proposed by Vincent (1995)). This proposed stage corresponds to “individual innovation” proposed by Wang and Minett (2004). However, this stage is hard to instantly capture for being used in individual corpus unless they are reached certain diffusion into larger community, so we are not going to study their characteristics in present work.

Words in diffusion should be those paid attention on in studies of Neologisms because Fischer (1998) has defined neologism as “A neologism is a word which has lost its status of a nonce-formation but is still one which is considered new by the majority



of members of a speech community.” The situation of losing nonce-formation is reached if the speaker is aware of having used or heard the form (Bauer, 1983). Kerremans (2015) described the attributes of “conventionalization” and “diffusion”: “conventionalization refers to the dynamic socio-pragmatic process by means of which a linguistic innovation becomes established in the language and the speech community..... diffusion denotes the dynamic spread of novel formations across the language and its speakers; it is therefore as much a socio-cognitive as a linguistic process, affecting both society and the language.” The proposed definition to “**Diffusion**” in this thesis is that the words in this stage should be those that can be comprehended by certain groups of people and are highly activated in certain registers. The grouping of people should be set boundary by the meta-information like age, gender, living area, or other social habits. Namely, it should be the fashion words that have been widely comprehended and used in contemporary. In this aspect, it is similar to the situation described as transitional conventionalization in Kerremans, but it should not be limited to those born from social events. The sources of fashion words proposed may come from two types: (1) The reviving of old lemmas. (2) Brand new coined usages. As stated earlier the first type should include three types of lemmas: (1.1) lemma with its original meaning (1.2) lemma with brand new meaning (1.3) lemma gains its new meaning from related meanings.

“**Conventionalization**” proposed here is different from the complicated levels in Kerremans (2015) but is closer to the concern on “language acquisition” proposed in Wang and Minett (2004). Kerremans (2015) has proposed that transitional conventionalization is characterized by a sudden significant increase in frequency and diffusion into various types of source and fields of discourse. The frequency curve will



show one steep wedge-like rise and decline within a short period. Most occurrences are stronger linked to the coiner and coinage event. Such words may then become inactivated, or become “recurrent semi-conventionalization” (to re-appear because the identical or similar events occur). The distinct between diffusion and recurrent semi-conventionalization does not have clear quantitative threshold, though she has pointed out that whether the diffused words become established as recurrent semi-conventionalized words depends on the regularity with which salience recurs and the degree of intensity. However, for current study words behave like this are only observed in short time period, so should be categorized as diffusion words for its increasing in use has not yet been stabilized. For Advanced conventionalization, which is illustrated as case study of “robosigning” in Kerremans (2015), it implies that the total amount high frequency within a short period indicates the higher intensity towards advanced conventionalization. “**Conventionalization**” in current study is proposed to be defined as words being stably used across different registers and across generation.

“**Inactivation**” should be those which are comprehensible but less active in being used. Stabilization in use is proposed in current study to be the behavioral indicator of the word is conventionalized or not for it include not only activation in frequency but also temporal information. Thus, a highly activated word may not resemble its being conventionalized for it may be just momentary burst in frequency. Similarly, if a lexical item is less active in its total frequency, but is stably used over time, then it is conventionalized, or it is inactivated. But, it should be notable that the observational resource from current study is synchronic PTT corpus, so the diffused words may have similar stabilized situation as the highly conventionalized lexical items for the diffused words are in fact diffused in this speech community. On the other hand, a lexical item

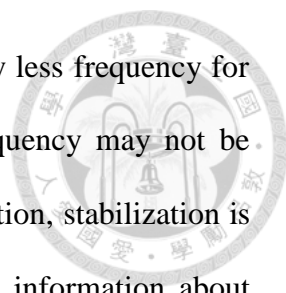
marked as inactivated in its low stabilization may still be able to come to revive.

3.1.5. Operational Definitions on Predicted Value

Frequency has been highly relied as reference of inclusion for dictionaries (Cook, 2010). Previous studies also have taken frequency as one of the predictive features (Kjellmer, 2000; Metcalf, 2002), or as the threshold for deciding being survival or not (Chang & Ahrens, 2008). Though frequency in the aspect of performance behaviorally represents the need of using, it can also reveal information about our comprehension. The highly frequent used ones are also those comprehensible ones. Additionally, frequency is not only the output result, but also the input influence: “frequency of complex words significantly influences the way in which we process and store them.” (Plag, 2004). In discussion on understanding morphologically complex words it is proposed that to access mental lexicon can be either by ‘whole word route’: directly access to the whole word representation, or by ‘decomposition route’: access to the decomposed elements (McQueen & Cutler 1998). Degree of decomposability of a given word (Hay, 2000; 2001) depends on relative frequency of the derived word and its base: the ratio of the frequency of the derived word to the frequency of the base.

$$f_{\text{relative}} = \frac{f_{\text{derivative}}}{f_{\text{base}}}$$

If the derivatives are more than the base, then it means the derivatives are no longer taking decomposition route, but the whole words route, which seems to be a good indicator for defining being conventionalized or not; however, it is hard to be adopted in present study. The first reason is that in Chinese definitions on “base” are still controversial. Besides, the elements for constituting disyllabic words are also included as target words in this study, so the f values of them are hard to retrieve appropriately based on this formula. Third, momentary total frequency may be hard to



capture real stabilization in use over time. Neologisms are relatively less frequency for their recency in coinage (Cook, 2010), so to focus solely on frequency may not be reliable. In addition, though frequency does imply valuable information, stabilization is the real one that embraces both activation of words and includes information about temporal aspect of activating, so it is more suitable to be the target predicting value for modeling. Hence, current study decides to adopt the Constant U from Wang (2010) to understand whether a word is stabilized or not.

Wang (2010) measures seasonal Constant U for words in teaching wordlists. She has claimed that with more temporal points in measuring, value of Constant U will be smaller, so the filter threshold would be stringent. Thus, instead of calculating Constant U by seasons, the values in current work would be calculated by month. It is termed as Revised Constant U in following discussion. For every lexical item its total frequency "x" (summed up from each month) is divided by the sum of total months in retrieved data to yield an average \bar{x} . Revised Constant U is calculated as the average frequency being divided by standard deviation of "x", as shown in following formula:

$$\text{Revised Constant U} = \frac{\bar{x}}{\text{stdev}(x)}$$

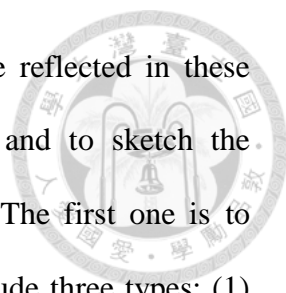
In our exploratory of data it is found that the correlation between Revised Constant U and total frequency is quite low, which signifies that these two values highlight different aspects of words. This corresponds to our assumptions that being used frequently does not mean being stabilized, for it may be just a flash in the pan. Meanwhile, small value in total frequency amount does not resemble that the word is not stabilized, for if it can still be constantly used across temporal period or when

certain event occurs. Revised Constant U can capture the phenomenon named as “Recurrent semi-conventionalization” in Kerremans (2015).

The differences in the values of Constant U in post and in comment, and measured in monthly or yearly are briefly summarized as following. The correlation of Constant U in posts by year or by month is highly correlated (0.85203), which is similar to situation in comments: the correlation of constant U by year and by month in comments is 0.9297476. The by-year correlation in comments and posts is less correlated (0.5964252), but by-month correlation is still high (0.7099764). The words captured as being used in posts are more than in comments.

Taking Revised Constant U as threshold to decide a word is being used or not, we can obtain following observations. In posts 31 words belong to those who are not used in selected boards for their value of Revised Constant U is zero. Among these 31 words it is only "大俗賣" has been used in comments but not in posts. There are 47 words that are not used in comments, which include 17 words that are used in posts. The by-month Revised Constant U in comments is chosen as the threshold value in quantitatively modeling stabilization in present work, for we assume that posts may be closer to formal written register in its larger context of arranging information structure, but relatively concise comments may be closer to spontaneously instant response in oral conversation. In order to capture how language is used in spontaneous way **Revised Constant U calculated in comments by-month** is taken as the indicator of behavioral performance in stabilization. Then, the underlying driven linguistic factors are investigated.

As mentioned in previous section, the adopted PTT corpus is to reflect more synchronic language using situation and is core originated speech community for the



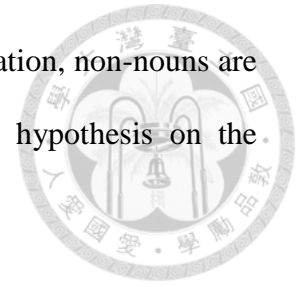
diffused words used in current work, so stabilization may also be reflected in these diffused words. Two strategies are adopted to compensate this and to sketch the differences between highly conventionalized and highly diffused. The first one is to control selected types of target words. Target words employed include three types: (1) Diffused words: highly diffused ones (words collected from Internet) (2) Words born before 1950: highly conventionalized (words used over a century) (3) Words born after 1950: recent conventionalized ones. The third type of words are those once diffused 50 years ago and adopted into dictionary in 1997~1998, so their stabilization value (Revised Constant U) can be used to understand how words are conventionalized or not after 50 years. Second, regression models are adopted to statistically sketch linguistic features of diffused words and conventionalized words. The collection and division on three sets of target words is discussed in detail in the next section.

3.2. Resource for Collecting Target words

When constructing prediction model or capturing life situations of words it would be great if we have resources to see how words fluctuates from its birth to its later development as in the way of Chesley and Baayen (2010) in comparing words first appearance frequency and 10-year later frequency. However, there is lack of available diachronic corpus as the newspaper corpus they use, so current study decides to compensate this by controlling conditions of selected target words and to explore their situations from the contemporary synchronic PTT corpus built by Liu (2014)². Thus, in order to reflect the features of general words used in daily life as proposed in previous section target words are collected from different sources and selection standards. Meanwhile, in order to avoid large effects coming from external referents and to focus

² <http://lopen.linguistics.ntu.edu.tw/PTT/concordance/>

on words that signify important linguistic information in communication, non-nouns are the targets included in current study. The retrieved sources and hypothesis on the selected words are discussed as following.

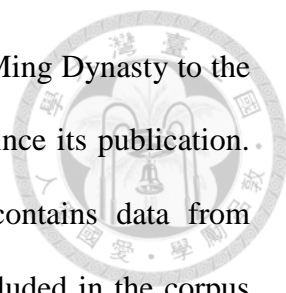


3.2.1. *Kim (2006) and Chang and Ahrens (2008)*

These are the new verbs proposed by National Languages Committee and they are included in the dictionary proposed by Ministry of Education around 1998 as “Collection of Neologisms I.” Kim (2006) and Chang and Ahrens (2008) have provided these wordlists. The reason to select these words is because they are once new words 17 years ago we can have the chance to understand their situations of stabilization and inactivation from angles of nowadays. They are all compounded verbs with single meaning. The list includes both loan and non-loan verbs. Meanwhile, they are good candidates to be the test data for machine learning because we already realized its life situation after 17 years.

3.2.2. *Google Books Ngram Corpus (GBNC)*

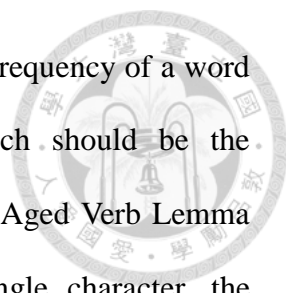
Data used in formal writing are retrieved because with the intention to be understood by readers formal writing should be more careful in using words that can be comprehended by the audience of that generation, so the words used are more representative of that generation. Besides, formal writing is more stringent in using words and less easily in giving away words, so if words in such register have been lost, then they would also no longer be used in oral condition. Thus, we can have the chance to understand the potential common ground between new born and died words. The main resource we use in this study is based on Google Books Ngram Database (GBNC, Michel et al., 2011).² The database has been available online since 2010, which supports data query across many languages. The huge dataset originates from the “Google Book



Project” that aims to digitalize books from 1500s, which is around Ming Dynasty to the present, and has facilitated many researches of digital humanities since its publication. As described in Lin et al.(2012), the new edition of GBNC contains data from 8,116,746 books, or over 6% of all books ever published. Data included in the corpus are only those ngrams that appear over 40 times across the books, where an ngram refers to the consecutive sequence of n items from texts (n = 1, 2, 3, etc.).

Google Books Ngram performs tokenization and sentence boundary detection for Chinese with a statistical system. The tokens for each language are listed as Table 1; however, current study has removed the non-Chinese tokens that have been collected in books, so the number of tokens is 8,535,128, and the number of types is 57,089. The data are tagged with parts of speech. It uses the universal POS tagset described in detail in Petrov et al. (2012). It is noted that for the Chinese section, data were retrieved from the books published in Mainland China, so it is simplified characters. But, the data would be transformed into traditional Chinese when adopted in current study. Meanwhile, a single lexical form may serve with more than one POS. For example, “burnt” can express either verbs (e.g., the house burnt) or adjectives (e.g., the burnt toast).

The goal is to sample conventionalized verbs from GBNC. In order to minimize potential tagging errors current study filter out ADJ and ADV but only use words that are tagged as VERB. By comparing the Top 10 Aged Verb Lemma and Top 10 Frequent Verb Lemma as well as from the overall correlation statistics we can realize that the correlation between frequency and birth of year is low (0.1766458). This indicates that the longer a word lives does not necessarily imply its popularity in being used, which corresponds to our hypothesis that there are inactivated cases in those long lived



conventionalized words. Meanwhile, this is also reasonable for the frequency of a word should be driven by many factors, for example, one of which should be the communicative needs from the speakers. Thus, by peeping Top 10 Aged Verb Lemma we can realize that the earlier born core verbs are mainly single character, the monosyllabic verbs. They should signify the earlier communicative needs from long time ago. Top 10 Frequent Verb lemmas provide the information about the communicative need from an overall picture. Interestingly half of them are compounding, which shows that the emergent compounding usages own some functional priority to win over those long-lived usages as shown in top 10 Aged Verb, thus this should imply that functionality of words should also be focused in influencing the conventionalization of a verb or not. In addition to aged and frequent verbs, verbs born in every 100 year are randomly sampled to enlarge the generality of selected target verbs. Error tagging ones are excluded.

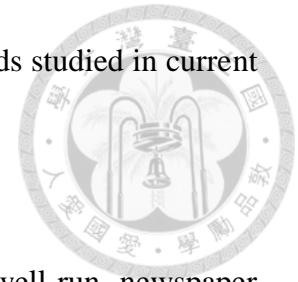
3.2.3. *Web*

The words collected from GBNC are born in sort of different ways from contemporary new words, which may cater for the new communicative needs in modern society, and are built up based on existed semantic representation, so recent new words are more constituted through compounding. The collection of new born diffusion words mainly come from PTTpedia³, which records frequent used terms in PTT as well as PTT events. The Internet Fashion terms reported in news⁴ are also included, though some of words are words that have already been recorded in Ministry of Education around 1998. As discussed in Kerremans (2015) Internet language has become the bed for giving birth of new words. Besides, Internet may even become the sources of news

³ <http://zh.pttpedia.wikia.com/wiki/PTT%E9%84%89%E6%B0%91%E7%99%BE%E7%A7%91>

⁴ <http://dailyview.tw/Daily/2014/10/20>

to the formal register like newspaper. Thus, the target diffusion words studied in current study may highly focus on new words from Internet.



3.2.4. Newspaper

United Daily News (UDN), which is a well-known and well-run newspaper published in Taiwan since 1951. The online resources they provide are rich, and are freely retrievable within the web area of National Taiwan University (NTU) because NTU has paid the fee for data resources. We have also manually consulted texts in 1951 and 2000 in United Daily News (UDN) database to be inspired with words that used in 50 years ago and words popping out in recent years in order to enlarge the generality of selected words.

3.2.5. Chinese Wordnet

Chinese Wordnet (CWN) is also used to retrieve information about synsets because current study hypothesizes that sense play an important role in sustaining the life of a lemma and giving birth of new expressions. Wordnet provides much more rich information on number of senses assigned to verbs than other classification schemes as shown in the comparison in English VerbNet, FrameNet, Levin's classification, Roget's thesaurus, and Wordnet conducted by Baker(2008).

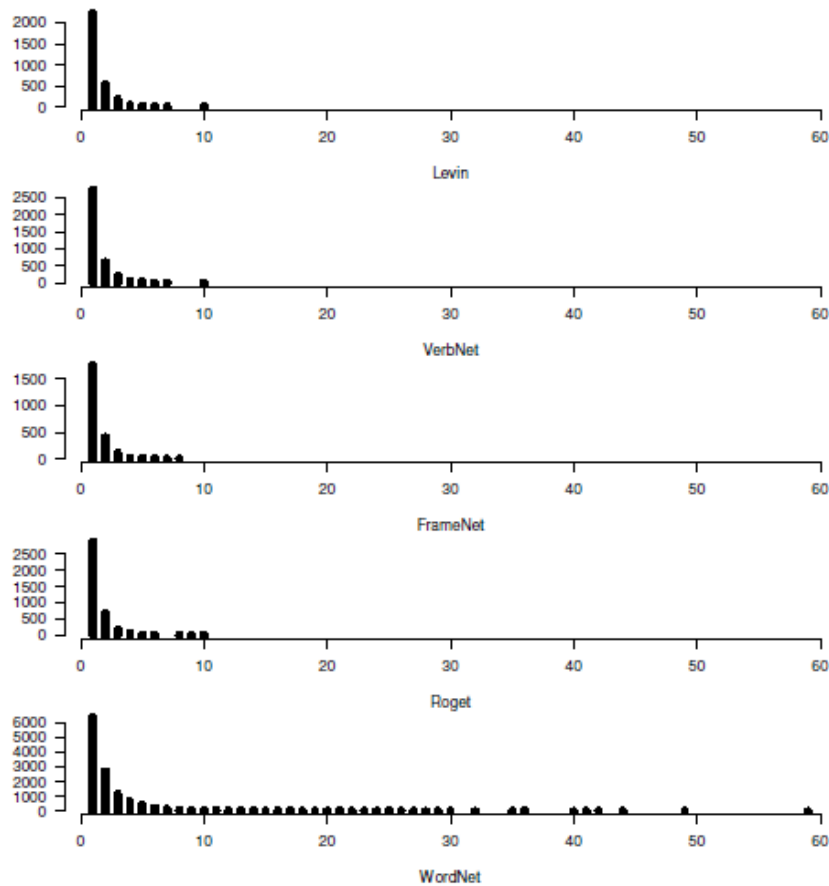
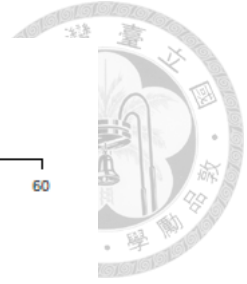


Figure 4 Comparison in English VerbNet, FrameNet, Levin’s classification, Roget’s thesaurus, and WordNet (Baker, 2008)

The brief quantitative information about CWN is shown in Figure 5. Current study is going to observe synsets with large members to understand the potential competitions in lexical items.

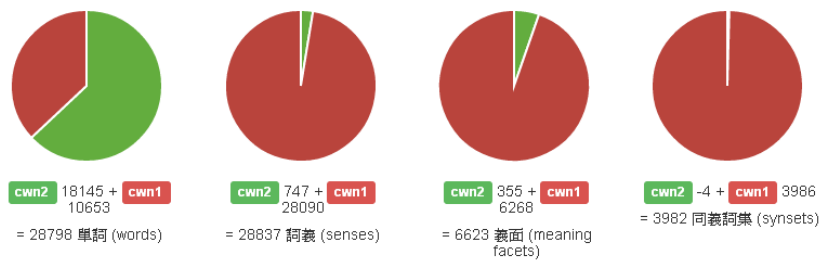
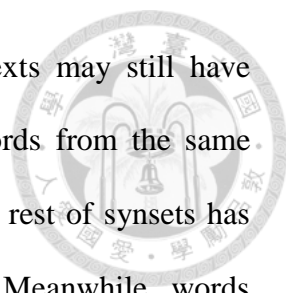


Figure 5 Quantitative Information About CWN



With the assumption that verbs in similar paradigmatic contexts may still have different behavioral characteristics current study also included words from the same synset. The largest two verb synsets have been chosen. Two of the rest of synsets has been randomly chosen. One set of variation words is included. Meanwhile, words derived from “hot” and “cold” are also included to understand the potential semantic connection between antonymic or near-synonymous words and words from the same embodied experiences because competitions among synonymous lexical items have been largely proposed (Meillet , 1958; Boulanger, 1997).

3.3. Categorization on Target Lexical Items

Based on the resources for selecting target words we can generally classify the words into those who should be in diffusion and those who have been once diffused in the past (past-diffused). This labeling is for facilitating the features observed in later discussion. The diffused ones are those expressions used on Web, so they are collected from either PTTpedia or news about Internet language. The past-diffused ones are those who have been through the stage of diffusion, and may or may not be conventionalized. Among these past-diffusion words those who are earliest traceable after 1950 are labelled for further discussion because they are relatively elder than the diffused ones but more recent than those words that have been used over century. On the other hand, some of the diffused ones may be semantic neologisms, reborn, or revived ones, so they will get extra labels for such meta-information. For factors like being loaned or not, being written in Chinese character or not, and having written variants or not are assumed as linguistic features, so they will be discussed in detail in section 3.4. The details on different types of target words are illustrated as following discussion.

3.3.1. Target Lexical Items for Understanding Diffusion

Lexical items from PTTpedia and reported Internet Language are used to observe behaviors of diffusion verbs for current study because they are currently new comers within 10 years. This temporal boundary is taken with reference to the suggestion from Kim (2008). Besides, these should be those who lost nonce formation as defined in Fischer (1998). These diffused words can be further delineated into several types as the proposal discussed in section 3.1.4. The re-born one may be the one who is homophonic and uses existed lemma, but bears totally different senses as the situation of “exaptation” defined in Vincent (1995). The revived ones are those who may exist in past usage, but turn out to be highly activated in contemporary Internet community. Besides, there are also cases coming from metaphorical sense extension or conversion. Such cases belong to semantic neologism. On the other hand, homophones may also influence the frequency of using words, which can be divided into two types. The first one is the lemma is with rich semantic information from its constituted elements as in the case of “神人.” The “神” can be a modifier to “人” to express “The person is amazing,” or it can also be a transitive verb with “人” as subject to express “find out the amazing person.” The second one may be the acronym from loan words. For example, “OP” can stand for “over post” to mean repeated posted content, and it may mean “over power” to express the modified one is powerful, or it can also stand for “One Piece,” a Japanese comic. There are also variations of written forms that can be further discussed and manipulated. For example, “ㄉㄉ” originally is an onomatopoeic symbol for laughing, but recently has been pragmatically enriched with ironic implication. The semantic neologisms are those who belong to diffusion words, but are using weighted lemma. Reviving old lemmas are those who belong to conventionalized words, but get

reactivated. The analysis on diffusion words are summarized in Table 4.

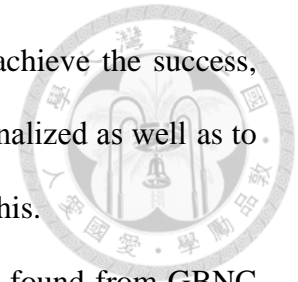
Semantic Neologism	Monosyllabic polysemy	"神","黑","萌"
	Homophone	"神人","CD","GG","OP"
	Metaphorical Extension	"關燈","開燈"
	Conversion	"筆記"
Reviving of Old Lemma		"中肯"
Reborn		"囧"
Variation in Written Forms		
	Synonymy	"厂厂","頗厂","根本厂厂" "頗呵","根本呵呵","頗喝" "科科","丐丐","顆顆"

Table 4 Types of Diffusion Words Included

3.3.2. Target Verbs for Understanding Conventionalization

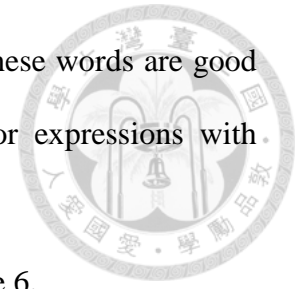
As stated in section about selections on target verbs, the sources to select verbs are very different. For understanding conventionalization the sources are mainly dependent on GBNC, CWN, and “Collection of Neologisms I.” However, in order to delineate the differences between “first emergent words,” those who exists centuries ago as defined in Wang and Minett (2004) and the recent emergent ones around 50 years, these verbs are further anchored their earliest traceable time in GBNC. For words whose earliest traceable is a century ago, the “first emergent words,” should be those “survivor” in

their contemporary competitor, so it is hard to observe how they achieve the success, but we can observe how their “living style” is after being conventionalized as well as to push further to understand how linguistic factors may contribute to this.



There are contradictory between the earliest traceable year we found from GBNC and the target words listed in “Collection of Neologisms I.” For example, for words like “跳槽,” or “搞笑,” which are included in National Languages Committee in 1998, but traced separately as born in 1959 and 1982 in GBNC. The traceable temporal information from GBNC would be taken. On the other hand, diffused words we collected from web may be actually reviving lexical items that have existed long time ago. For example, “反串” and “搞定,” which may be already in use around 1959, so it is decided to put them into conventionalized wordlists. Those who can be traceable after 1950 in GBNC would be categorized as words born after 1950 for they may be once diffused around 1950, and are good models for predicting possible living situations of diffused words in contemporary. For those who are in 1900 ~ 1949 we categorize as words before 1950. Both words after 1950 and before 1950 should be viewed as conventionalized, but they may provide different aspects of information for words born before 1950 in current study are those lived over century, so they can illustrate more rooted conventionalized situation. Words born after 1950 should own qualities closer to contemporary diffused words. Furthermore, “中肯” and “淡定” are reviving terms without additional meaning, so they should still be categorized as reviving conventionalized words. Besides, words proposed by being inspired by reading newspaper in 2000 are categorized as words born after 1950 since they are recently conventionalized rather than diffused around 2000. For words randomly selected from CWN and cannot be anchored temporal point from GBNC and from “Collection of

Neologisms I’ are all categorized in as words born before 1950. These words are good targets to observe how different members in the same synset, or expressions with antonymic relationship behave similarly or differently.



The number of three types of target words is shown as in Figure 6.

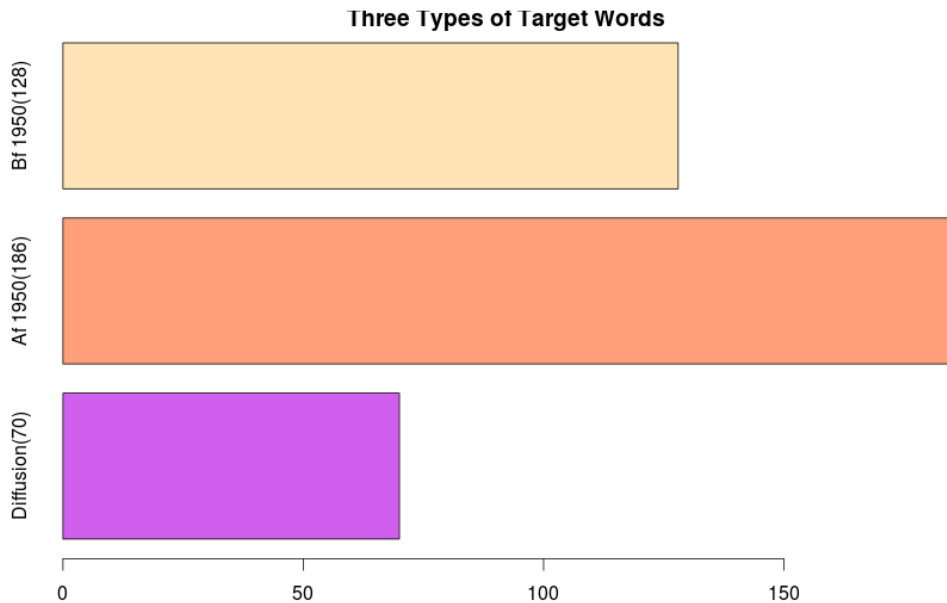
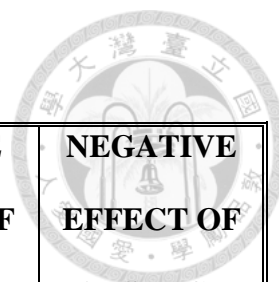


Figure 6 Number of Three Types of Target Words Observed in Current Study

3.4. Proposed Linguistic Predictors for Understanding Stabilization

Chang and Ahrens (2008) computed three prediction models on judging words’ survival and fail. The first two are separately based on factors of Kjellmer (2000) and of Metcalf (2002), and the third one is the hybrid of the two. The features used in the three models are summarized in Table 6 for features proposed by Kjellmer (2000), Table 66 for FUDGE model of Metcalf (2002), and Table 77 for predictors selected by Chang and Ahrens (2008). These summarized charts are quoted from Chang and Ahrens (2008).

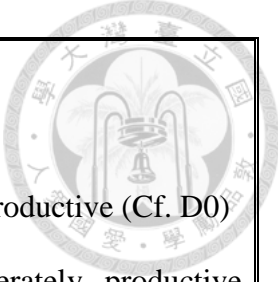


FACTORS	POSITIVE EFFECT OF PRESENCE	NEGATIVE EFFECT OF ABSENCE
S1. It has semantic parallels in the language.	1	-1
S2. It is transparent to the layman.	0	-1
Ph1. It has phonological parallels in the language.	0	-3
Ph2. It is easy to pronounce.	0	-1
M1. It has morphological parallels in the language.	3	-1
M2. It follows morphological principles.	0	-3
M3. Its derivative affix is highly productive.	3	-1
M4. Its derivative affix is compatible with the stem.	2	-1
G1. It has graphematic parallels in the language.	0	-3
G2. Its spelling agrees with its pronunciation.	0	-1
O1. It has prestigious and/or exotic connotations.	2	0
O2. It is concise.	1	0
O3. It has humorous connotations.	2	0

Table 5 Factors from Kjellmer (2000), Summarized by Chang and Ahrens (2008)



	Full Names	Definitions (according to Metcalf)	Scores*
F	Frequency of Use	This factor can also be expressed as popularity, plain and simple.... The kind of popularity a new word needs is attention.	0: Friends, family, coworkers, or only one person 1: 1000~100,000s 2: Widely used
U	Unobtrusiveness	“In plain English, you don’t notice it....[A new word] camouflages itself to give the appearance of something we’ve known all along.”	0: Conspicuous (exotic, clever) 1: Noticeable 2: Unobtrusiveness
D	Diversity of Users and Situations	“[A new word] also needs to be used by a variety of people in a variety of situations.”	0: Specialized terms 1: In general conversation, with explanations 2: Widely used

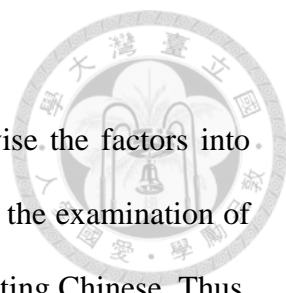


G	Generation of Other Forms and Meanings	<p>A new word that generates others also generates a greater chance for its own success.</p> <p>(a) “Variety of meanings”.(level 2)</p> <p>(b) “Generating new forms”(level 2)</p> <p>(c) N→V or V→N. (level 1)</p>	<p>0: Not productive (Cf. D0)</p> <p>1: Moderately productive (POS)</p> <p>2: Productive (New forms and meanings)</p>
E	Endurance of the Concept	The endurance or the durability, not of the word itself, but of what it stands for.	<p>0: Nonce form, archaism</p> <p>1: Historical references</p> <p>2: Long enduring</p>

**Table 6 FUDGE scale from Metcalf (2002),
Summarized by Chang and Ahrens (2008)**

Factors	Weightings	Factors	Weightings
Frequency	-1 0 1	Morph. Rules & Gaps	Y: 1
Productivity	-1 — 0 —▶ 2	Productive Affixes	Y: 2
Semantic Gaps	— Y: 1 —▶	Spelling/Pronunciation	Y: 1
Transparency	Y: 0	Prestige	Y: 1
Endurance	Y: 0	Conciseness	Y: 0

Table 7 Factors used in Hybrid Model from Chang and Ahrens (2008)

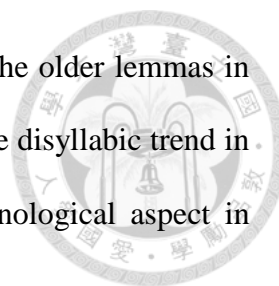


Different from Chang and Ahrens (2008) we are going to revise the factors into more linguistic orientated angle and to delete some factors based on the examination of Chang and Ahrens (2008) or consideration on applicability in evaluating Chinese. Thus, following discussion illustrates the operational definition of the features used in current study as well as clarifies corresponding features in previous studies. Features proposed in current study are categorized from the angle of linguistic domains in order to fully connect with insights from linguistics. On the other hand, among the proposed features, what Metcalf (2002) has proposed as “Endurance of the Concept” should be reflected in the calculation in Revised Constant U because enduring concept should be those who can be used stably. On the other hand, when building models given the fact that some lexical items’ frequency is zero, the value of Revised Constant U would be “NA,” so current study has replaced these NAs as zero for calculation convenience.

3.4.1. Phonology

Number of Syllable

Previous studies have taken many hypotheses like: “Ph1. It has phonological parallels in the language,” “Ph2. It is easy to pronounce,” or “G2. Its spelling agrees with its pronunciation,” but these features may highly Indo-European Language oriented, and if they are used in current Chinese target words, these features may only be related to the loan words for though we can capture loan words written in Chinese characters there are mismatches to the real pronunciation for phonotactic constraints across different languages, so instead of viewing previous phonological features current study would like to focus on the syllable part. The main reason is that from previous



discussion on selecting target words in section. It has implied that the older lemmas in GBNC are most monosyllabic. Besides, there are studies pointing the disyllabic trend in contemporary Chinese, so current study would like to probe phonological aspect in words by counting their numbers of syllables.

Assumptions in this section are summarized as in Table 8:

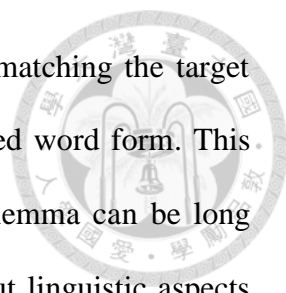
Proposed features in current study	Features in Kjellmer (2000)	Features in Metcalf (2002): FUDGE	Features in Chang(2008)	Features in Kerremans (2015)	Delete
Phonology					
number of syllable					
	Ph1. It has phonological parallels in the language.	Unobstrusiveness			Delete: Not meet features in Chinese
	Ph2. It is easy to pronounce.				
	G2. Its spelling agrees with its pronunciation				

Table 8 Summarized Chart for Phonological Predictor

3.4.2. Morphology

Component Richness

Previous studies mostly emphasized the importance in morphological productivity. The “M3. Its derivative affix is highly productive,” (Kjellmer, 2000) and “Generating new forms (level 2) of Generation of Other Forms and Meanings”(Metcalf, 2002) are included in the discussion of morphological productivity as in Chang and Ahrens (2008); however, it is controversial in defining affixes in Chinese when previous studies are taking derivational affixes as the calculating base. The phenomena that a morpheme has been actively used in constituting other lexical items should still be considered. As



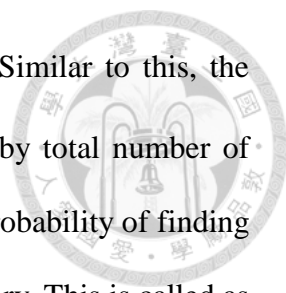
stated previously the relatively complicated and limited aspect in matching the target words is that it is hard to anchor the sense activated by the matched word form. This may be able to turn into an advantage in discussing why certain lemma can be long lived since that we are not just focusing on the frequency factor but linguistic aspects driven behind. The higher activation on forms stored in our lexicon can be a linguistic clue for incorporation into mental lexicon. Thus, the method adopted in current study is slightly different. Instead of following the approach of Chang and Ahrens (2008) in counting the number of lexicalized items constituted by certain monosyllabic character is over 10 or not, current study decides to adopt method similar to the way in understanding realized productivity and token-type ratio. The assumption is that a lexical item is easier to be passed down through generations if its components have constituted in other lexical items. This is due to the rich senses of this component for being used to form disyllabic or more syllabic lexical items should base on the sense it owns. Thus, the highly activation in being used in consisting other lexical items reflect both the activation in this morpheme as well as importance in its sense. On the other hand, as the example “Boobgate” comes from “Watergate” discussed in Kerremans (2015) the schema (a pattern, a rough outline, a coarse-grained, less-fully-specified version of a concept which the elaborations render, each in a different way, in finer, more elaborate detail) in conceptualization is encoded in morphemes constituting new lexical items (Tuggy 2005; Kemmer 2003), so current study would like to term such phenomenon as a factor called **component richness**. If a lexical item has more than one component, then the values of the higher one would be adopted as the richness in schema of that target words. Google Book Ngram Corpus (GBNC) is selected as the

resource for calculating value of each components in all lexical items because its formality.

The formula adopted is how previous studies calculate morphological productivity. There is rich discussion on productivity of morphological rules in previous studies. Previous studies have emphasized on its correlation with many other linguistic factors like phonological aspects, semantic transparency etc. (Plag, 2004) There are mainly five ways to approach this issue (Plag, 2004). First, the type-frequency V , which is called as realized productivity in Baayen (2009): with a text corpus or a large dictionary, productivity can be measured by counting the number of attested different words (type) with a particular affix. The greater the type-frequency is, the higher the productivity of the affix is. This measure indicates the past achievement, rather than present productivity (Plag, 2004; Baayen, 2009). Second, counting the number of neologisms in a given period can show an aspect of productivity. The greater the number of neologisms indicates the higher the productivity of a given affix in that period (Plag, 2004).

In addition to aforementioned method, the following measurements involve the concern on hapax legomenon, which is those rare words of language (instead of a newly coined derivative), or some weird ad-hoc inventions by an imaginative speaker or found in poetry or advertisement with respect to a given corpus. Hapax legomena is not the same as neologisms (Baayen, 2009). However, even not all of the hapaxes with a given affix are neologisms, Plag (2004) assumes that it is among the hapaxes (as against words with higher frequency) that we can find the highest proportion of neologisms (Baayen & Renouf, 1996; Plag, 2003).

Thus, the third method involves that the higher the number of hapaxes with a given



affix (n_1) in a large corpus shows the greater the productivity is. Similar to this, the fourth one is to divide the number of hapaxes with a given affix by total number of tokens with that affix, the P can be arrived at , which indicates the probability of finding new words among all the tokens of a particular morphological category. This is called as expanding productivity in Baayen (2009) and as compared with realized productivity the expanding productivity is viewed to illustrate the present producing power of the affix as in following formula. n_1^{aff} for the number of hapaxes with a given affix and N^{aff} stands for the number of all tokens with that affix. The formula of realized productivity is as following:

$$P = \frac{n_1^{aff}}{N^{aff}}$$

Expanding productivity may show category's contribution to the growth rate of the vocabulary in a corpus. However, in addition to the present productivity it may also care about the future productivity, so here comes to the final measuring aspect: the potential productivity (Baayen, 2009), which is highly sensitive to markedness relations. It is based on the assumption, "once an affix has saturated the onomasiological market, it has no potential for further expansion." (Baayen, 2009) This measurement is also indirectly sensitive to the compositionality of the words. The formula is the ratio I of the estimated size of the category S in an infinitely large corpus and the observed number of types V in a corpus of size N : $I = S/V(N)$.

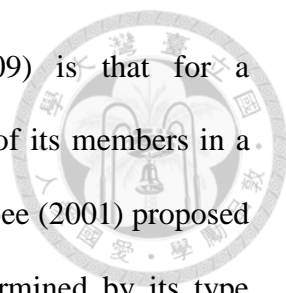
The first four methods are compared in Plag (2004) as shown in Table 9. It shows that the productivity ranking in different measures seem to contradict each other. This is reasonable for different measures highlight different aspects of productivity.



Rank	V		N		n_1		P		OED ne-ologisms	
1	<i>-ness</i>	2466	<i>-ion</i>	1369116	<i>-ness</i>	943	<i>-wise</i>	0.061	<i>-ion</i>	625
2	<i>-ion</i>	2392	<i>-ity</i>	371747	<i>-ion</i>	524	<i>-ish</i>	0.0338	<i>-ist</i>	552
3	<i>-ity</i>	1372	<i>-ness</i>	106957	<i>-ist</i>	354	<i>-ness</i>	0.0096	<i>-ity</i>	487
4	<i>-ist</i>	1207	<i>-ist</i>	98823	<i>-ity</i>	341	<i>-less</i>	0.0088	<i>-ness</i>	279
5	<i>-less</i>	681	<i>-less</i>	28340	<i>-less</i>	272	<i>-ist</i>	0.0036	<i>-less</i>	103
6	<i>-ish</i>	491	<i>-ish</i>	7745	<i>-ish</i>	262	<i>-ity</i>	0.00092	<i>-ish</i>	101
7	<i>-wise</i>	183	<i>-wise</i>	2091	<i>-wise</i>	128	<i>-ion</i>	0.00038	<i>-wise</i>	12

Table 9 Comparisons on Methods used in Calculating Morphological Productivity

Realized productivity is used in present work to measure component richness because the expanding productivity proposed by Baayen (1993) is more about the ability to produce new words rather than the ability to sustain the living situation of new words, but what we concern about is how words are supported from existed lexical representation. Meanwhile, the target words used in current study include both the affix itself and the compounds using these elements, so the purpose is not about predicting their productivity for the future or present, but to understand the supporting power accumulated from the past achievement, also the formula may be sort of different for the most of the target words are compounds. The higher potential or expanding productivity do not indicate the sustaining power on words' living. On the other hand, loan words are more complicated issue, so they will not into the discussion in this factor, but being dealt from factor about mixed originated morpheme.



The formula for realized productivity in Baayen (2009) is that for a morphological category C is estimated by the type count $V(C, N)$ of its members in a corpus with N tokens. Instead of viewing from token frequency, Bybee (2001) proposed that the productivity of a word formation schema is largely determined by its type frequency, for assessing the productivity of a schema in terms of token frequency would be counterproductive as in the case of comparing 1600 monomorphemic English verbs, of which 146 were irregular and 1454 regular (Baayen/Moscoso del Prado Martin, 2005). For in this case the summed frequency of all irregular verbs, 1793949, exceeds the summed frequency of the much larger set of regular verbs (732552). The driven reason is explained as unproductive categories use high token frequency to protect irregular forms from being regularized (Baayen, 2009). Differing from the resort to account for the ability of affix to produce what is concerned in current study is what the affix has produced may be the support to the new comers and be the strong connecting points for the existed.

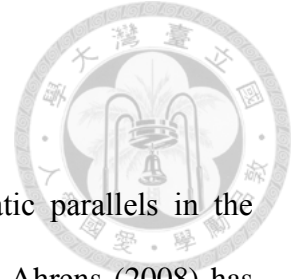
On the other hand, Fernández-Domínguez (2010) compared the morphological productivity of lexicalized ones and new-formed non-lexicalized compounds. The same formula is used to separately understand the degree of lexicalization of lexicalized ones and the profitability of non-lexicalized ones: number of individual lexical unit being divided by the sum of frequency of the lexical units.

$$\pi = \frac{V}{N}$$

In Fernández-Domínguez (2010), the resulted values are relatively higher in non-lexicalized ones than in lexicalized ones. Meanwhile, there are distinctive threshold

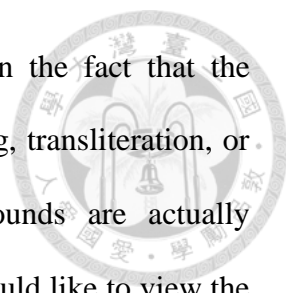
within the lexicalized ones. Thus, there may be some significant implications. Hence, component richness for target words in current study will be understood by evaluating both the realized productivity and the type-token ratio proposed by Fernández-Domínguez (2010).

Previous discussions all stand from affixation or compounding in Indo-European languages, but there is less discussion in the compounding productivity or affixation in Chinese. To simplify this issue the measuring unit in current study is syllable-based. Every syllable is viewed as a morpheme and its value is calculated. Then, for lexical units own more than one morpheme, namely those who are not disyllabic, instead of taking multiplication of each of its element, current study would like to choose the highest value among the elements to be the value of richness for the whole unit, which may seem to be arbitrary, but the underlying assumption is that higher morphological productivity is proved to be highly correlated with many linguistic aspects (Plag, 2004), and it may also imply how rich certain element is in our mental lexicon in constituting words ,which should be a facilitating effect on activating or integrating into existed mental lexicon because with more words related to this element it may indicate the probability in faster activation and the rich semantic root of the element. Meanwhile, in measuring this factor, the loan words should not be included for two reasons. First, the words borrowed from other languages may be morphologically complex in the donor languages but not necessarily decomposed in the borrowing process (Plag, 2004). Second, differing from the studies discussion on Indo-European languages loan words included in Chinese may be only transliteration from donor language, so the discussion on morphological productivity may be meaningless, so their parts will be further discussed on the factor about mixed origin morphemes.



Number of Graphematic Variation

Kjellmer (2000) proposed the rule that “G1. It has graphematic parallels in the language” should be one of the evaluated features, and Chang and Ahrens (2008) has testified this feature by searching the radicals, but in Chinese the origin of creating characters are quite complicated. The characters may have semantic bearing components or phonetic components, which though should be a good index to explore for loan words or monosyllabic characters, it is hard to use in current study because there are many compounds included in our target words, which are relatively less related to the graphemic components, so in the part of graphemic features instead of focusing on the single written character, current study would like to view from group-based angle: graphemic variations among target words. It is not necessary for every language to have written language, but for languages having written system how to write should be an important issue in reflecting many other linguistic factors, such as historical phonological changes or sociolinguistic factors. Given upon that written system is an important part in Chinese as well as the ways to retrieve research data are highly dependent on written sources, so current study decides to consider written variations as an independent variable that can contribute to the use of a lexical item. Hence, every written variation will be treated as an independent lexical item to probe their frequency and other proposed linguistic features for the way they are written down should influence their linguistic life. Meanwhile, number of variants a lexical item owns recorded in Chinese Wordnet (CWN) would also be included as one of linguistic features. New expressions like “ㄅㄅ” or “顆顆” are also updated in CWN, so the variation resource is reliable. On the other hand, however, the written forms of loan



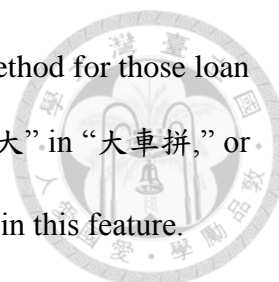
words should be various in chosen written ways, but given upon the fact that the strategies included in recording loan words may be direct borrowing, transliteration, or translation, namely, the variation in encoding the phonetic sounds are actually overlapped in the feature of being loaned or not, so current study would like to view the forms found in reachable resources (Zhu, 2000; Kim, 2006; Chang & Ahrens, 2008) as the standard form of loan words, and view them as non-varied. Lastly, there would be qualitative discussion on which win-over variations in later discussion.

Graphemic feature for being encoded by Chinese character or not

Kjellmer (2000) also proposed, “M1. It has morphological parallels in the language” and “M2. It follows morphological principles.” Current study testifies these by understanding whether the lexical item is written as Chinese characters or not. This is more about how borrowed words are encoded in the target language. It may adopt different translation strategies. The loan words may be direct borrowing as in using “lag,” or may be transliteration as “累格” for “lag,” but sometimes the translation may consider both semantic and phonetic features in the case of “快活,” so this feature may also highly related to loan words, but there is also exception as in “不解釋” is denoted as “BJ4,” which may be for humorous or easy to type down reason. To this feature those who are written in its donor language or in Zhuyin, or those who mix Zhuyin and Chinese characters will be marked as N as not being coded in Chinese character. Thus, this feature may be a view to understand how words activation in use.

Mixed originated morphemes or not

Kjellmer (2000) proposed, “M4. Its derivative affix is compatible with the stem,” which is testified by Chang and Ahrens (2008) as “words should not have morphemes of mixed origins,” and as stated in Chang and Ahrens (2008) this feature is more related



to those borrowed ones, so current study would like to follow this method for those loan terms that has been mixed with Chinese semantic morphemes like “大” in “大車拼,” or aspect marker as “了” in “卯死了.” These words will be marked “Y” in this feature.

Assumptions in this section are summarized as in Table 10:

Proposed features in current study	Features in Kjellmer (2000)	Features in Metcalf (2002): FUDGE	Features in Chang(2008)	Features in Kerremans (2015)	Delete
<u>Morphology</u> Component Richness of the monosyllabic verb or of the elements in the disyllabic verb constructions	M3. Its derivative affix is highly productive.	“Generating new forms (level 2)”of “Generation of Other Forms and Meanings”	Productive Affixes		Source of Data: Google Book N-gram Corpus (GBNC)
<u>Morphology</u> Number of graphematic variation	G1. It has graphematic parallels in the language.				
<u>Morphology</u> be encoded by Chinese character or not	M1. It has morphological parallels in the language. M2. It follows morphological principles.				
<u>Morphology</u>	M4. Its derivative		words should not		
Mixed originated morphemes or not	affix is compatible with the stem.		have morphemes of mixed origins		

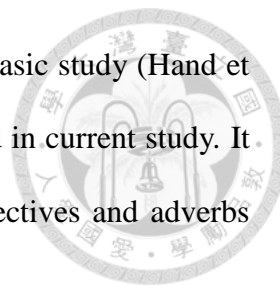
Table 10 Summarized Chart for Morphological Predictors

3.4.3. Syntax

Part of Speech

Syntactic factors are less discussed in previous studies, but the attribute of categories of words should be taken into consideration because the part of speech itself denotes the specific functions and behavioral distributions of lexical items. Though there are other non-noun lexical items, among them verbs are highly related to syntactic

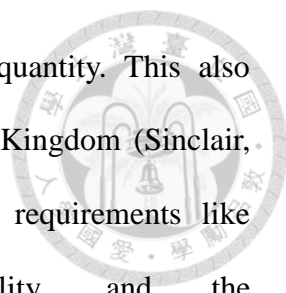
structure, and thus more complicated than nouns as indicated in aphasic study (Hand et al., 1979). The constructions of verbs are the main concern explored in current study. It is notable that current study would like to adopt the view that adjectives and adverbs should be delineated as verbs in Chinese.



Co-occurrence

The proposal of Metcalf (2002) in “Generation of Other Forms and Meanings” has been adopted as number of co-occurrences in Chang and Ahrens (2008)’s design. This may also correspond to the H6: The early development of syntagmatic lexical networks, represented by collocations in the present study, promotes conventionalization.in Kerremans (2015). This process of developing syntagmatic lexical networks in mental lexicon is called as network-building in Aitchison (2003). It is importantly shown from how children deal with words in similar sigmatic context, such as those near synonyms or antonyms. For children and adults collocations all show certain degree of importance in words’ identification and learning (Aitchison, 2003). Thus, though some linguistics may devalue the importance of collocation by emphasizing selectional restrictions/preferences, and take lexicon as a list of interchangeable words (Stefan, 2004), from previous discussion it should not deny what Firth (1957) said, “You shall know a word by the company it keeps!” Namely, word meaning can be learned from the words come alongside. Collocational links may be those optional candidates that commonly associated as in, "rude adolescents," or "fresh-faced youths." Some frequent associations may become fixed order as in "bride and groom, "or become clichés (Gibbs and Gonzales, 1985; Fenk-Oczlon, 1989).

Stefan (2004) has introduced that “Collocation” can be explored mainly in two approaches: distributional approach and intensional approach. The former one proposed



by the Neo-Firthians is to define collocations with observable quantity. This also become reference for corpus-oriented lexicographic in the United Kingdom (Sinclair, 1991; Lehr, 1996; Williams, 2003). The latter one may meet requirements like “semantic non-compositionality, syntactic non-modifiability, and the non-substitutability of components by semantically similar words” (Stefan, 2004). Stefan has termed distributional notion as “cooccurrences,” which employs co-occurring frequency information and statistical association. Cooccurrences may be positional or relational (Stefan, 2004). Positional cooccurrences are co-occurred words within certain distance, the collocational span (Sinclair, 1991). Relational cooccurrences are concerned with linguistic views in the structural relationship involved by co-occurring words.

The issue is simplified in present work by calculating number of different types of co-occurrences to illustrate the horizontal connections of the lexical items. The co-occurrence used in the thesis refers to co-occurred words without setting arbitrary threshold on co-occurring frequency because for some newly emergent expressions with low frequency the co-occurrence phenomenon may easier to be filtered out for it is hard to compute meaningful association scores for less frequent data (Stefan, 2004; Cook, 2010). There are 22 boards used for calculating co-occurrence from comments in PTT: LoL, ToS, PuzzleDragon, MenTalk, WomenTalk, Boy_Girl, Hate, happy, Sad, NBA, Baseball, movie, Food, BuyTogether, home_sale, Stock, StupidClown, joke, ask, Kaohsiung, Keelung, TaichungCont.

Assumptions in this section are summarized as in following chart:



Proposed features in current study	Features in Kjellmer (2000)	Features in Metcalf (2002): FUDGE	Features in Chang(2008)	Features in Kerremans (2015)	Delete
<u>Syntax</u> Co-occurrence		“Generation of Other Forms and Meanings”	Productivity words having more than ten collocates would be scored as two; those with less than ten collocates but having more than three Word Sketch functions would be considered moderately productive and scored as one; and those with less than ten collocates and having less than or equal to three Word Sketch functions would be scored as zero	H6: The early development of syntagmatic lexical networks, represented by collocations in the present study, promotes conventionalization.	

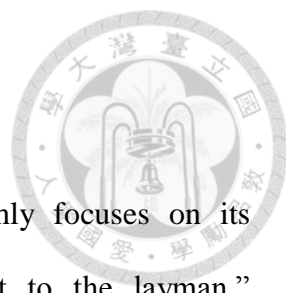
Syntax
Parts of Speech

Proposed features in current study	Features in Kjellmer (2000)	Features in Metcalf (2002): FUDGE	Features in Chang(2008)	Features in Kerremans (2015)	Delete
<u>Syntax</u> Co-occurrence		“Generation of Other Forms and Meanings”	Productivity words having more than ten collocates would be scored as two; those with less than ten collocates but having more than three Word Sketch functions would be considered moderately productive and scored as one; and those with less than ten collocates and having less than or equal to three Word Sketch functions would be scored as zero	H6: The early development of syntagmatic lexical networks, represented by collocations in the present study, promotes conventionalization.	

Table 11 Summarized Chart for Syntactic Predictors

3.4.4. Semantics

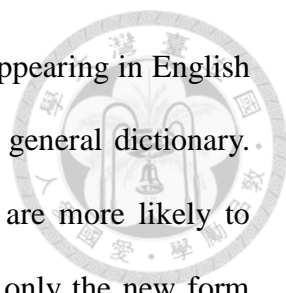
Number of senses



The semantic part in hypothesis from previous studies highly focuses on its semantic transparency in comprehension: “S2. It is transparent to the layman,” (Kjellmer, 2000) and the “Unobtrusiveness” in Metcalf (2002), which has been interpreted as “the meanings of transparent words should not be specialized and must be clearly inferable from the form” in Chang and Ahrens (2008), and corresponds to the “H1: Semantic ambiguity” in Kerremans (2015). However, in real language use semantic ambiguity could be solved by contextual information. Besides, transparency can actually be reflected in the range of use and number of senses for based on frequency effect as well as studies on mental lexicon it can be inferred once the sense of the form has been highly activated, no matter this meaning is morphological or metonymic originated, then it becomes automation. Thus, there are no significant activating differences in reaction time as in those cases of entrenched metaphors. Namely, the quality or origination of sense may not be the core factor influence the sense being adopted or not because once adopted in use it is in use. On the other hand, the number of senses, or the relationships with other lexical items may contribute more, so they are included in discussion of current study.

Number of synonym, Number of near synonym, Number of near synonym, Number of antonym, Number of holonym, Number of hyponym

In addition to syntagmatic angle indicated in co-occurrence, the paradigmatic view is also important, so various semantic relationships should all be included to understand the use of certain lexical items because words in the same synset may in the relationship of competition (Boulanger, 1997). Boulanger in her studies on comparing survived



words and fade-away words, which is defined by observing words appearing in English new-word dictionary in 1990 and later inclusion of these words in general dictionary. She finds that new words that have competing established forms are more likely to succeed because in competition case the concept is established, so only the new form need to run for supporting from speakers, but in non-competition case both the new referent and the new form need to be accepted. Besides, Though words listed in the same synset may have partial overlap in use, they should also have their unique living as shown in the study on different suitable contexts for “chase” and “pursue” (Aitchison and Lewis, 1996), so in order to explore connection in this part words in the same synset are compared with their Revised Constant U as well as their information across variables proposed in current study qualitatively.

Similarly, antonyms that are defined as words of opposite sense are also interested in understanding because members of antonyms are interchangeable syntagmatically or frequently co-occurred (Charles and Miller, 1989; Fellbaum, 1995). There are different types of antonyms: binary antonym, gradable antonym, and converseness (Lyons, 1981; Cruse, 1992; Murphy and Andrew, 1993; Kreidler, 1998; Cruse, 2011).

Superordinate relations are most available if only the group members are fairly prototypical and the superordinate label is commonly used (Johnson-Laird, 1983; Hurdord et al., 2007; Cruse, 2011). The information is less reachable in CWN, so current study would not focus on discussion in this part. One notable thing is that to all target words the meronymic relation cannot be retrieved, so this variable is excluded in current study.



Assumptions in this section are summarized as in Table 12:

Proposed features in current study	Features in Kjellmer (2000)	Features in Metcalf (2002): FUDGE	Features in Chang(2008)	Features in Kerremans (2015)	Delete
<u>Semantics</u> Number of senses		“Variety of meanings(level 2) ”of “Generation of Other Forms and Meanings”(Metcalf 2002)			
<u>Semantics</u> Number of synonym	S1. It has semantic parallels in the language. O2. It is concise		Semantic Gaps	if there are no competing synonyms, then we consider the word filling up a semantic gap.	
<u>Semantics</u> Number of near synonym	S1. It has semantic parallels in the language. O2. It is concise		Semantic Gaps	if there are no competing synonyms, then we consider the word filling up a semantic gap.	
<u>Semantics</u> Number of antonym					
<u>Semantics</u> Number of holonym					
<u>Semantics</u> Number of hyponym	S2. It is transparent to the layman.	Unobtrusiveness	Transparency: we adopt identical operational definitions as in Metcalf’s model, i.e., the meanings of transparent words should not be specialized and must be clearly inferable from the form.	H1: Semantic ambiguity	Delete: This can be reflected in the dissemination across language users and number of senses for to investigate the meaning is morphological or metonymic originated is not so meaningful because based on frequency effect as well as studies on mental lexicon once the sense of the form has been highly activated, then it becomes automation.
					so there is not significant activating differences in reaction time as in those cases of entrenched metaphors.

Table 12 Summarized Chart for Semantic Predictors

3.4.5. Sociolinguistics



Loan words or not

Language contact is an important path in enlarging lexicon of language users. Being loaned from other languages is controversial in whether it is inhibiting or facilitating effect in adopting the loan word (Kjellmer, 2000; Metcalf, 2002; Kerremans, 2015). Borrowing from other languages have been richly discussed (Betz, 1949; Haugen, 1950; Weinreich, 1953) can be delineated into different situations as shown in the schematic classification summed up and modified by Duckworth (1977) in Figure 7 .

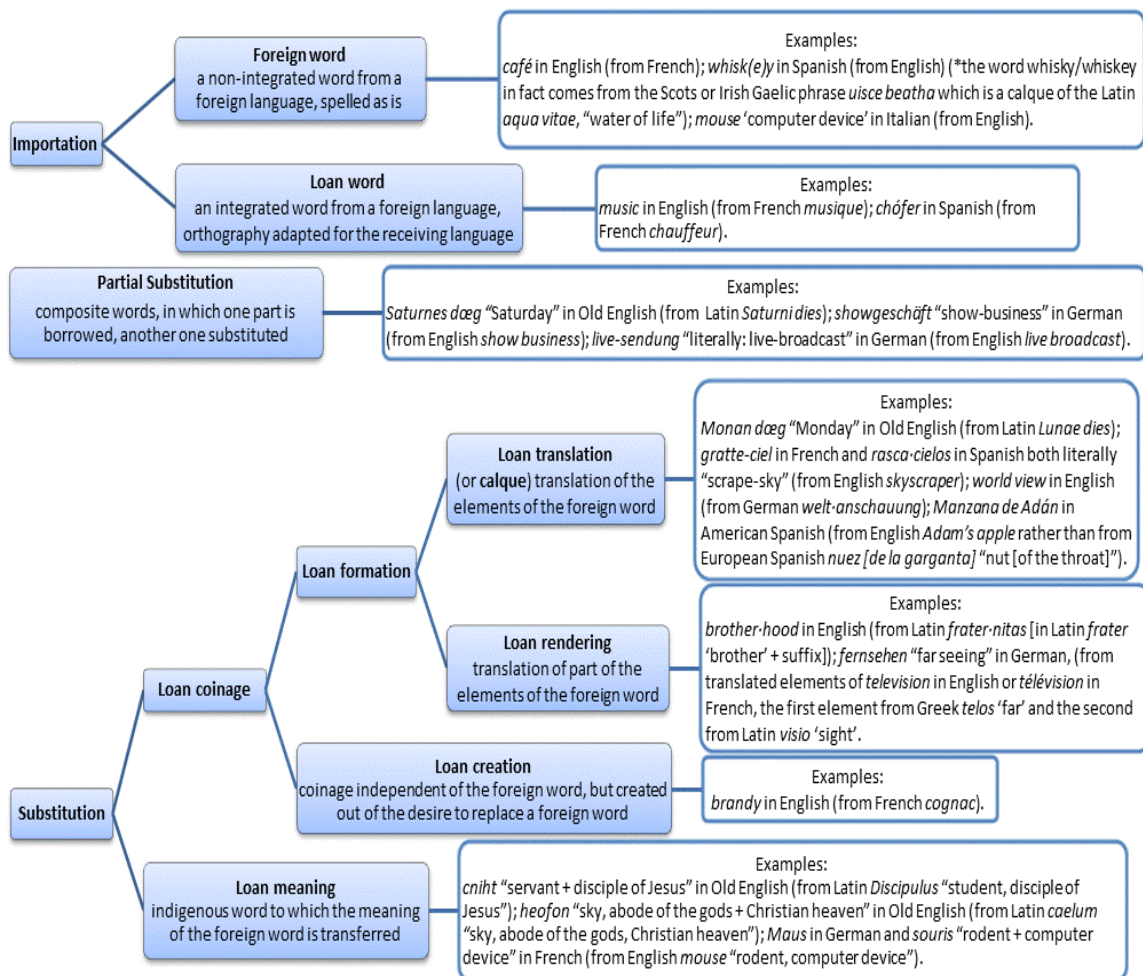
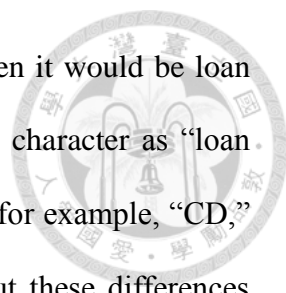


Figure 7 Schematic Representations About Classification on Borrowing by Duckworth (1977)

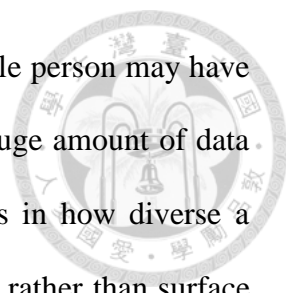


Thus, if only words meet situations summarized in Figure 7, then it would be loan words referred in current study. Thus, those recorded into Chinese character as “loan word,” and those recorded in its donor language as “foreign word,” for example, “CD,” are all referred with generic term “loan word” in current study, but these differences may be captured in the features like whether it is recorded in Chinese characters, or morphological features like being with mixed morphemes or not. Current study assumes that loan word should have its significance in entering lexicon. It may be limited in its phonotactic constraint, but which may also be overcome if it captures unique information what has existed in our experience, but has not lexicalized in native language.

Borrowing in the definition of Thomason and Kaufman (1988) refers to lexical items that can be fluently adopted by speakers of recipient language. The donor languages for targeted loan words included in current study include language system with written forms (English) and without written forms (SouthernMin). Thus, transliteration bearing by Chinese Characters and translation are both included. Besides, lexical borrowings include idiomatic and multi-word expressions are also not excluded. The discussion on this part is aimed to have similar understanding as in Chesley and Baayen (2010) to realize the entrenchment of loan words.

Dissemination across Language Users

Dissemination of users are also taken into consideration. Similar to the calculation of indexicality in Altmann et al. (2011) the dissemination is calculated by dividing the amount of users by the total frequency of the use of the words. This information is based on post author across different boards. The limitation in posts may be its limited sample writing style; however, if a word can be highly disseminated across posts, then it



should be significant in some way. On the other hand, though a single person may have multiple IDs in writing posts, this bias can be less serious for the huge amount of data used in current study. These comparisons all may provide insights in how diverse a word is in being used in different situations and by different people rather than surface using frequency. There are 23 boards used for calculating this value: LoL, ToS, PuzzleDragon, MenTalk, WomenTalk, Boy_Girl, Hate, happy, Sad, NBA, Baseball, movie, Food, BuyTogether, home_sale, Stock, StupidClown, joke, ask, Kaohsiung, Keelung, TaichungCont, and Gossiping.

Assumptions in this section are summarized as in Table 13:

Proposed features in current study	Features in Kjellmer (2000)	Features in Metcalf (2002): FUDGE	Features in Chang(2008)	Features in Kerremans (2015)	Delete
<u>Sociolinguistics</u> loan words or not	O1. It has prestigious and/or exotic connotations.	Unobtrusiveness			
<u>Sociolinguistics</u> Dissemination across users	Number of User IDs/total frequency	Diversity(variety of users and situations)			Posts is assumed to be relatively more stringent in word use

Table 13 Summarized Chart for Sociolinguistics Predictors

3.4.6. Pragmatics

Number of Involved Conceptual Relation Type, Number of Related Concept Words

Semantic relations are important in signifying paradigmatic interaction among lexical items, but the conceptual experiences should also be captured. Different from semantic relations conceptual experiences are habitually linked, so this habitual entrenchment plays a key to revealing human cognition. The experiential concepts the

lexical items involved can be achieved by retrieving data from ConceptNet5⁵. ConceptNet5 originally is built for computers to know about the world and understand humans' written text by constructing a knowledge representation network. The ideal overview of the representation is shown as in Figure 8.

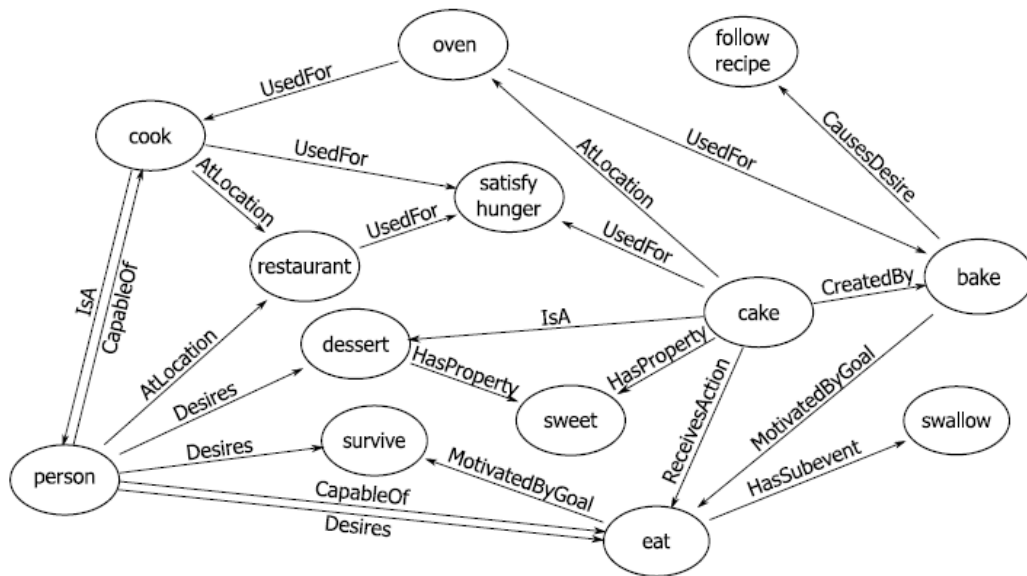
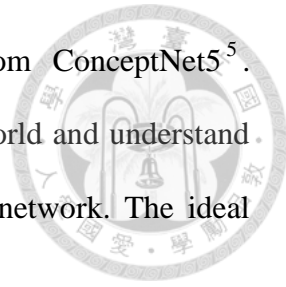
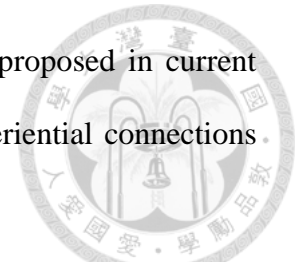


Figure 8 Cluster of Concepts Adopted from Speer and Havasi (2012)

In this representation the scope include both words and phrases in written language and across different languages. The relationships among these words are not just based on lexical definitions but also include the general common knowledge, namely, the related experiences lexicalized in natural language. For example, knowledge about “jazz” should not just lexical defining like “Jazz is a genre of music,” which is caught in the IsA relation defined in ConceptNet5, but also includes facts like AtLocation: “Jazz comes from New Orleans,” or UsedFor: “Saxophone is used for jazz.” The total 21 types of relationships and the sentence pattern to capture these relationships are listed as

⁵ <http://conceptnet5.media.mit.edu/>

in table 14. The different types of relationships the target words proposed in current study involved would be captured to understand the potential experiential connections sustaining the words from contemporary language in use.



Relation	Sentence pattern	Relation	Sentence pattern
IsA	<i>NP</i> is a kind of <i>NP</i> .	LocatedNear	You are likely to find <i>NP</i> near <i>NP</i> .
UsedFor	<i>NP</i> is used for <i>VP</i> .	DefinedAs	<i>NP</i> is defined as <i>NP</i> .
HasA	<i>NP</i> has <i>NP</i> .	SymbolOf	<i>NP</i> represents <i>NP</i> .
CapableOf	<i>NP</i> can <i>VP</i> .	ReceivesAction	<i>NP</i> can be <i>VP</i> .
Desires	<i>NP</i> wants to <i>VP</i> .	HasPrerequisite	<i>NP VP</i> requires <i>NP VP</i> .
CreatedBy	You make <i>NP</i> by <i>VP</i> .	MotivatedByGoal	You would <i>VP</i> because you want <i>VP</i> .
PartOf	<i>NP</i> is part of <i>NP</i> .	CausesDesire	<i>NP</i> would make you want to <i>VP</i> .
Causes	The effect of <i>VP</i> is <i>NP VP</i> .	MadeOf	<i>NP</i> is made of <i>NP</i> .
HasFirstSubevent	The first thing you do when you <i>VP</i> is <i>NP VP</i> .	HasSubevent	One of the things you do when you <i>VP</i> is <i>NP VP</i> .
AtLocation	Somewhere <i>NP</i> can be is <i>NP</i> .	HasLastSubevent	The last thing you do when you <i>VP</i> is <i>NP VP</i> .
HasProperty	<i>NP</i> is <i>AP</i> .		

Table 14 Interlingual relations in ConceptNet Adopted from Speer and Havasi

(2012)

On the other hand, in addition to conceptual relationships current study would also like to probe the experientially correlated words captured in ConceptNet5. The underlying assumption is similar to “O3. It has humorous connotations” proposed in Kjellmer (2000), or “Unobstrusiveness” in Metcalf (2002), namely, the extra connotation or functions in use may influence how a word is being adopted. Connotation is highly associated with experiences occurring in the world but it is usually hard to be quantified in values, so current study would like to understand this part by understanding different types of experiential words that may collocate with the target words. For example the “New Orleans” for “jazz” implies the pragmatic situations for mentioning “New Orleans” and the connotation “jazz” contains. The experiential word types that are related with target words would be retrieved to

represent the pragmatic situations or connotation a word may have.

Activeness in Different Writing Styles, Activeness in Different Themes

The feature “Diversity (variety of users and situations)” proposed from Metcalf (2002) the comparison on activeness in different writing styles and themes are proposed in current study to understand how different information context may correlate to the activeness of words. Different writing styles are the posts in PTT and the comments in PTT because posts are mostly intended monologue for presenting information, which if viewing genre as a continuum should locate closer to the written genre, but comments are more communicative oriented to give dialogue-like feedbacks, which should more like instant response in oral conversation. Given upon the fact that diffused words may hard to be captured in GBNC or Sinica Corpus, the investigation on used writing style may provide an equal comparison base on the using divergence of words. Hence, current study would like to have activeness observation in this aspect in order to capture the living style of different words with the assumption that some words may be more suitable in using in dialogue like feedbacks, but others may be alive in both monologue and dialogue.

Similarly, activeness in different themes provides similar information in exploring possibility of theme-bonding words, which is similar to idea of “topicality” in Altmann et al. (2013), but the calculating method adopted is much closer to Kerremans (2015) in calculating the number of themes that are activated. The judgement on activeness in above factors is based on the normalized accumulative frequency and slope proposed in Chang and Ahrens (2008) in each theme and writing style. There are 9 themes: Games, Gender, Mood, Sport, Lifestyle, Business, Story, Ask, and Geography. The themes are incorporated from 22 boards. The boards are: LoL, ToS, PuzzleDragon, MenTalk,



WomenTalk, Boy_Girl, Hate, happy, Sad, NBA, Baseball, movie, Food, BuyTogether, home_sale, Stock, StupidClown, joke, ask, Kaohsiung, Keelung, and TaichungCont. Posts and Comments are separately calculated their activeness. The summarized information about the posts, comments, themes, and their corresponding boards are in Appendix 2.

On the other hand, as illustrated in the discussion in previous section the quantitative value “frequency” has been proposed as predictive feature in many studies. Frequency may signify important information such as the nameworthiness in Kerremans (2015). However, given on the fact that Revised Constant U is calculated from frequency across time, and the interpretations on frequency can be reflected in factors like activeness in writing styles, themes, and dissemination across users, so this factor is excluded.

Assumptions in this section are summarized as in following chart:

Proposed features in current study	Features in Kjellmer (2000)	Features in Metcalf (2002): FUDGE	Features in Chang(2008)	Features in Kerremans (2015)	Delete
Pragmatics					
Number of Involved Conceptual Relation Type (ConceptNet)					
Pragmatics					
Number of Related Concept Words (ConceptNet)	O3. It has humorous connotations	Unobtrusiveness	Frequency of Use	H5: The nameworthiness of the represented concept or its salience in society promotes conventionalization.	This aspect has been reflected in activeness over themes writing style and dissemination.
Pragmatics					
Activeness in Different Writing	Diversity(variety of users and situations)				



Styles:			
Total frequency and Slope in PTT Posts (Excluding Gossiping for its inclusiveness in various topics)			
Total frequency and Slope in PTT comments (Excluding Gossiping for its inclusiveness in various topics)			
Pragmatics			
Activeness in Different Themes: Number of Activation Themes (Total frequency		Diversity(variety of users and situations)	Posts take the lead in directing themes, so the information retrieved from posts
and Slope in different theme boards (posts) (Including Gossiping for its inclusiveness in various topics)			
Sociolinguistics	Number of User IDs/total frequency	Diversity(variety of users and situations)	Posts is assumed to be relatively more stringent in word use
Dissemination across users		Endurance of the Concept	This has been reflected in Constant U

Table 15 Summarized Chart for Pragmatic Predictors

The predictors adopted in current study and their correspondences in previous studies are all summarized in in the appendix 1.



Chapter 4.

Exploratory Analysis and Modeling

This chapter presents quantitative results of experiments on the three sets of target words as well as qualitative analysis on interactions among words. In section 4.1 and section 4.2 results of Revised Constant U and performance of linguistic factors in these three types of target words are presented. Linguistic regression models for different sets of targets are evaluated in section 4.3. In section 4.4 Competition of words from the same synset is qualitatively discussed. Section 4.5 proposes suggested standards in including words in lexicology by testifying results of inclusion from 8000 Chinese Words.

4.1. Revised Constant U in Three Types of Targets

From Figure 9 the overall distribution of Revised Constant U for all target words is presented. It can be observed that there is a peak for those whose Revised Constant U is zero, which leads to the left skew of the distribution. With this information in mind we look closer at lexical items into three types according to their resources.

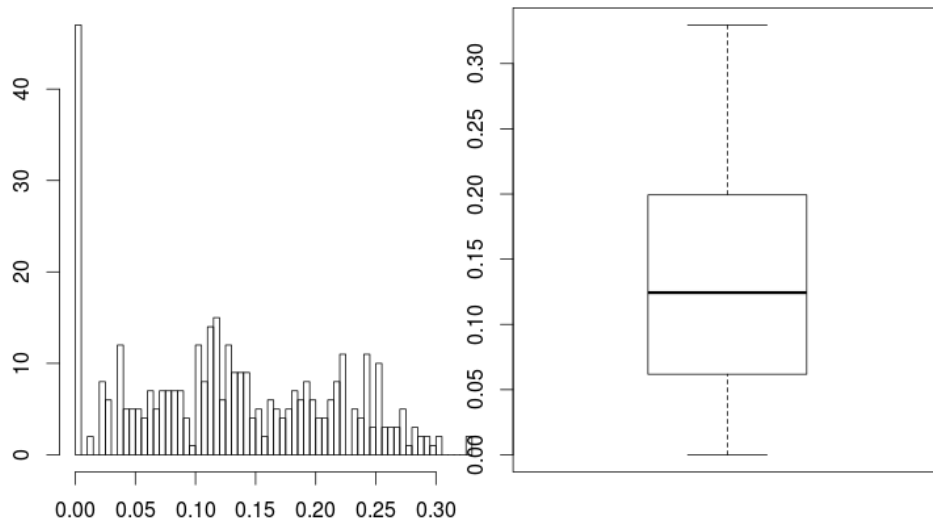


Figure 9 Distribution of Revised Constant U for all Target words

First, **lexical items that are before 1950** are retrieved. There are 8 lexical items with zero Revised Constant U. Among these 8 lexical items five of them are found sporadically used in posts. These words are theoretically existed over century; however, from the perspective of contemporary language use, there are still less stabilized ones in being used in writing style of comments. This indicates that lexical items though may still be comprehensible for their being used or included in more formal written genre, from the aspect of natural language performance they may less tend to be used.

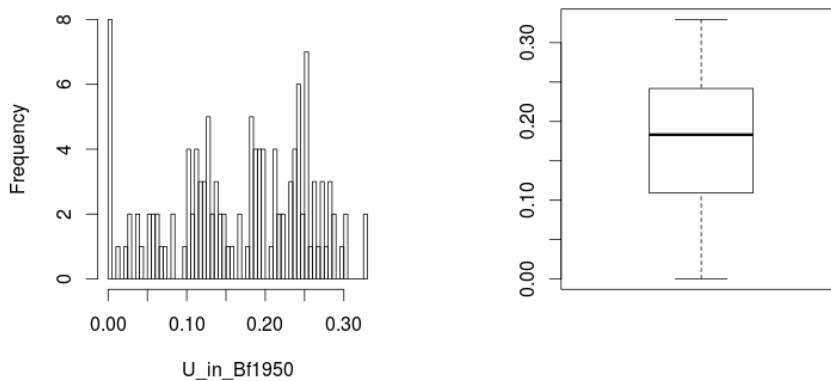
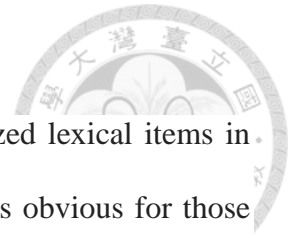


Figure 10 Distribution of Revised Constant U for Target words Before 1950



Taking a look at the top 10 stabilized and the 10 least stabilized lexical items in comments with its log frequency, the difference in overall pattern is obvious for those highly stabilized. Though number of essays in earlier time may be fewer for regular essay deletion in PTT as shown in the lower frequency at the right side in Figure 11, the lexical items are constantly used over time, which is different from those less stabilized ones with sporadically burst point in Figure 12. These burst points may show activation in the view of total frequency, but the real stabilization in use in comments of these words is low.

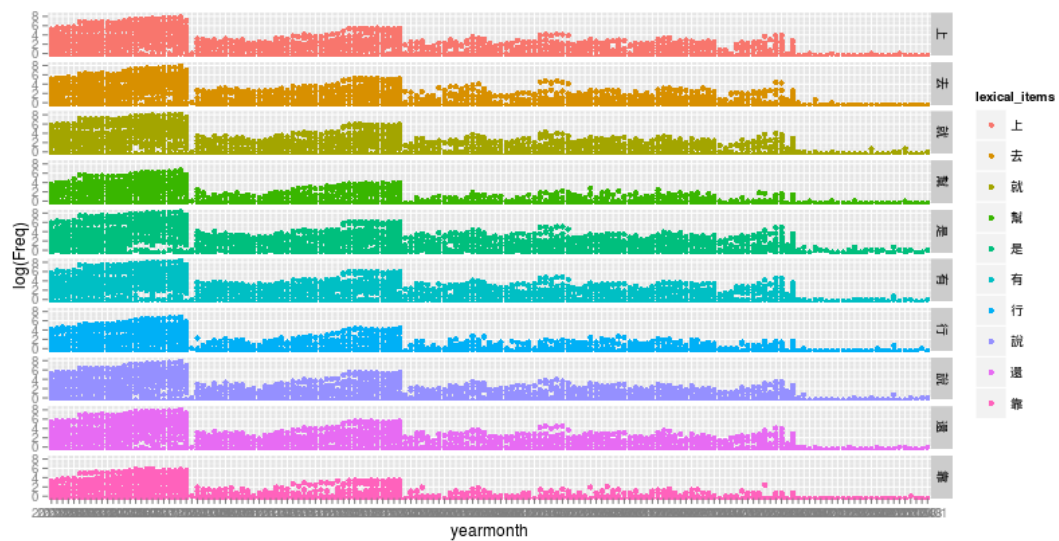


Figure 11 Cross Month Frequency Distribution for Top 10 Target Words Before

1950

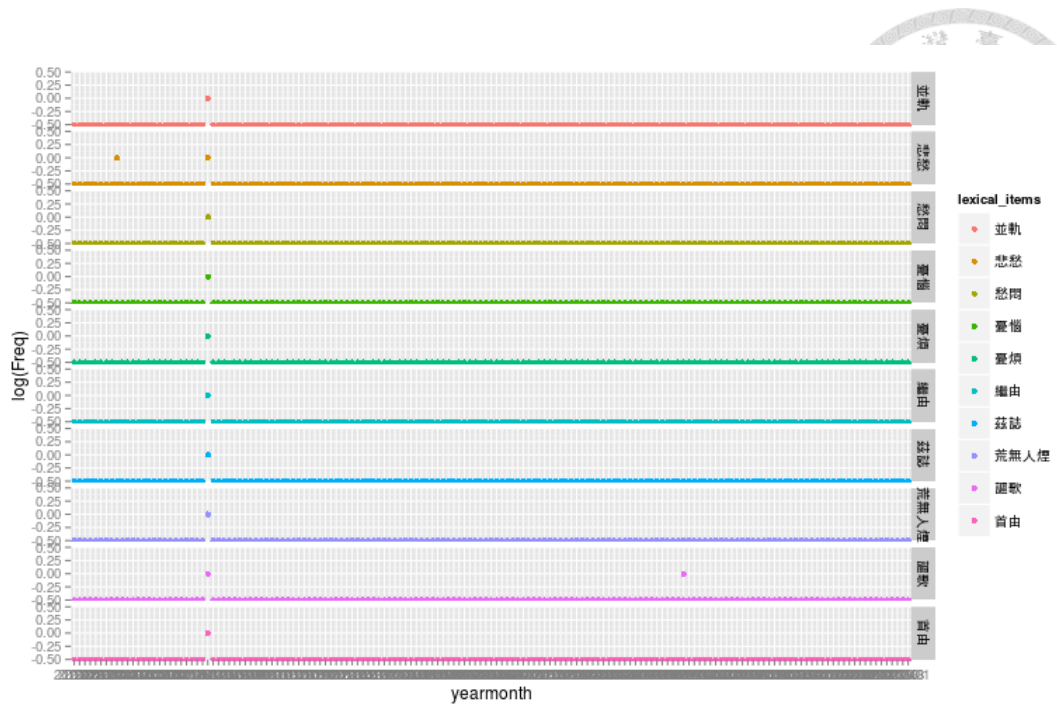


Figure 12 Cross Month Frequency Distribution for Tail 10 Target Words Before 1950

Then, we can take a look at those whose earliest traceable time is **after 1950**, which include 186 lexical items. It seems that there is still left skewed for many of them are with zero value in Constant U.

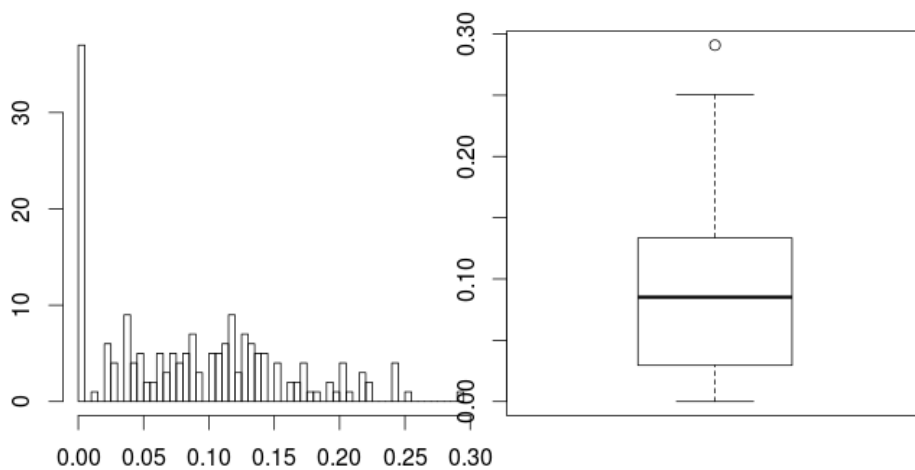


Figure 13 Distribution of by Month Constant U for Target Words After 1950

The lexical items denoted as loan word, such as “哈,” ”讚,” ”菜” are high in Constant U, but given the fact what is loaned is its new sense, and what the Constant U could reflect is the stabilization of its lemma, which should not be singly contributed by their loaned senses. Thus, lexical items denoted as loan word, but is with more than one senses are selected out. There are six such words: "煞到," "三八," "鐵齒," "菜," "讚," "哈". The left are 180 lexical items. The median of Constant U in these words is 0.02950. The Constant U larger than 0 is considered to be stabilized in use. There are 37 lexical items with zero Constant U, which indicates that though they are once diffused around 1950s, they are flash in a pan, and would have less opportunity to be passed down over generation.

Similarly, the cross month frequency patterns of the top 10 stabilized and the 10 least stabilized lexical items in these lexical items are presented in Figure 14 and Figure 15. The pattern difference is also obvious. However, notably the stabilized lexical items in this set are less as hugely used as those top 10 stabilized lexical units in words born before 1950.

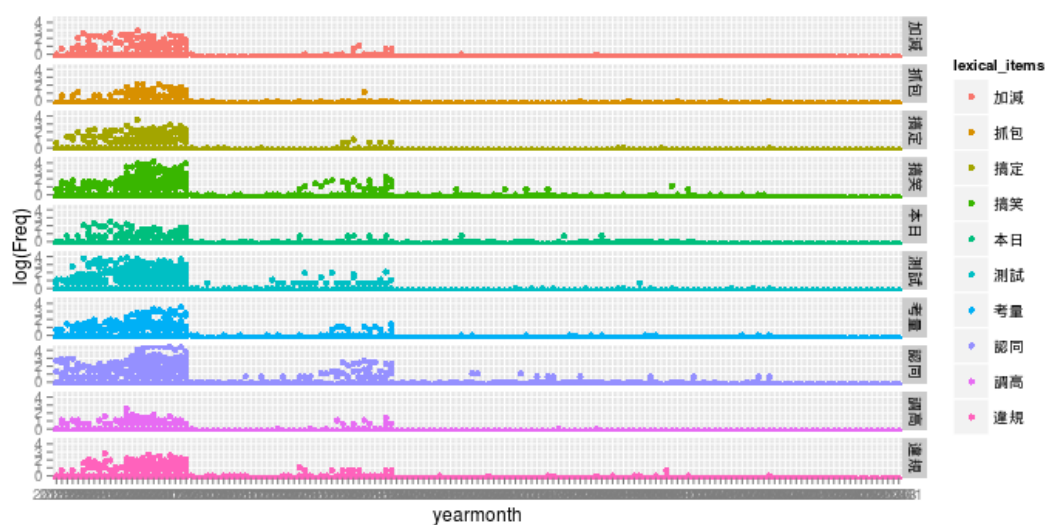


Figure 14 Cross Month Frequency Distribution for Top 10 Target Words After

1950



Figure 15 Cross Month Frequency Distribution for Tail 10 Target Words After

1950

Most of **diffused words** collected from internet corpus are stabilized in use except two lexical items. But, these two lexical items can still be captured its stabilization in being used in posts. This result should be reasonable for they are “born” from internet corpus. Nevertheless, whether they can be entered into lexicon for next generation is still an issue.

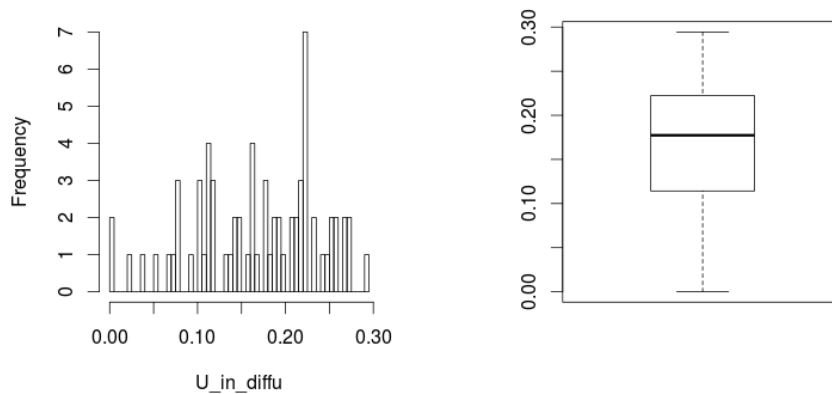


Figure 16 Distribution of by Month Constant U for Diffused words

The stabilization pattern in Figure 18 is similar to those words after 1950s in Figure 14 , but it is generally more activated in single temporal point. Comparing with Figure 11 the diffused words are less activated than lexical items that have been existed over a century. The comparison with words after 1950s may imply that lexical items from Internet may be facilitated with the community it originates in being stably used. But, its stabilization is still less as those who have existed over a century.

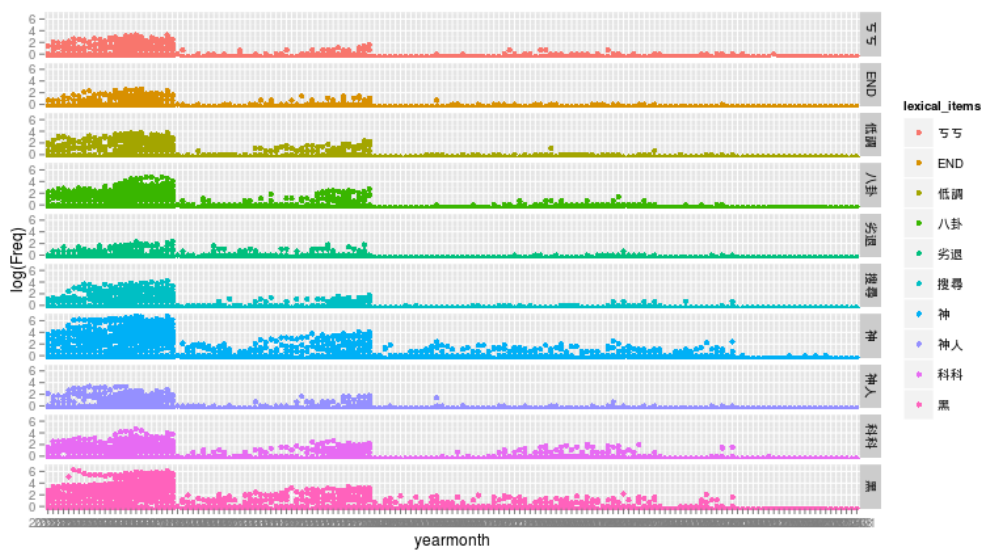


Figure 17 Cross Month Frequency Distribution for Top 10 Diffused Words

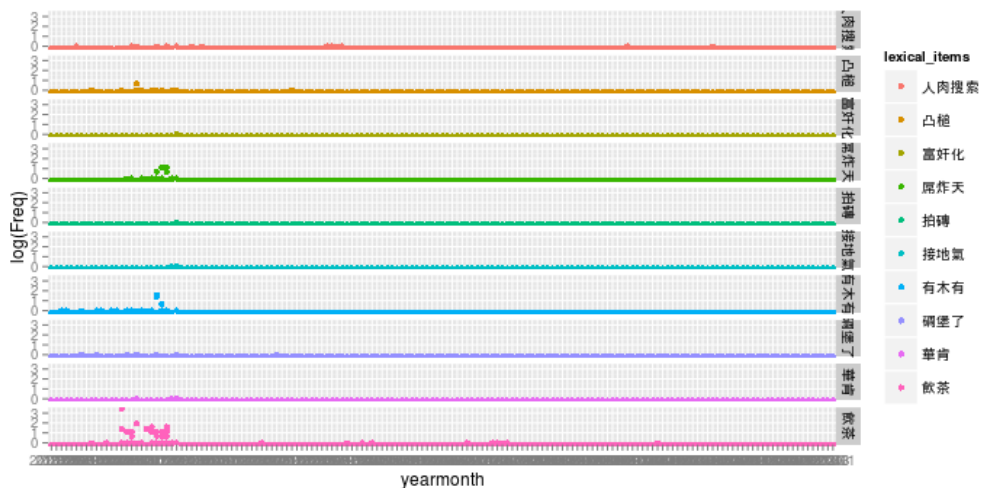


Figure 18 Cross Month Frequency Distribution for Tail 10 Diffused Words

Comparing the Revised Constant U in these three groups in Figure 19, it can observe that lexical items existed over century are stabilized in nearly normal distributed way, but the diffused words are slightly right skewed and the words existed over 50 years are sort of left skewed. Meanwhile, the lexical items traceable after 1950s are relatively less stabilized than those have been existed over a century. The boxplot has shown that the maximum stabilization value of lexical items traceable after 1950s locates around the median part to the lexical items existing over a century. Those words that have been existed around 50 years may best resemble the potential stabilized situation for the presently diffused words.

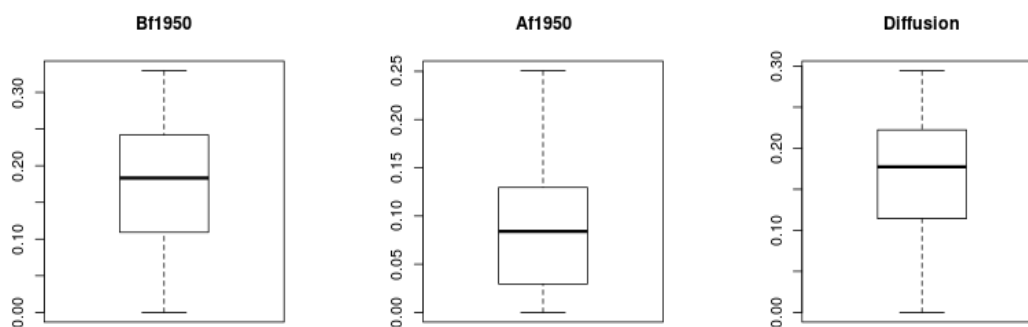


Figure 19 Distribution of Revised Constant U of Three Sets of Target Words

4.2. Performance of Linguistic Factors in Target Words

This section briefly summarizes linguistic characteristics of each set, and some of them may show differences among these sets. The differences are not testified with statistics for the main purpose is to perceive and explore potential differences among different sets of words.

The **number of syllables** for three sets of words has shown that though disyllabic are rich in all sets, words before 1950 are with more monosyllabic lexical items and

with relatively homogeneous syllable type. Words born after 1950 and diffused words are similar in with relatively various syllable types.

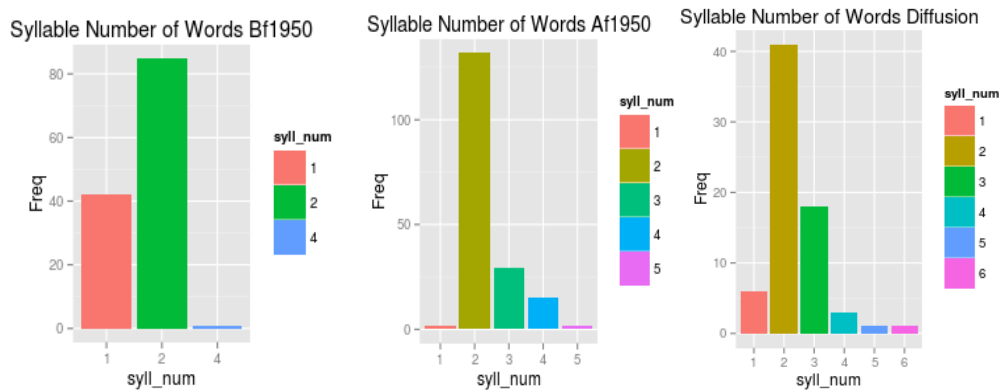
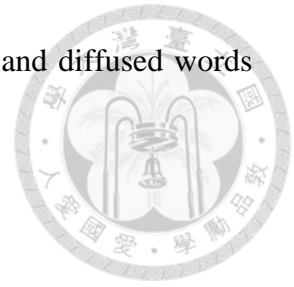


Figure 20 Number of Syllables for Three Sets of Target Words

Morphologically, the results show that words born after 1950 and diffused words are quite similar to each other in having **mixed originated morphemes**.

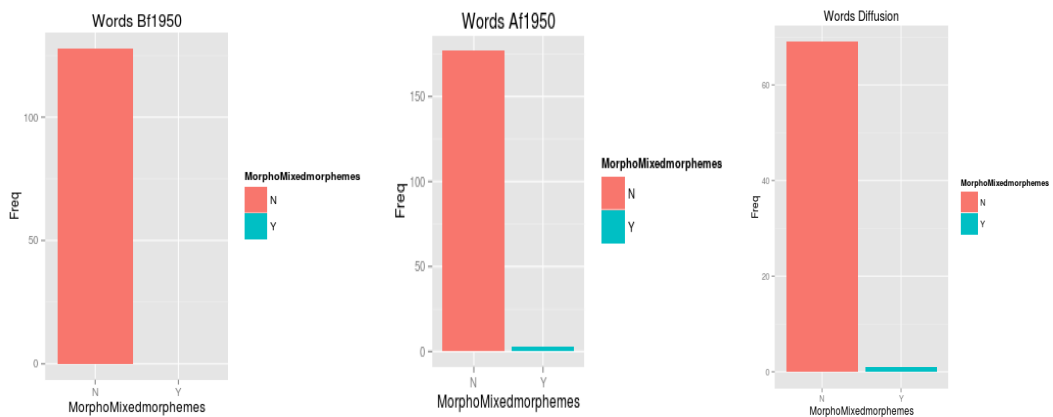


Figure 21 With Mixed Originated Morphemes or not

But, from currently collected data as shown in Figure 22 only diffused words involve morphemes not **encoded in Chinese**.

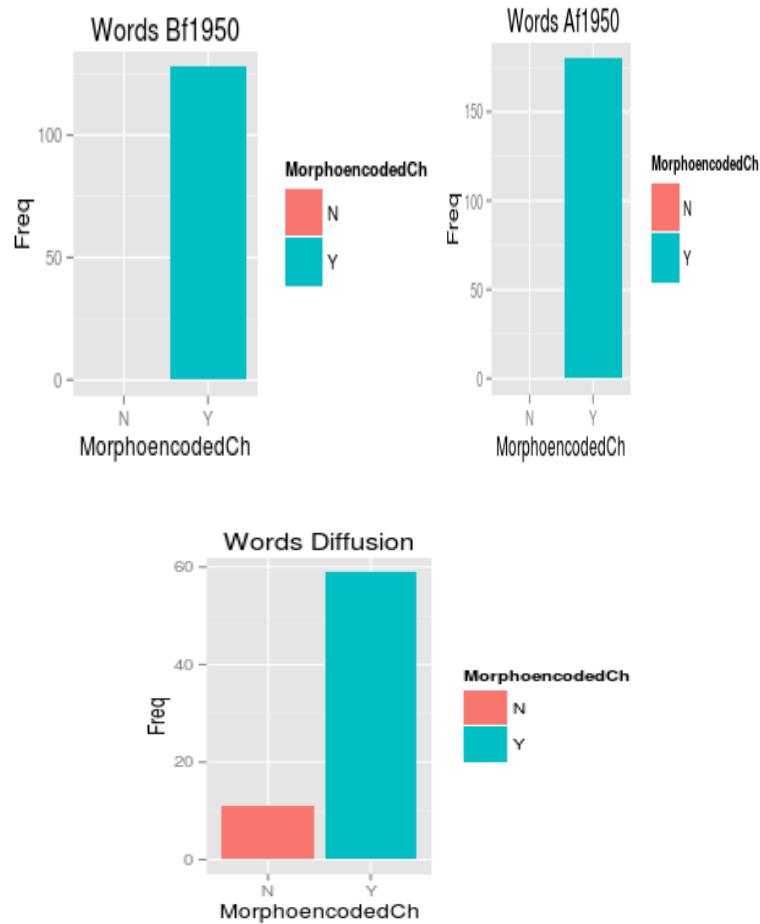


Figure 22 Encoded in Chinese or not

Component richness from the angle of realized productivity, it shows that words born after 1950 and diffused words are with relatively lower realized productivity in overall than those born before 1950, which may be due to the fact that there are most words from loaned words in these two sets of words.

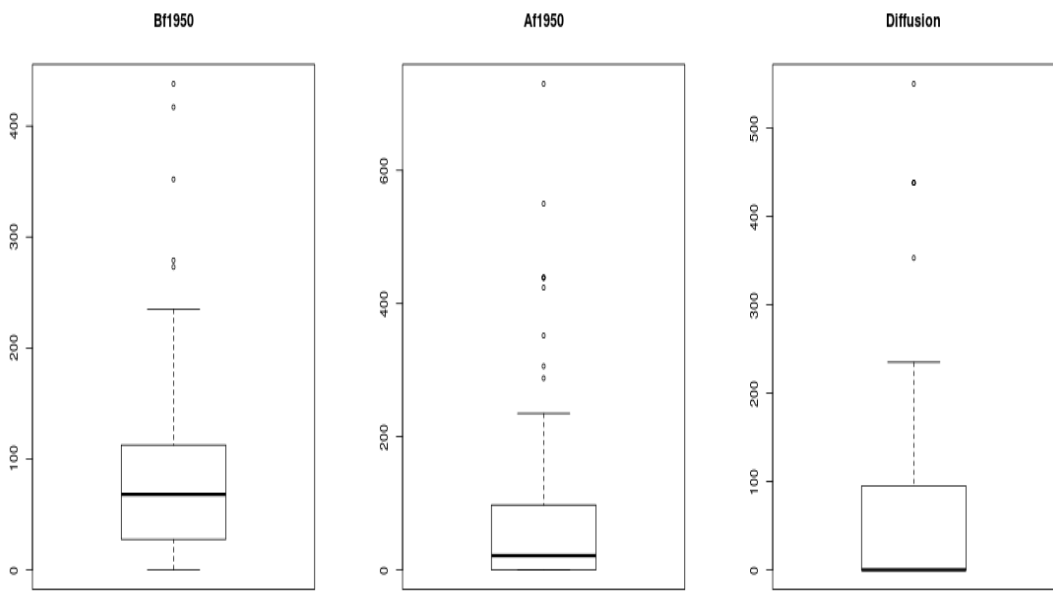


Figure 23 Component Richness: Realized Productivity

When viewing from type-token ratio, it shows that words born before 1950 and words born after 1950 are relatively less than diffused words.

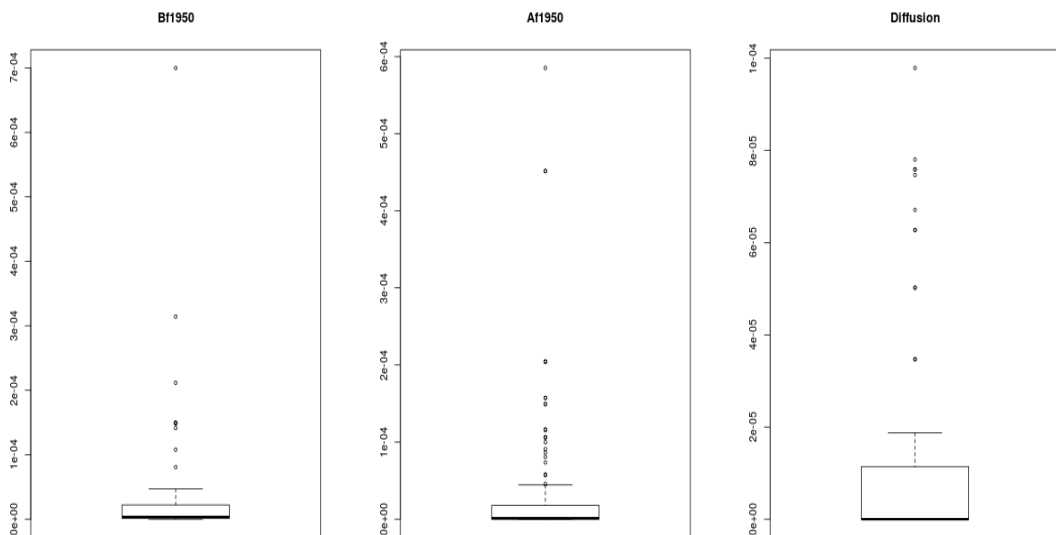
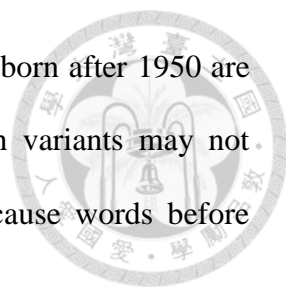


Figure 24 Component Richness: Type- Token Ratio



The **variants** may be most rich for words before 1950. Words born after 1950 are less with variants. This may be with implication that words with variants may not weaken the possibility of being conventionalized into lexicon because words before 1950 are the set with members own most variants.

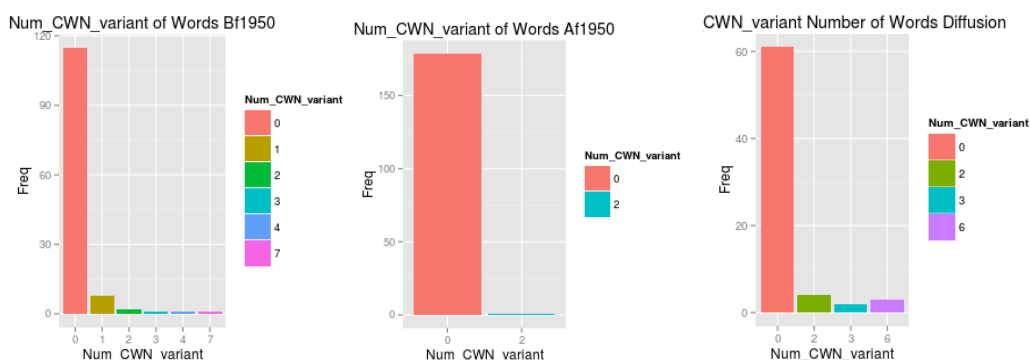


Figure 25 Distribution of Variants

The distribution of **parts of speech** for each set is shown in Figure 26. Verbs are most rich ones as intended design.

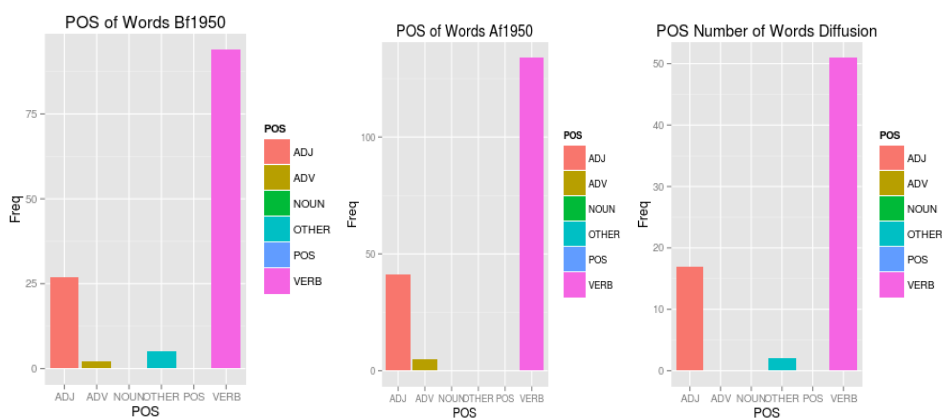


Figure 26 Distribution of Parts of Speech

The syntagmatic information on **co-occurrences** shows that words born before 1950 do have extreme high number of different types of words collocating before or after it, but most of them are similar to the way of words born after 1950 and diffused words. The summary statistics for before and after co-occurrences for each set are

summarized in and the boxplots are shown in Figure 27 and Figure 28.

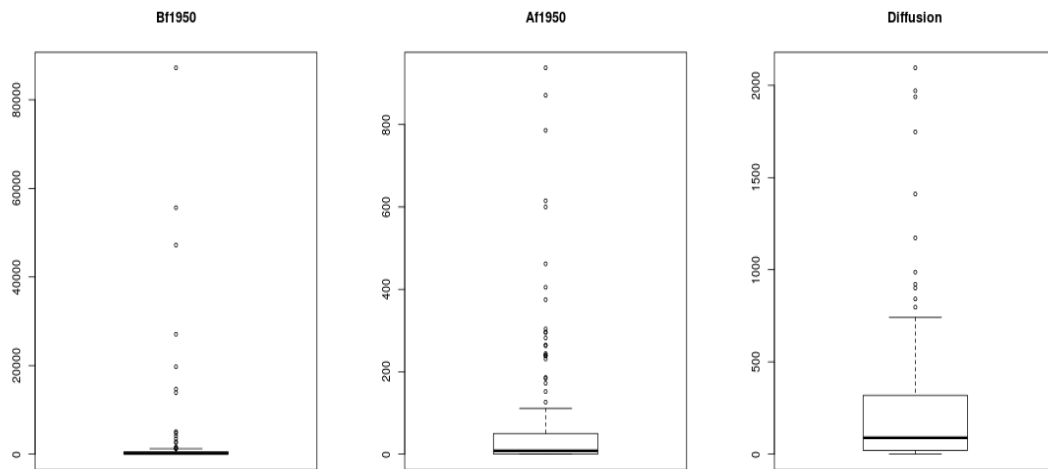
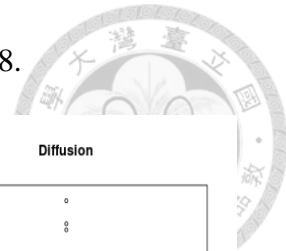


Figure 27 Distribution of Co-occurring Types (Before Target Words)

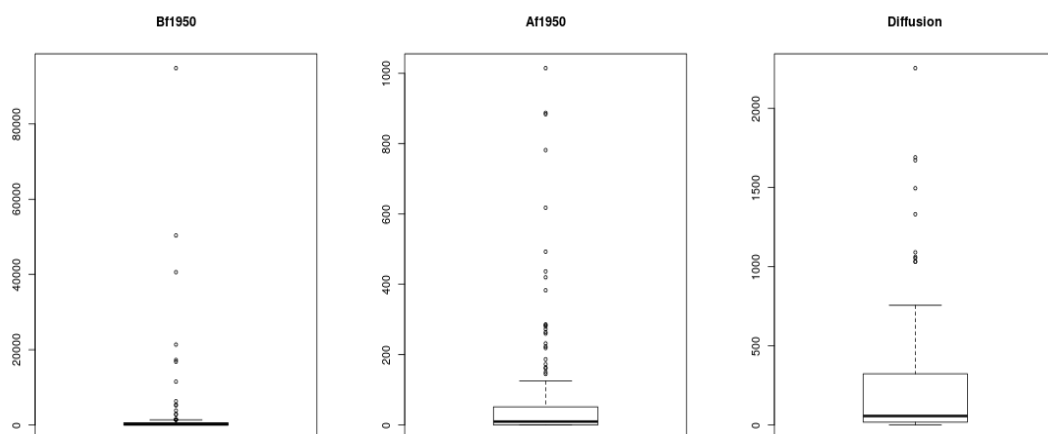


Figure 28 Distribution of Co-occurring Types (After Target Words)

The summarized statistics indicate that words born after 1950 are with less co-occurring word types than the other two sets.

Summary Bf_co-occurrences Set of Target Af_co- occurrences

Statistics

Min.	0.00	Words before 1950	0.0
1st Qu.	21.25	Words before 1950	22.0
Median	133.00	Words before 1950	137.0
Mean	2479.00	Words before 1950	2420.0
3rd Qu.	502.20	Words before 1950	564.2
Max.	87250.00	Words before 1950	94850.0
Min.	0.00	Words after 1950	0.00
1st Qu.	0.00	Words after 1950	0.00
Median	8.00	Words after 1950	9.50
Mean	65.27	Words after 1950	67.78
3rd Qu.	50.00	Words after 1950	51.25
Max.	938.00	Words after 1950	1015.00
Max.	0.00	diffusion	0.0
Min.	20.25	diffusion	18.0
1st Qu.	88.50	diffusion	56.5
Median	319.60	diffusion	298.0
Mean	306.00	diffusion	315.0
3rd Qu.	2096.00	diffusion	2255.0

Table 16 Summary Statistics for Before and After Co-occurring Word Types for Each Set

Semantically, words born before 1950 are with richer **number of senses** than other sets. The outliers in diffused words are “黑,””腿,”and”神.” They are semantic neologies that have existing lemma with multiple meanings.

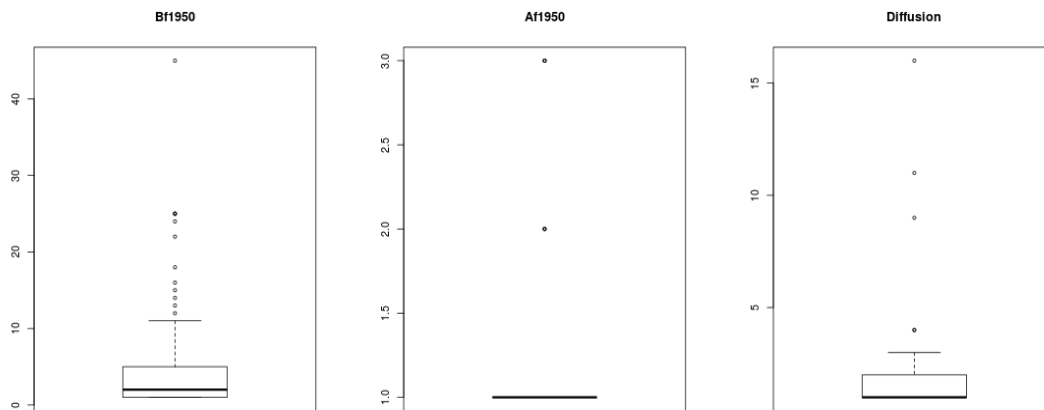
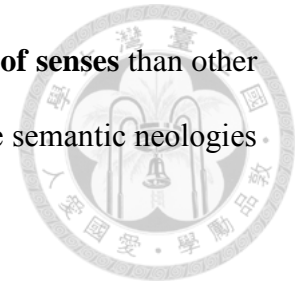


Figure 29 Distribution of Number of Senses

The results of **semantic relation** have indicated that words born before 1950 are with more near synonyms, antonyms and hypernyms than others. Words born after 1950 and diffused words involve none of hyponymic relation. The synonymic relations also show similar trend with richer information for words that have existed over a century as shown in Figure 30 . The sum of total relations also reflects this phenomenon. Though the limited relations in words after 1950 may be due to the limited information from CWN, it is still reasonable to suppose that words existing over a century are with richer semantic network as bonding.

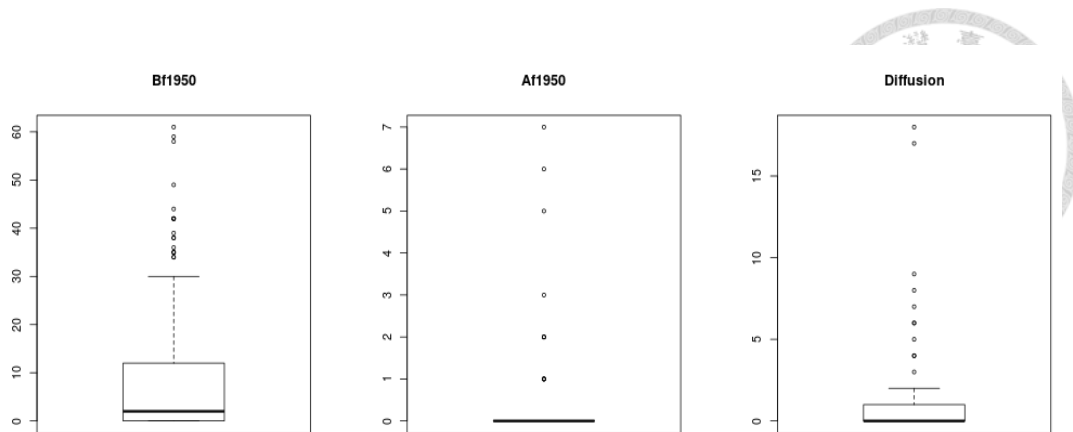


Figure 30 Distribution of Number of Involved Synonymous Relation

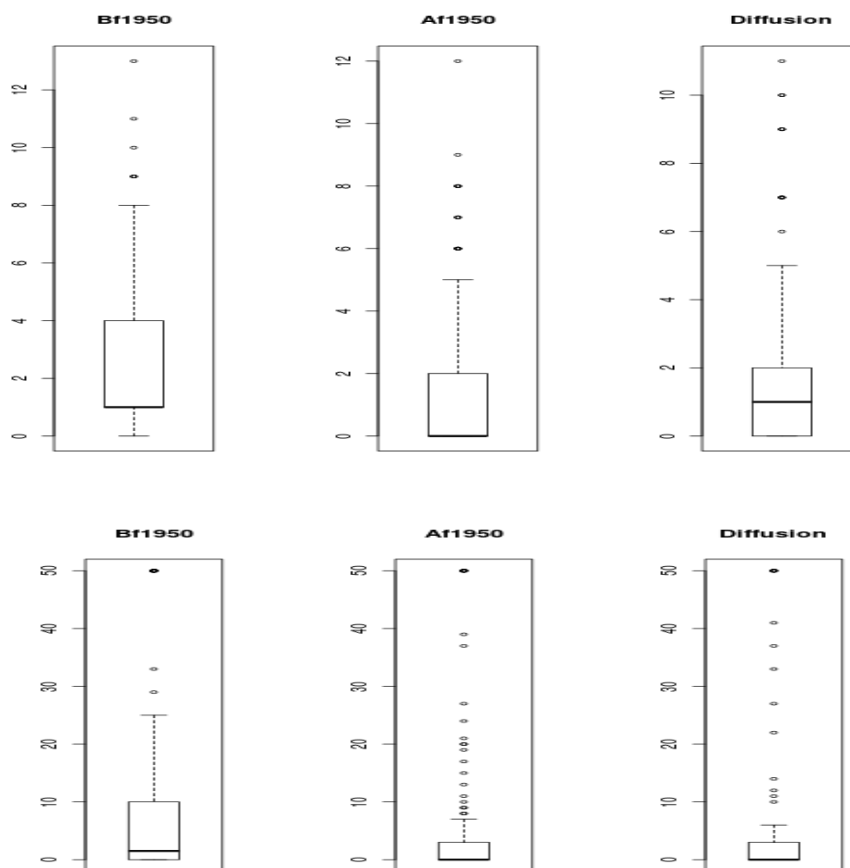
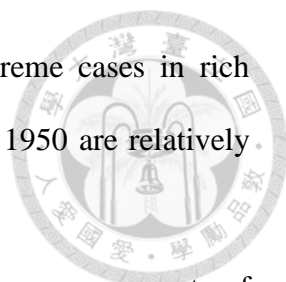


Figure 31 Upper Panel: Distribution of Involved Conceptual Relationships, Lower Panel: Distribution of Related Conceptual Words

As shown in Figure 31, different **conceptual relation types** and **related**



conceptual words it shows that three sets of words all with extreme cases in rich conceptual relations or related conceptual words, but words before 1950 are relatively higher than others.

Active situations in posts and comments show some differences across sets of words and across writing style. The activeness defined here adopts the threshold value proposed by Chang and Ahrens (2008) as discussed in section 2.2.2. If the target word is active in one of the retrieved themes in post or comment, then it will be categorized as active in that writing style. It shows that words born before 1950 are relatively higher than others in both writing style, especially in post style. Words born after 1950 are relatively lower in both styles as compared with the other two sets of words, and they are also more active in posts than in comments. Diffused words tell a different story. They are relatively active in both styles than words born after 1950, but less active than words born before 1950. The more active style for them is in comments. This may imply two points. First, different usages of words in different oriented writing styles may exist. Second, if we take posts as with information structure closer to formal writing, and comments as with information structure closer to casual oral speaking way as well as recognize that comments are more feedback oriented than posts, then it may imply that diffused words are more correlated in oral style and “diffused” in interaction.



Figure 32 Actively used in Posts or not

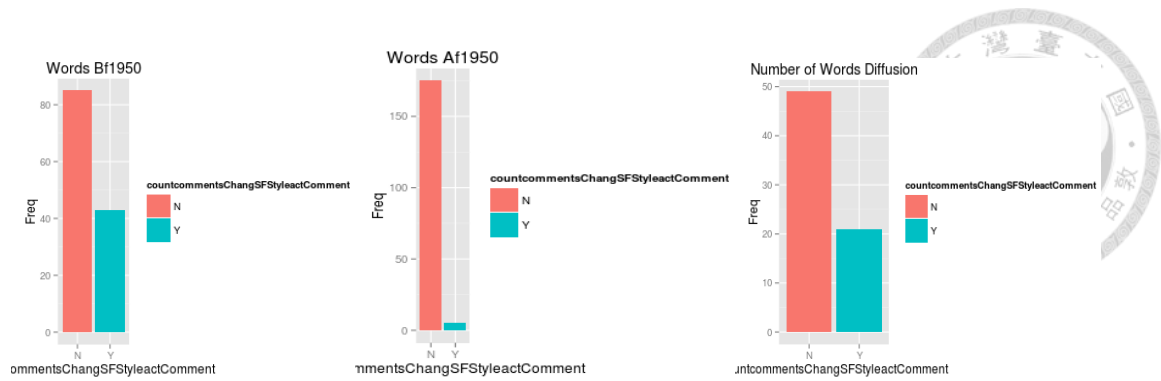


Figure 33 Actively used in Comments or not

The distribution of **loan words** for each set is shown in Figure 34, it has signified that loan words are more in words after 1950 and diffused words in our collected data. These two sets are much more similar in this aspect.



Figure 34 Distribution of Loan Words in Each Target Word Set

Though diffused words have the highest outlier which is with 6.059 in **dissemination**, the overall disseminated value in words before 1950 is higher than the other two sets, which indicates their highly entrenchment across different users.

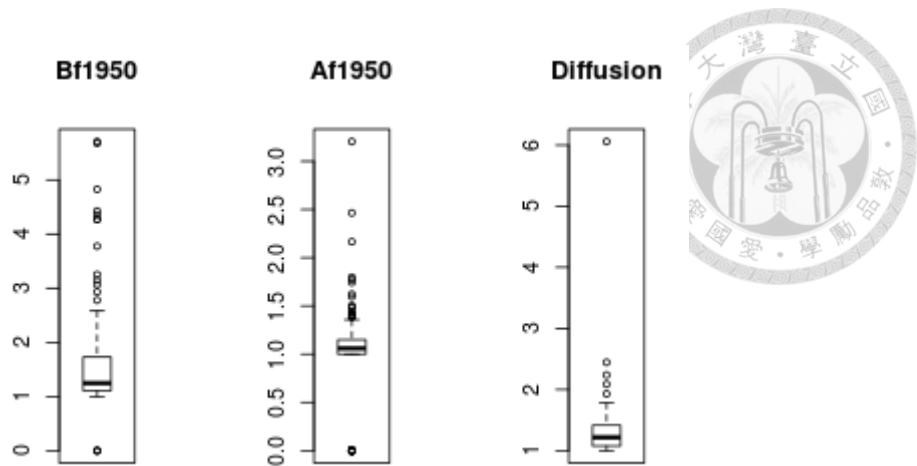


Figure 35 Dissemination in Each Target Set of Words

As shown in above exploratory data, words after 1950 and diffused words are more similar to each other in many linguistic aspects, but they have quite different Revised Constant U. This fact has two implications. First, the power of being in diffuse contributes a lot in being used stably. Second, words born after 1950 are comparatively able to reflect the possible future living situation of currently diffused words.

4.3. Linguistic Regression Models for Three Sets of Words

Before moving on building regression models, the density plot for checking normality of Revised Constant U is presented in Figure 36. From the top to the bottom it presents separately original data, square root, log transformation, and inverse in transforming data for easiness in modeling. From the plots it shows that the original data is not in normal distribution, but it looks better with log transformation, though is still slightly left skewed.

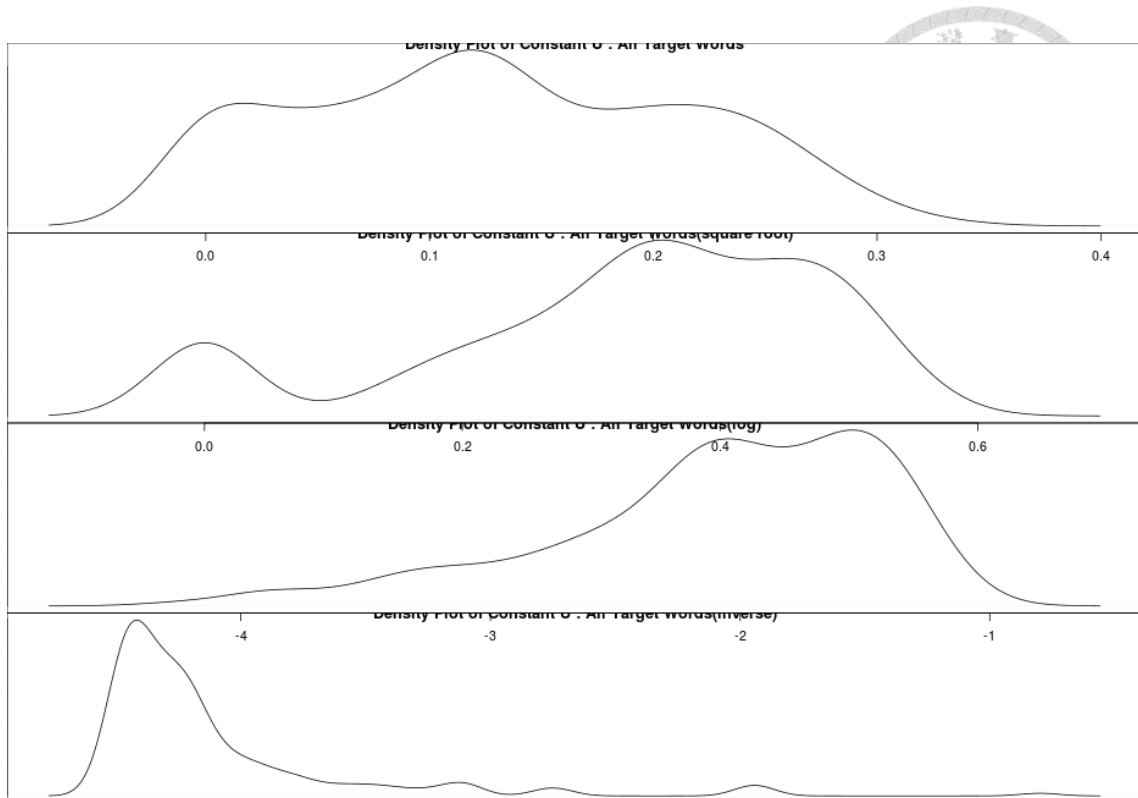


Figure 36 Density Plots for Revised Constant U of all Target Words: from top to the bottom shows separately the distribution of original data, of log transformation, of square root, and of inverse in transforming data

Density plots of Revised Constant U for words born before 1950, born after 1950, and diffused words are also shown in Figure 37. Similar to the boxplot presented in Figure 19, words before 1950 and diffused words are highly similar. But, all of them are not normal distribution, so the log transformation is adopted as shown in Figure 38, which though is still slightly left skewed.

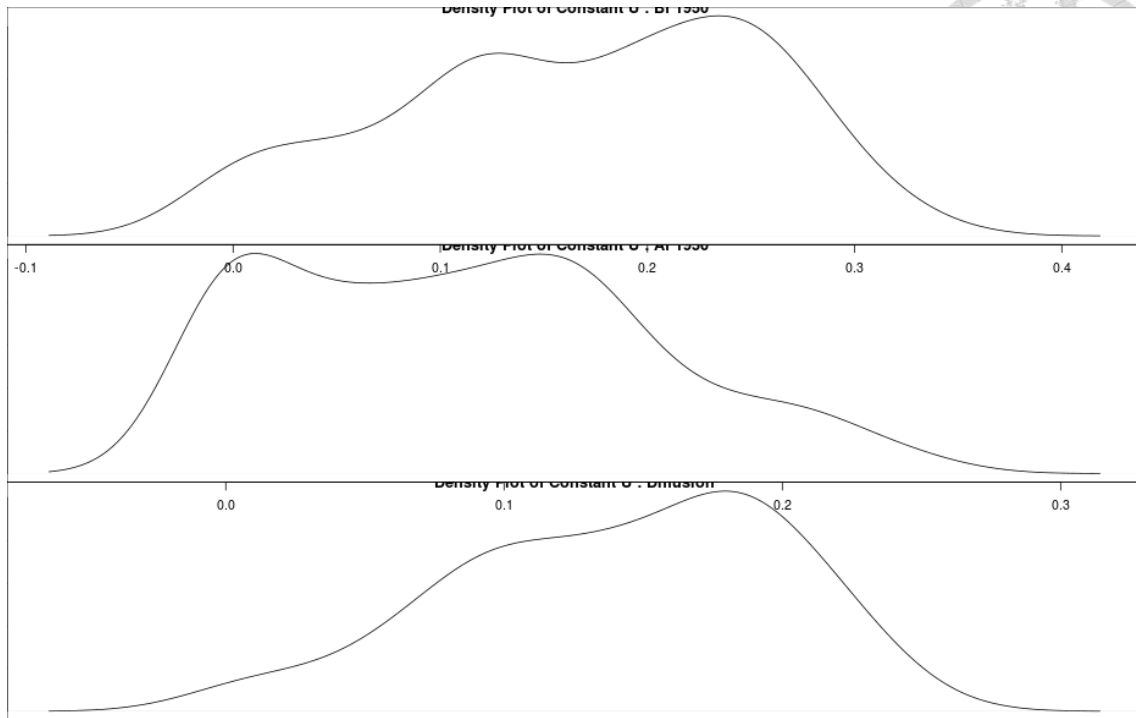


Figure 37 Density Plots for Constant U of 3 Sets of Words: from top to the bottom shows separately the distribution of original data in Words Before 1950, Words after 1950, and Diffused Words

Similar to the prediction model built for understanding entrenchment of loan word by Chesley and Baayen (2010) there is also non-normality in our response variable, the Revised Constant U. However, their conducting in non-parametric random forests has shown the reliability of the results from regression model. Thus, current study will still adopt regression models to understand linguistic factors driven behind Revised Constant U. Different from their choosing only main effects and two-way interactions. There are 384 lexical items and total 19 predictors under 6 proposed linguistic aspects in current exploratory. With concern on degree of freedom current study is going to build model for each linguistic aspect. Models for all 384 target words, words after 1950, words before 1950, and diffused words are separately presented.

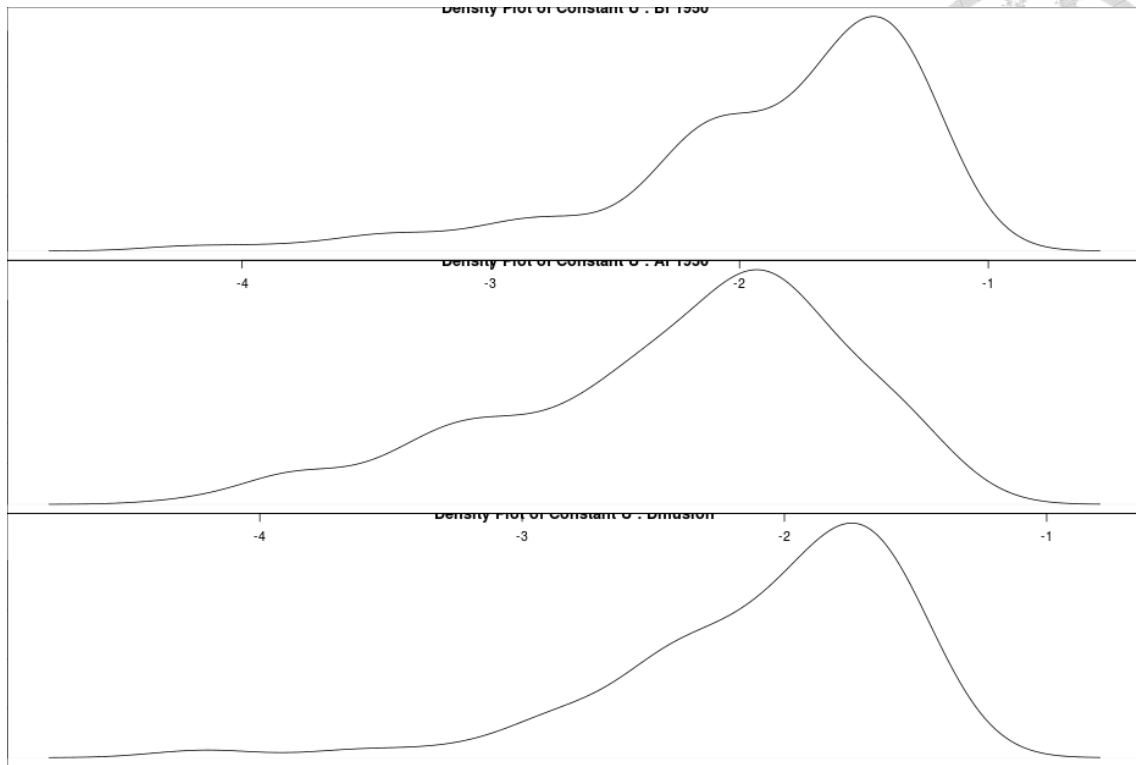
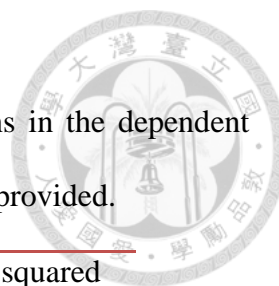


Figure 38 Density Plots for Constant U of 3 Sets of Words: from top to the bottom shows separately the distribution of log transformation of Revised Constant U in Words Before 1950, Words after 1950, and Diffused Words

In addition to linear regression models for understanding highlighted linguistic aspects in each set of words, logistic regression models are built to sketch differences between words existing over a century and diffused words as well as a proposed prediction model based on words born after 1950.

4.3.1. Revised Constant U and Phonology

From density plots in previous section the non-normality of Revised Constant U has been revealed, and as anticipated the residual plots are not ideal, which is just as the way in Chesley and Baayen (2010). However, given the fact that current study testifies six linguistic aspects separately, so the less ideal in residual plots of every model should be reasonable for a single linguistic aspect may not be enough to explain the surface performance in Revised Constant U. The linear regression between Revised Constant U



and Phonology of different target words is summarized in Table 17. Multiple R-squared indicates the percentage of variations in the dependent variable explained by the model, and the Adjusted R-squared is also provided.

Type of Target Words	Multiple R-squared	Adjusted R-squared
All Target Words	0.2129	0.2109
Words Before 1950	0.3106	0.3052
Words After 1950	0.163	0.1583
Diffused Words	0.05621,	0.04233

Table 17 Revised Constant U and Phonology

This indicates that number of syllable plays a relatively larger role in explaining variation to words before 1950 than those diffused recently.

4.3.2. Revised Constant U and Morphology

In exploring morphological aspects, present work performs backward variable selection starting with main effects and interaction for all predictors in morphological aspect, and used the Akaike Information Criterion (Akaike 1974) to eliminate superfluous predictors. For different target word types the yielded formula of powerful predictors is different.

Type of Target Words	Multiple R-squared	Adjusted R-squared
All Target Words	0.102	0.06877
Words Before 1950	0.2012	0.1684

Words After 1950	0.03522	0.02432
Diffused Words	0.1575	0.1056



Table 18 Revised Constant U and Morphology

Morphological variables also show advantageous explanation ability to words before 1950. Among the predictors the relative important ones are type-token ratio of component richness as well as interaction between type-token ratio of component richness and realized productivity of component richness.

4.3.3. Revised Constant U and Semantics

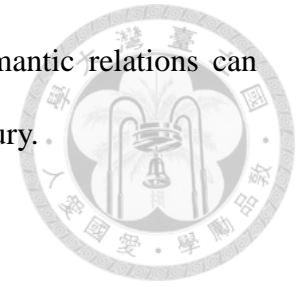
In exploring semantic predictors, present study performs backward variable selection starting with main effects and interaction for all predictors in semantic aspect, and used the Akaike Information Criterion (Akaike 1974) to eliminate superfluous predictors. For different target word types the yielded formula of powerful predictors is different.

Type of Target	Multiple R-squared	Adjusted R-squared
Words		
All Target Words	0.4535	0.3844
Words Before 1950	0.6074	0.4695
Words After 1950	0.1042	0.07847
Diffused Words	0.3048	0.2263

Table 19 Revised Constant U and Semantics

Semantic variables also show significant advantageous explanation ability to Words

before 1950. This may imply that the richness in senses and semantic relations can explain the constantly in use for words that have existed over a century.



4.3.4. Revised Constant U and Syntax

In exploring syntactic predictors, current study performs backward variable selection starting with main effects and interaction for all predictors in syntactic aspect, and used the Akaike Information Criterion (Akaike 1974) to eliminate superfluous predictors. For different target word types the yielded formula of powerful predictors is different.

Type of Target	Multiple R-squared	Adjusted R-squared
Words		
All Target Words	0.3127	0.2846
Words Before 1950	0.5029	0.451
Words After 1950	0.6629	0.6408
Diffused Words	0.5522	0.485

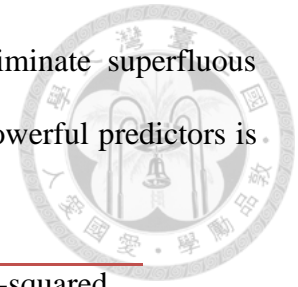
Table 20 Revised Constant U and Syntax

Different from previous models, syntactic predictors show advantageous explanation ability to words after 1950. The main effects and the interaction among the three variables, parts of speech, number of before-word co-occurring type, and number of after-word co-occurring type are all significant in the model.

4.3.5. Revised Constant U and Pragmatics

In exploring pragmatic predictors, current study performs backward variable selection starting with main effects and interaction for all predictors in pragmatic aspect,

and used the Akaike Information Criterion (Akaike 1974) to eliminate superfluous predictors. For different target word types the yielded formula of powerful predictors is different.



Type of Target	Multiple R-squared	Adjusted R-squared
Words		
All Target Words	0.5786	0.5685
Words Before 1950	0.7606	0.7424
Words After 1950	0.3012	0.2812
Diffused Words	0.5003	0.4157

Table 21 Revised Constant U and Pragmatics

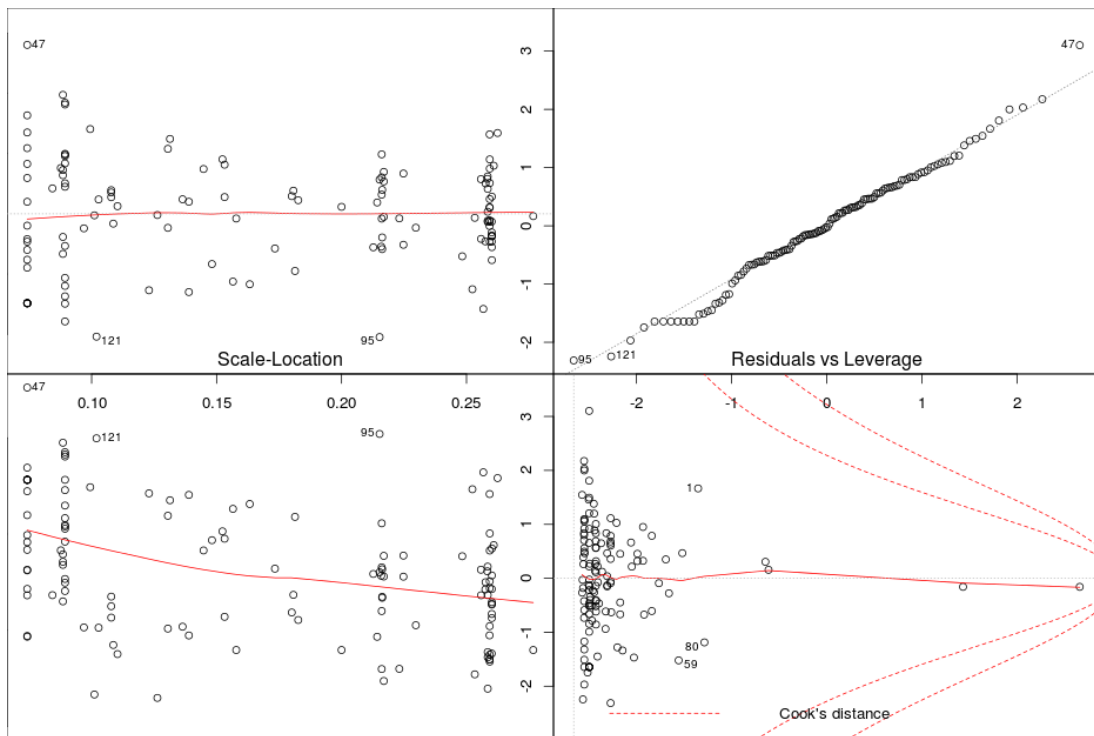
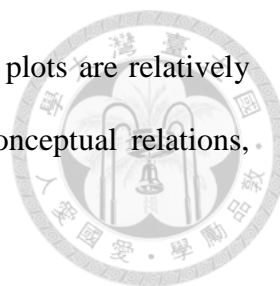


Figure 39 Residual Plots for Pragmatic Model for Words Before 1950

Similar to previous models, pragmatic predictors show advantageous explanation



ability to words before 1950. Besides, the requirements in residual plots are relatively meet in Figure 39. This implies the importance of experiential conceptual relations, writing styles, and themes in contributing being stably used.

4.3.6. Revised Constant U and Sociolinguistics

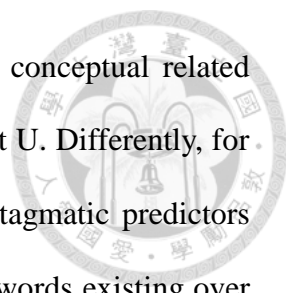
In exploring sociolinguistic predictors, current study performs backward variable selection starting with main effects and interaction for all predictors in sociolinguistic aspect, and used the Akaike Information Criterion (Akaike 1974) to eliminate superfluous predictors. For different target word types the yielded formula of powerful predictors is different.

Type of Target	Multiple R-squared	Adjusted R-squared
Words		
All Target Words	0.3765	0.373
Words Before 1950	0.4768	0.4684
Words After 1950	0.2891	0.277
Diffused Words	0.2599	0.2263

Table 22 Revised Constant U and Sociolinguistics

Similar to previous models, sociolinguistic predictors show higher explanation ability for words before 1950, but they are not so significant as pragmatic factors.

When comparing all these factors we can discover that words born before 1950 can be best statistically accounted by pragmatic factors. Activeness in comments, activeness

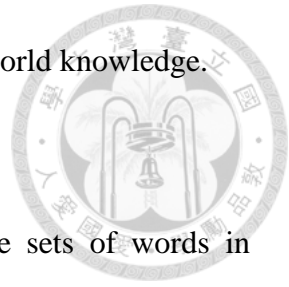


in posts, number of involved conceptual relations, and number of conceptual related word type account 76% behavioral performance of Revised Constant U. Differently, for words born after 1950, those who exists only about 50 years, syntagmatic predictors show advantageous ability to account. These two results imply that words existing over centuries are highly correlated with rich pragmatic experiences accumulated in daily life as well as using context selected by language users. Thus, habitual experiential association plays an important role in understanding whether a word can live longer and be used over generations. Besides, the context a word is used may imply the spread of usage of that word. To be used in a variety of contexts highlights the adoptive ability of word in language and its important role in conveying messages. Nevertheless, for words coined in more recent years the syntactic compatibility is the key. Types of word collocate with the target become important indicator. With the temporal information and the correlated linguistic features we may propose that the usability of lexical expressions may first be decided by their compatibility with already existed words. Such compatibility is more than being paradigmatically antonymous or synonymous, but more about whether the target words semantically and syntactically cooperate with other words or not. Stronger structural compatibility means that the word is being accepted by existed lexicon and its significant role in conveying information. Then, the further sustainability relies on deeper entrenchment with world knowledge as well as suitability in being used in different registers as indicated by the outstanding performance of words before 1950 in pragmatics. Results in current discussion show that as the days progress the important factor influencing life of a word may move from more context-limited syntactic relation to larger pragmatic information related to world knowledge we have entrenched with the word. A word is more than a sign carrying

literal meanings, but a crystal of human cognition, experience, and world knowledge.

4.3.7. Logistic Regression Model

From the multiple linear regression models built for three sets of words in previous section we can realize that though both diffused words and words before 1950 all have relatively high values in Revised Constant U, the linguistic factors driven behind are different, thus it is with interest to go beyond this surface behavioral phenomenon in order to understand what factors can distinguish the two. A logistic linear regression models is conducted, in which words before 1950 are viewed as conventionalized, and diffused words are viewed as not conventionalized. The main effects are evaluated except the total number of semantic relations because it is statistically collinearity with other semantic variables. With the results of parametric statistic Wald test in Table 23 we can realize that number of syllables, number of synonymic relations, number of near synonyms, whether it is actively used in content of comments, and whether it is from other language are variables statistically significant in distinguishing these two set of words.



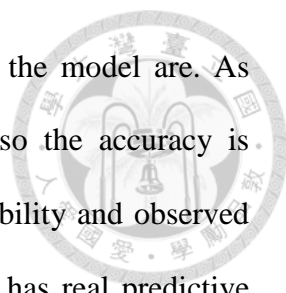
Wald Statistics		Response: hypothesis		
Factor	Chi-Square	d.f.	P	
syllable_num	7.49	1	0.0062	
max_morpho_prod	2.71	1	0.0995	
MorphoencodedCh	0.03	1	0.8674	
MorphoMixedmorphemes	0.00	1	0.9808	
POS	4.35	3	0.2257	
max_tyt_k_ratio	1.09	1	0.2964	
Num_CWN_variant	0.97	1	0.3257	
bfcolloc	0.75	1	0.3863	
afcolloc	0.81	1	0.3681	
Num_sense	0.15	1	0.6952	
Num_CWN_antonym	0.45	1	0.5035	
Num_CWN_hyponym	0.05	1	0.8237	
Num_CWN_hyponym	0.02	1	0.8802	
Num_CWN_nearsyn	3.86	1	0.0493	
Num_CWN_syn	5.05	1	0.0247	
Conceptnum_relationtype	0.15	1	0.6981	
Conceptnum_relatedwordtype	2.37	1	0.1238	
actinboards_post	0.00	1	0.9950	
countcommentsChangSFStyleactPost	0.12	1	0.7270	
countcommentsChangSFStyleactComment	4.47	1	0.0346	
loan.word	26.23	1	<.0001	
diss_dissemination_value_post	0.33	1	0.5680	
TOTAL	38.33	24	0.0321	

Table 23 Parametric Statistic Wald test for Logistic Model of Conventionalized and Unconventionalized Words

```
Logistic Regression Model
lrm(formula = hypothesis ~ syllable_num + Num_CWN_nearsyn + Num_CWN_syn +
countcommentsChangSFStyleactComment + loan.word, data = CDtwords,
x = T, y = T)
```

		Model Likelihood		Discrimination		Rank Discrim.	
		Ratio Test		Indexes		Indexes	
Obs	198	LR chi2	106.95	R2	0.574	C	0.875
past_diffusion	128	d.f.	5	g	3.145	Dxy	0.751
diffusion	70	Pr(> chi2)	<0.0001	gr	23.222	gamma	0.780
max deriv	2e-04	gp		gp	0.353	tau-a	0.345
		Brier		Brier	0.119		
		Coef	S.E.	Wald Z	Pr(> Z)		
Intercept		-5.6458	1.1424	-4.94	<0.0001		
syllable_num		2.1668	0.5106	4.24	<0.0001		
Num_CWN_nearsyn		0.3192	0.1479	2.16	0.0309		
Num_CWN_syn		-0.1366	0.0622	-2.20	0.0280		
countcommentsChangSFStyleactComment=Y		1.7439	0.5078	3.43	0.0006		
loan.word=Y		3.2535	0.6421	5.07	<0.0001		

Table 24 Statistic Information for Logistic Model of Conventionalized and Unconventionalized Words



In logistic model, R^2 indicates how accurate the predictions of the model are. As shown in Table 24, the R^2 in current model is more than 0.5, so the accuracy is concurred. C is the index for concordance between predicted probability and observed response, if its value is above 0.8, then it may indicate the model has real predictive capacity. Dxy is a rank correlation between predicted probabilities and observed responses, which is 0.751. The values R^2 , C, and Dxy are high. They are values of gauging predictively of model, so the conclusion draw from this model may have its reliability. The bootstrap validation test also indicates the reasonable of current model. The fast backwards elimination algorithm reports that all predictors are retained. This indicates that though behaviorally with similar performance on Revised Constant U , these two sets of words are different in linguistic aspects.

On the other hand, it is also with interest to understand the important factors that can predict a words' future life in being conventionalized or not. It may be inappropriate to build a single model to all target words because the diversity of words included in present work. Words born before 1950 are those similar to what Wang and Minett (2004) called as "first emergent words." They are earlier coined for purposes different from recent diffused words. Words born after 1950 are characteristically similar to recent diffused words as shown in previous discussion. Thus, they can better shed lights in understanding the future life of present diffused words.

To look closer at words born after 1950 we can build a final multiple linear regression model incorporating all linguistic aspects by selecting those predictors with higher interpretative power in each separate linguistic perspective. We perform backward variable selection starting with main effects and interactions that are selected from each of above models. This multiple linear regression model for words after 1950

explains 82% of the variations with its adjusted r square as 0.7992. The residual plots are relatively appropriated in being correlated.

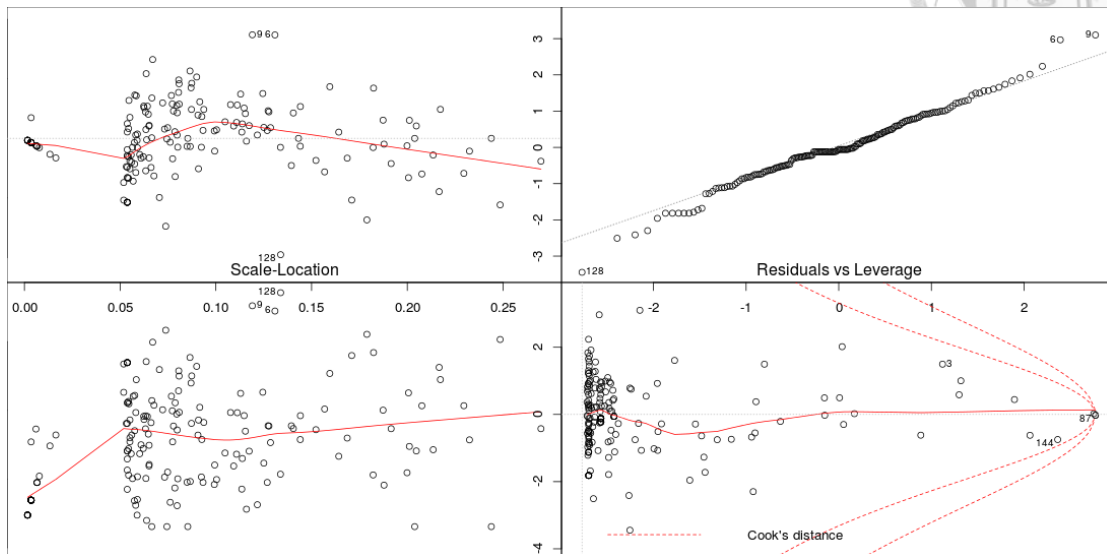


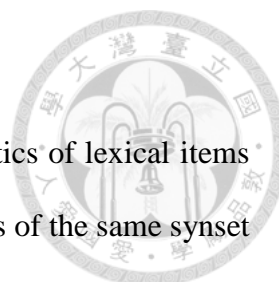
Figure 40 Residual Plots for Multiple Linear Regression model for Diffused Words

The detailed formula yielded from this multiple linear regression model is shown in Table 25 . Syntactic and pragmatic factors show significant influence.

Revised Constant $U = 1.631e-03 + 9.914e-03 \text{POSADV} + 1.874e-03 \text{POSVERB} + 1.917e-04$
 $\text{bfcolloc} + 1.387e-03 \text{afcolloc} + 3.085e-03 \text{Conceptnum_relationtype} - 1.353e-01$
 $\text{actinboards_post} + 8.380e-03 \text{countcommentsChangSFStyleactPostY} + 5.116e-02$
 $\text{countcommentsChangSFStyleactCommentY} + 5.022e-02 \text{diss_dissemination_value_post} - 9.504e-$
 $02 \text{Num_CWN_hypernym} - 4.969e-02 \text{Num_CWN_syn} + 5.534e-02$
 $\text{Num_CWN_totalrelation} - 1.995e-03 \text{POSADV:Pwithoutlw\$bfcolloc} + 2.659e-04$
 $\text{POSVERB:Pwithoutlw\$bfcolloc} + 2.081e-03 \text{POSADV:Pwithoutlw\$afcolloc} - 1.157e-03$
 $\text{POSVERB:Pwithoutlw\$afcolloc} - 5.368e-06 \text{bfcolloc:Pwithoutlw\$afcolloc} - 1.084e-02$
 $\text{Conceptnum_relationtype:Pwithoutlw\$countcommentsChangSFStyleactPostY} - 2.234e-03 \text{Num_}$
 $\text{CWN_syn:Pwithoutlw\$Num_CWN_totalrelation} + 4.064e-06 \text{POSADV:Pwithoutlw\$bfcolloc:P}$
 $\text{withoutlw\$afcolloc} + 4.730e-06 \text{POSVERB:Pwithoutlw\$bfcolloc:Pwithoutlw\$afcolloc}$

Table 25 Formula of Multiple Linear Regression Model of Words Born After 1950

In addition to sketching characteristics of words born after 1950, the features that can be used to decide its conventionalization are testified by building logistic model. Words after 1950 are classified into two sets. Words whose Revised Constant U is zero are considered to be not conventionalized, and those who are with Revised Constant U higher than zero are considered to be conventionalized. Among the total 180 lexical items, 37 are unconventionalized and 143 are conventionalized. Though this is a small data set, they are still randomly split into test data and train data. With stepwise back selection it shows that type number of co-occurring word before target word is singly good enough as predictor. The accuracy on test data is 0.7955. In order to ensure its real effect from syntagmatic relation words born before 1950 are used as conventionalized words and diffused words used as diffused words to testify the model. The accuracy is 0.6335.



4.4. Qualitative Analysis on Members of Synset

In addition to quantitative understanding in linguistic characteristics of lexical items how lexical items compete with each other is another issue. Members of the same synset are good target for understanding, for they share semantic representation and paradigmatic network, but they are different in situations of stabilization.

There are 15 members in the synset to express depression. Members from the same synset are behaving differently as shown in their cross month frequency in Figure 41.

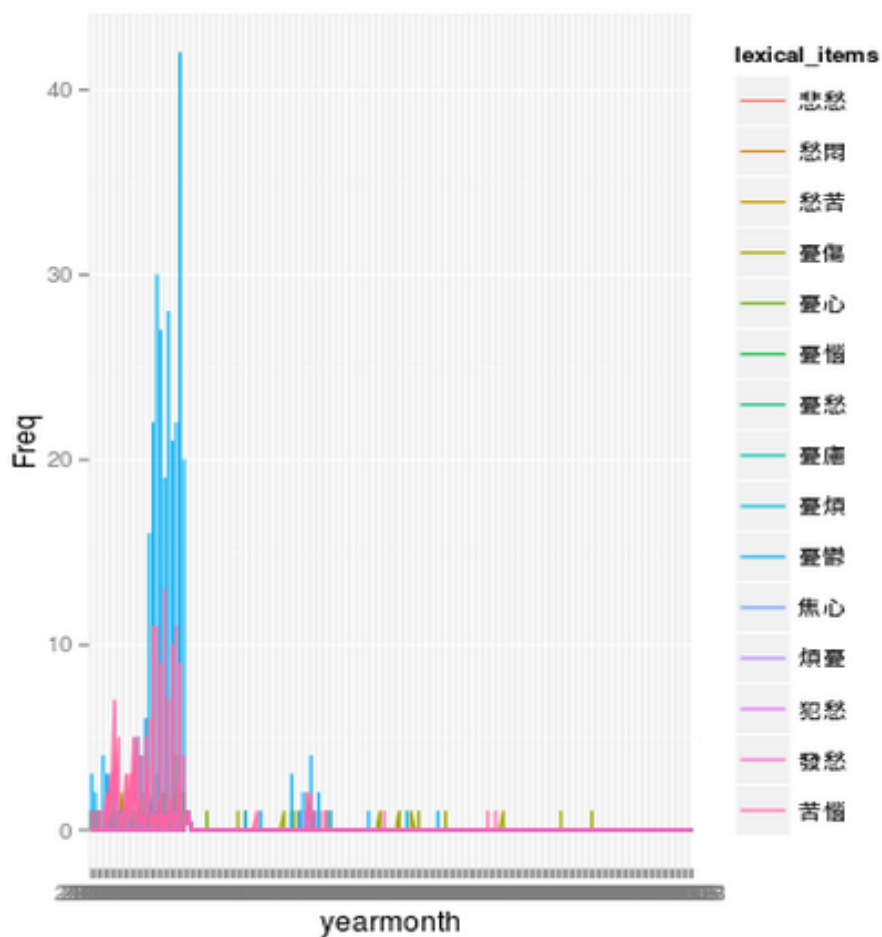


Figure 41 Cross Month Frequency of Synset Members

Their total frequency and their Revised Constant U are presented in Figure 42.

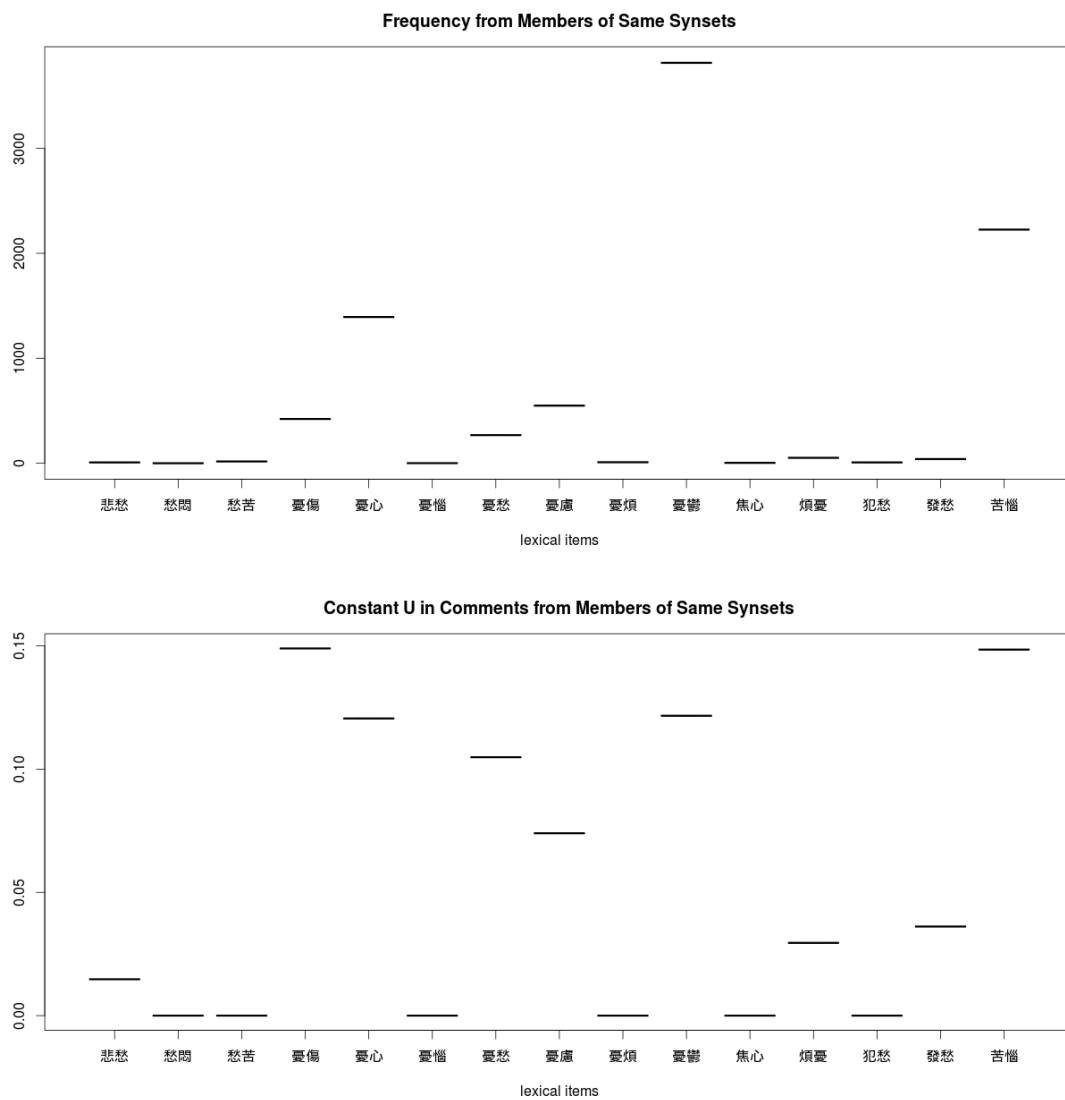


Figure 42 Distribution of Total frequency and Revised Constant U

These words are in paradigmatic relation for being as members of the same synset, so the syntagmatic view should be invited to understand how the words work differently in their co-occurring companies as well as pragmatically conceptual relation and related words that habitually linked in experiences. The words that are stably used include "悲愁," "煩憂," "發愁," "憂慮," "憂愁," "憂心," "憂鬱," "苦惱," and "憂傷."

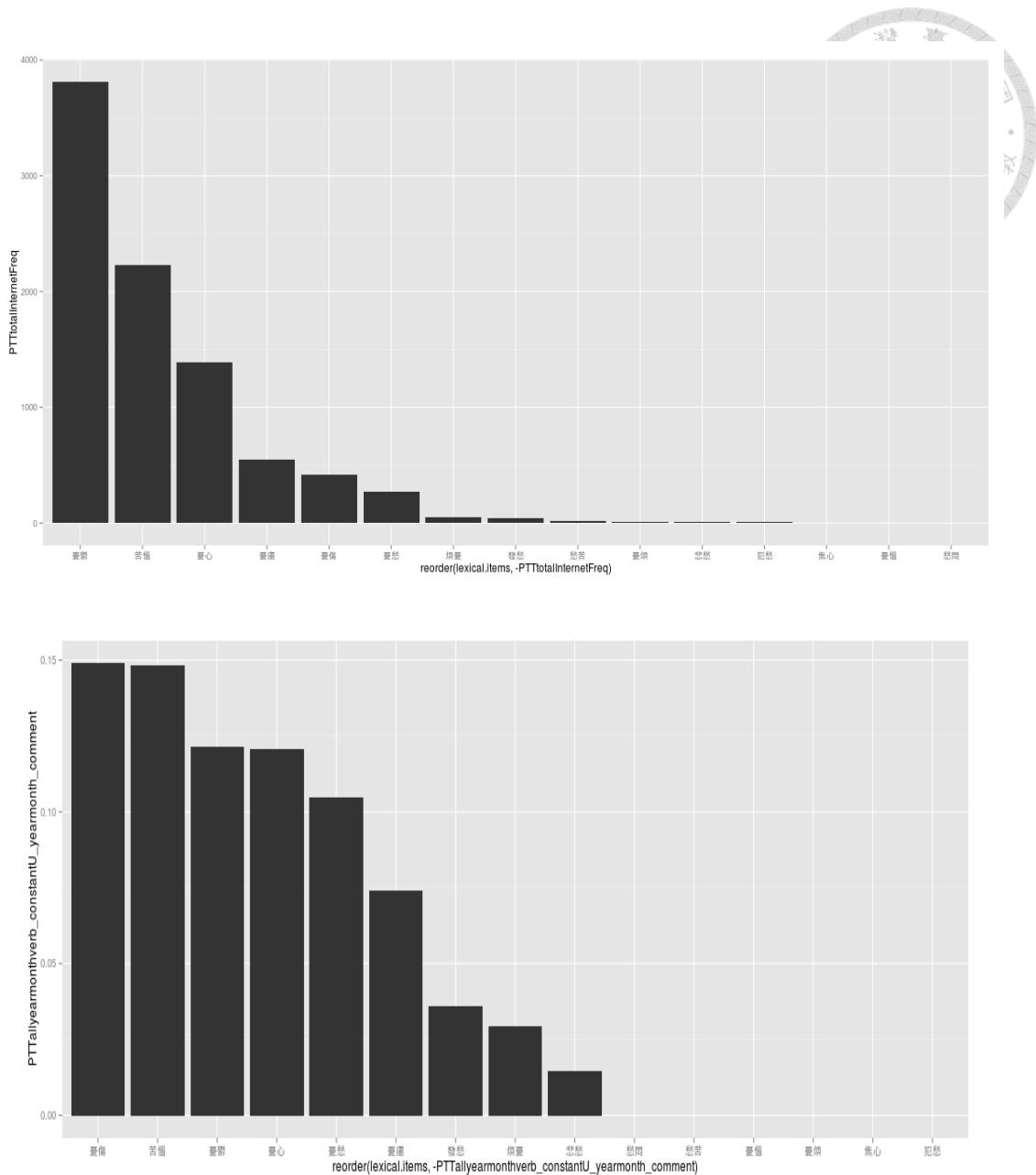


Figure 43 Synset Members Separately Ordered by Frequency (Upper panel) and Revised Constant U Value (Lower panel) Decreasingly

In Figure 44 and Figure 44 the words are ordered by Revised Constant U value decreasingly, and the y lab is presented with number of different types of co-occurring words. The plot shows that Revised Constant U and number of co-occurring accompany seem to be correlated.

number of related conceptual words is shown in Figure 46 with ordering by value of Revised Constant U decreasingly.

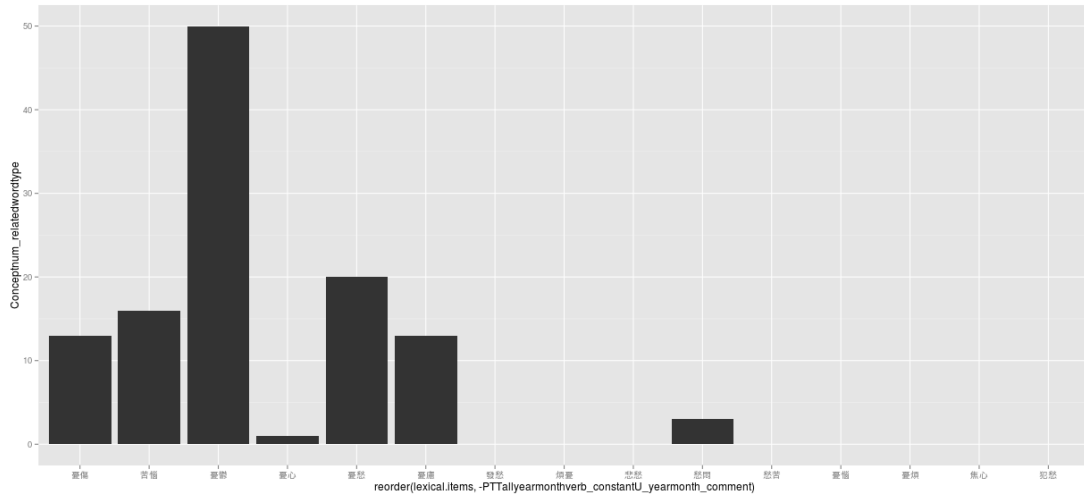


Figure 46 Number of Related Conceptual Words with Target Words Ordered by Revised Constant U Value Decreasingly

The number of conceptual relations is probed to understand conceptual contribution in standing out from other usages of the same sense.

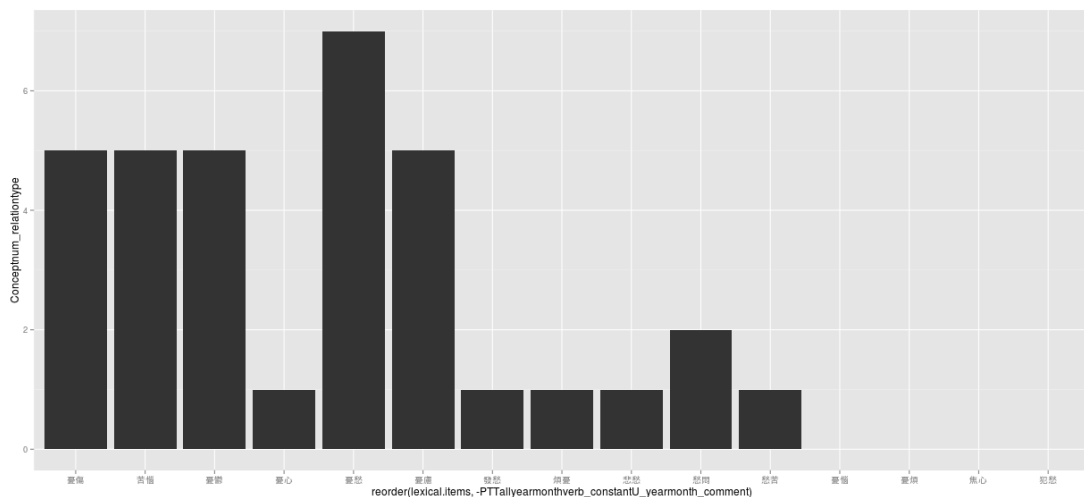
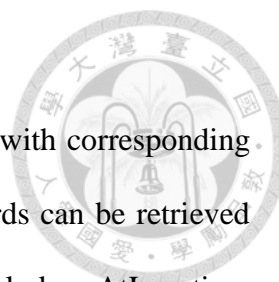


Figure 47 Number of Involved Conceptual Relations with Target Words Ordered by Revised Constant U Value Decreasingly



All of the concepts used in this sense are shown in Figure 48 with corresponding synset members that have lexicalized the concepts. Only seven words can be retrieved conceptual information from ConceptNet5. The concepts include: AtLocation, CapableOf, Causes, CausesDesire, Desires, HasProperty, HasSubevent, IsA, MotivatedByGoal, and SymbolOf.

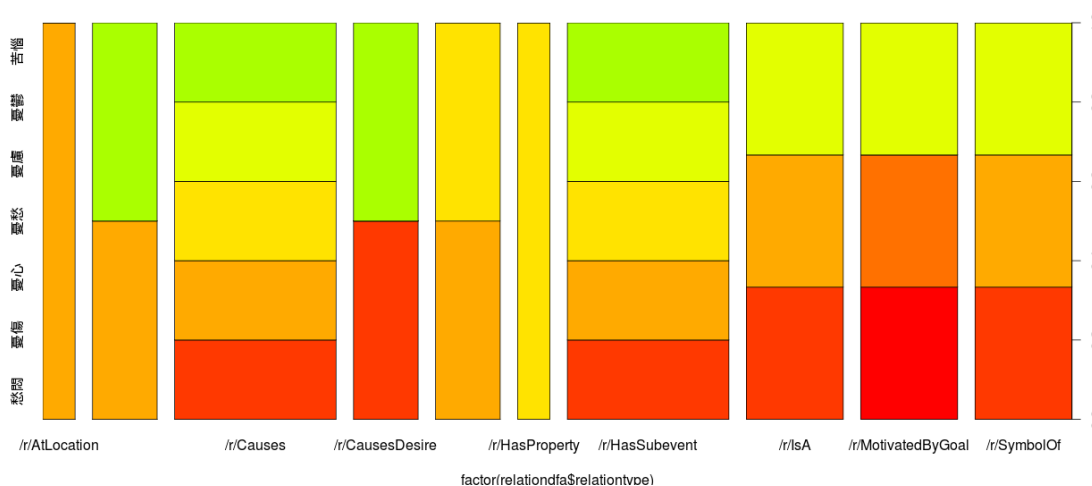


Figure 48 Distribution of Involved Conceptual Relation and Synset Members

We can move forward to their interaction with Revised Constant U in Figure 49 and

Table 26. It shows that words with high Revised Constant U may not be captured its conceptual relations on ConceptNet5, which may be a limitation on resources; however, for those reachable data we can find that except “憂心” the rest of those who are with high Constant U values are with rich conceptual relations.

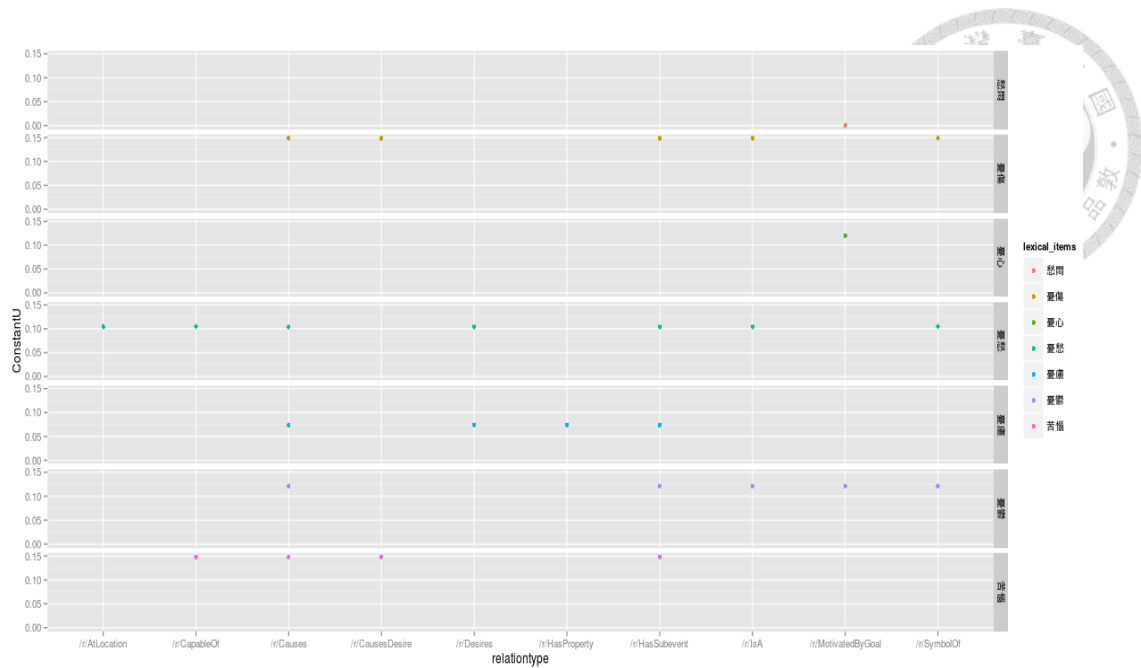
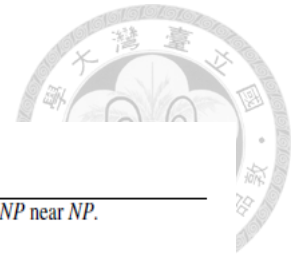


Figure 49 Synset Members and their Corresponding Involved Conceptual Relations

Lexical items	憂傷	苦惱	憂鬱	憂心	憂愁	憂慮	愁悶
Number of	5	4	5	1	7	4	1
Conceptual Relations							
Revised Constant U	0.1489	0.1484	0.1216	0.1205	0.1048	0.0740	0.0000

Table 26 Number of Conceptual Relations and Revised Constant U Value of Synset Members

The meaning each of conceptual relation stands for is shown in Table 27. In this synset there are 10 conceptual relations out of the total 21 provided conceptual relation.



Relation	Sentence pattern	Relation	Sentence pattern
IsA	<i>NP</i> is a kind of <i>NP</i> .	LocatedNear	You are likely to find <i>NP</i> near <i>NP</i> .
UsedFor	<i>NP</i> is used for <i>VP</i> .	DefinedAs	<i>NP</i> is defined as <i>NP</i> .
HasA	<i>NP</i> has <i>NP</i> .	SymbolOf	<i>NP</i> represents <i>NP</i> .
CapableOf	<i>NP</i> can <i>VP</i> .	ReceivesAction	<i>NP</i> can be <i>VP</i> .
Desires	<i>NP</i> wants to <i>VP</i> .	HasPrerequisite	<i>NP VP</i> requires <i>NP VP</i> .
CreatedBy	You make <i>NP</i> by <i>VP</i> .	MotivatedByGoal	You would <i>VP</i> because you want <i>VP</i> .
PartOf	<i>NP</i> is part of <i>NP</i> .	CausesDesire	<i>NP</i> would make you want to <i>VP</i> .
Causes	The effect of <i>VP</i> is <i>NP VP</i> .	MadeOf	<i>NP</i> is made of <i>NP</i> .
HasFirstSubevent	The first thing you do when you <i>VP</i> is <i>NP VP</i> .	HasSubevent	One of the things you do when you <i>VP</i> is <i>NP VP</i> .
AtLocation	Somewhere <i>NP</i> can be is <i>NP</i> .	HasLastSubevent	The last thing you do when you <i>VP</i> is <i>NP VP</i> .
HasProperty	<i>NP</i> is <i>AP</i> .		

Table 27 Interlingual relations in ConceptNet Adopted from Speer and Havasi

The most shared concepts involved are Causes and HasSubevent, which may imply that being able to involve in causation and several sub-events should play a role in being stably used because from , it shows that except “憂心” the rest of those who are with high Constant U values are all involved with these two conceptual relations.

The comparison in this section implies that words in the same synset, with same paradigmatic qualities, may be in the relationship of competition. The potential key to winning over the contest is the structural compatibility and involved conceptual relations.

4.5. Application: Inclusion of Lexical Items for Lexicology

Above findings we have discussed so far are further applied on proposing suggestions on inclusion of lexical items for lexicology. In addition to quantitative assumptions on constructing wordlist (Kessler, 2001), here we adopt linguistic consideration: pragmatically stable in use, syntactic compatibility, and semantically number of senses are taken as standard to expanding inclusion of words. The target words studied in present work has been compared with 8000 Chinese Words provided

by Steering Committee for the Test of Proficiency—Huayu⁶. Though the correlation between teaching level of words included in Huayu 8000 Chinese Words and their Revised Constant U is not highly correlated because the standards for assigning teaching level for each lexical item relies highly on frequency (Hunston, 2002; Tseng, 2013), the words all have Constant U values, namely, they are all stably used in contemporary as shown in Figure 50, in which Revised Constant U values are ordered by teaching levels from 5 to 1. Most of words are those who have existed before 1950, so current study would like to propose some updates on the wordlists in order to testify the proposed suggestions of inclusion from current study is appropriate or not. The updates are aimed to more than words before 1950 with the intention to supply that what is taught to the learner should synchronize with what is stably used in contemporary language speakers.

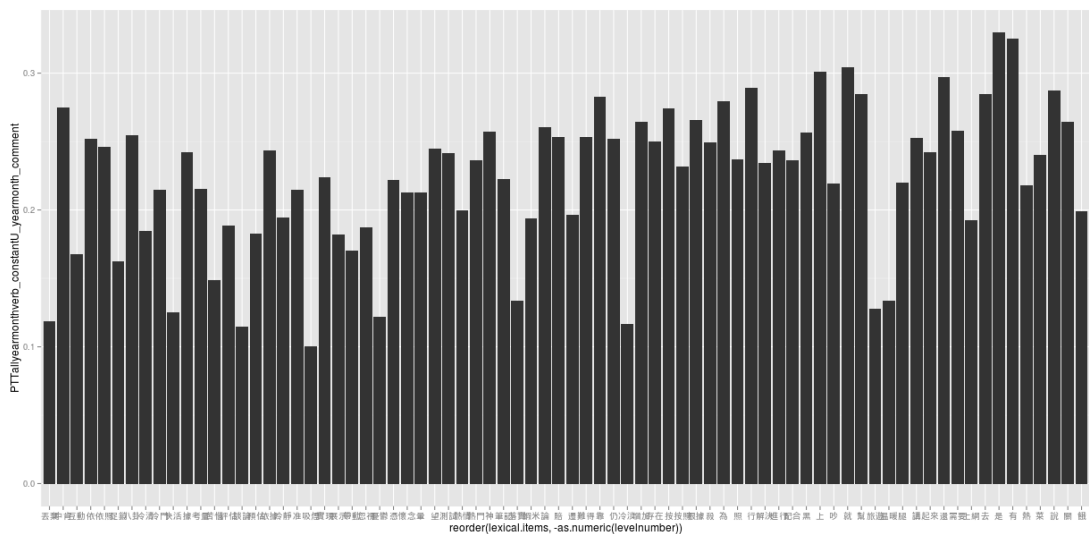
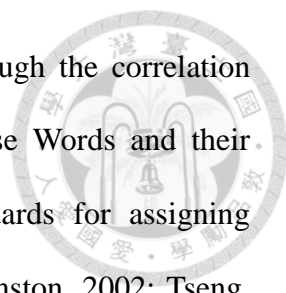


Figure 50 Revised Constant U Values of Target Words in 8000 Chinese Words

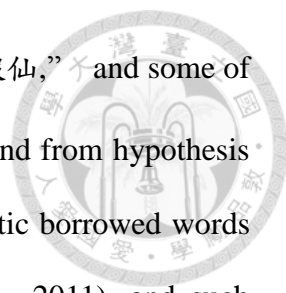
The stabilization is set as with Revised Constant U value more than 0.1005, which is

⁶ <http://www.sc-top.org.tw/english/download.php>

the minimum Revised Constant U value of words that have been included in wordlists. Besides, in order to avoid including lemma that is weighted in Revised Constant U values for its rich senses, word with number of senses more than 10 are excluded. In addition, given the fact that the significant role of syntagmatic relation shown in discussion of regression model building and qualitative analysis, type number of before target co-occurring accompanies and type number of after target co-occurring accompanies are also included as filtering features. The minimum value of type number of before target co-occurring accompanies and type number of after target co-occurring accompanies are 22 and 19, which is set as the filter value. There are 30 words from words before 1950 and 50 words from words after 1950 match above criteria. 50 words from words after 1950 are displayed in word cloud according to their total frequency.



Figure 51 Words Suggested to be Included in 8000 Chinese Words from Words after 1950



Most of these words are slang words, such as “暗爽,” “抓狂,” “假仙,” and some of them are borrowed from SouthernMin. From sociolinguistic angle and from hypothesis proposed by Kjellmer (2000) and of Metcalf (2002), the use of exotic borrowed words is with implication on establishing social identity (Altmann et al., 2011), and such exotic feature plays decisive role in whether the word is adopted or not. In addition to these social implications, current study supposes that the reason why these words are stably used is because the states they denoted are in daily human emotion experiences, but are not captured in a single lexical item in Chinese. Hence, their important function in signifying human cognition illustrates why they should be included.

Meanwhile, new inclusion from words before 1950 contains variants and synset of “吸煙,” which is included as level 4 in 8000 Chinese Words. Hong (2005) has probed the collocational limitations and distributional differences in variants at character level, but the activating differences at word level may have additional implications on human cognition. Variants of words share the same paradigmatic aspects, thus their syntagmatic perspective in co-occurring accompanies and involved conceptual relations as well as related conceptual words may provide insights.

From Figure 52 it shows that as a variant “抽菸” is much more stable than “抽煙,” and “抽菸” is much more stable in use than “吸煙” as in the same synset group. This is also reflected its number of conceptual relations and related conceptual words as shown in Figure 53.

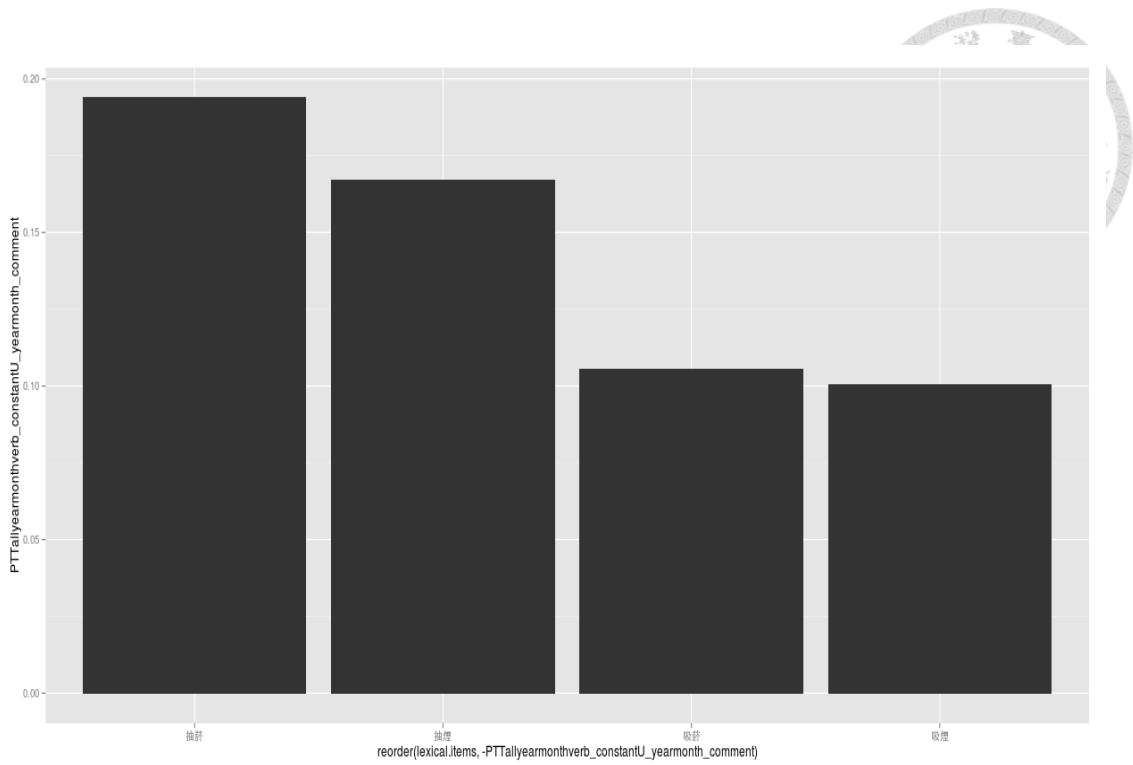


Figure 52 Variants and Synset of “吸煙” Ordered by Revised Constant U

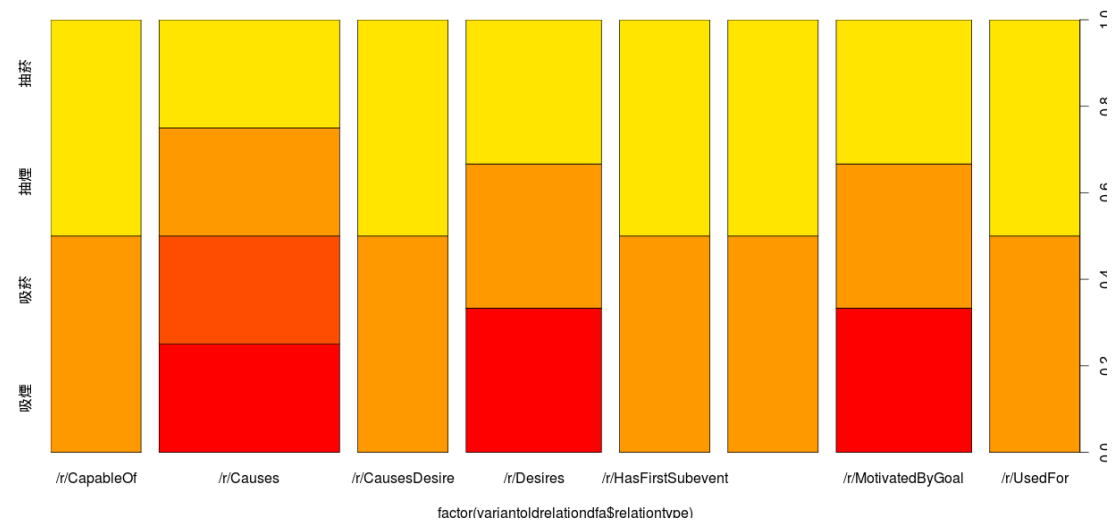
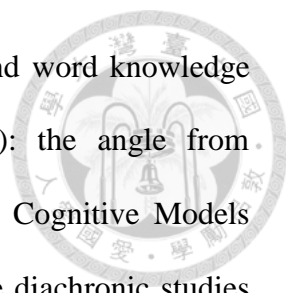


Figure 53 Involved Conceptual Relations for Variants and Synset of “吸煙”

On the other hand, with the criteria proposed words coming from the same conceptual embodiment can also be included. From the perspective of Cognitive



Semantics, to understand the relation between world knowledge and word knowledge can be approached by a variety of methods (Geeraerts, 2010): the angle from prototypically and salience (Rosch, 1973), the proposed Idealized Cognitive Models (Lakoff, 1987), the angle of Frame Semantics (Fillmore, 1985), the diachronic studies based on Invited Inferencing Theory of Semantic (Traugott and Dasher, 2005), or the approach from Conceptual Metaphor and Metonymy (Lakoff and Johnson, 1980). Among these angles, in the studies of the relation among meaning, concept, and embodiment, the topic about emotion language is very popular. Kövecses (2000) has introduced that emotion language can be classified into expressive or descriptive. In the descriptive emotion language, it can be further classified into literal language and figurative languages (Conceptual Metaphor and Conceptual Metonymy). When studying the issue of emotion, Lakoff and Kövecses (1987) have proposed the universal metonymic principle: the physiological effects of an emotion stand for the emotion. The physiological effect, temperature, is well discussed. For example, there is the operation of Anger is Heat metaphor in language (Lakoff and Kövecses, 1987; Yu, 1998). It is proposed that anxiety and fear are different from each other in the aspect of temperature from the angle of corpus linguistics (Ulrike, 2010), or anxiety is related to heat and fear to cold (Yu, 2002). When comparing with 8000 Chinese Words it is found that not all of lexicalized experiences of temperature are included, but with the proposed criteria in present study words from embodiment of temperature are all qualified to be included as shown in Figure 54.

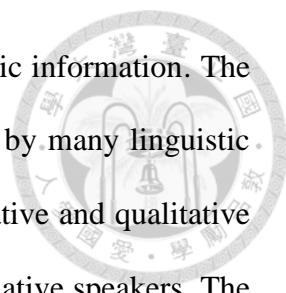
Chapter 5.

General discussion and conclusion



5.1. Conclusion

Previous studies have many insights in describing life of lexical items; however, there is rarely study providing both qualitative and quantitative perspectives in profiling life stages of words. The proposed life stages in present study are diffusion, conventionalization, and inactivation. They are not sequential stages, but may cycle from each other. The diffused ones may be just flash in the pan and swift into inactivation. The inactivated ones may be revived into being used. With adopting Revised Constant U the stably used ones can be clearly quantified from those inactivated ones. Manipulation on different types of target words gives opportunity to realize linguistic factors driven behind Revised Constant U and words coined in different temporal points. Pragmatics significantly accounts stabilization of words before 1950, but for words after 1950 the decisive factor is syntax. Though diffused words are highly stably used in PTT corpus, their underlying driven linguistic factors are different from words existing over centuries in five aspects: number of syllables, number of synonymic relations, number of near synonyms, whether it is actively used in content of comments, and whether it is from other language. Based on these findings appropriate prediction model for foretelling possible future life of currently diffused words is proposed with the aid of syntactic information. Additionally, with an aid from quantitative information qualitative understanding on potential competition within synset is probed to delineate potential picture in lexical competition. With these findings criteria for further extending Chinese lexicological resources are proposed with

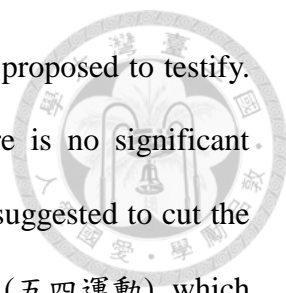


consideration on stability in use, syntactic compatibility, and semantic information. The Revised Constant U is a representative behavioral indicator driven by many linguistic factors, and syntagmatic factor, which is significant in both quantitative and qualitative analysis, plays key role in assisting including words stably used by native speakers. The update words are meaningful for they appropriately reflect variants that are used widely and lexical items that are conceptual or semantic related to words already listed in 8000 Chinese Words.

5.2. Implication and future study

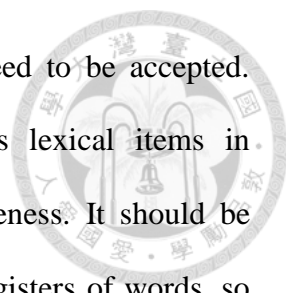
This thesis is intended to propose appropriate methodological design of understanding linguistic factors influencing conventionalization of lexical items as well as potential foretelling of diffused words. Though present study works hard to investigate related issues around conventionalization, there are still many directions can be further probed.

First, the intersection between language use and language comprehension is an important issue. Though spontaneous language performance can be retrieved in present report, it would be more comprehensive with inviting comprehension part to further understand mechanisms of lexicon. For example, survey on ironic or sarcastic tune of words, or other significant connotations words may bear. Secondly, the details of the found important factors should be further touched. The syntagmatic relations can be further probed by comparing different types of co-occurring words to different target words. For example, what are the types of co-occurring words that co-occur only with stabilized ones, or only with the non-stabilized ones? Besides, larger window size and information about POS in understanding syntagmatic relation of the target words can all be included. Such subtle differences may be abstracted to understand detailed features



deciding conventionalization of words. Third, more features can be proposed to testify. For example, from angles of phonology and of morphology there is no significant difference between words from 1950 and words after 1950, so it is suggested to cut the temporal line at 1900s, for it is the year for May Fourth Movement (五四運動), which influences people to move from writing in Classical Chinese/Literary Chinese (文言文) to writing in vernacular Chinese (白話文). Thus, from available resources we can reconsider the done investigation by changing temporal boundary of retrieved words. For example, to get words from 臺灣民報 in 1923. The discussion on phonological features can be further extended with consideration on syllable structure. Words with non-existed syllables in Mandarin Chinese should be included in observation. For example, "ㄅㄨㄛˊ" and "ㄅㄨ" are good examples of having no parallels in borrowing languages, though they may have Chinese characters standing for them as the way "ㄅ" stands for "ㄅㄨ". Or, to testify the activeness or Revised Constant U of the target words in different registers to further delineate the spreading directions of being conventionalized is also an important direction. The other example is to compare written variants in detail to understand driven cognitive reasons for choice of bearing word form in Chinese.

Besides, though competition among lexical items is generally revealed, it should be further probed with anchoring temporal information of the appearance of every synset member in order to understand how we incorporate new member and replace old ones within the same paradigmatic network. Related direction can be started from synonym blocking to understand appropriate range of synonymous member for a word to sustain in lexicon. Boulanger (1997) proposes that new words that have competing established form are more likely to succeed because the concept is established, but in




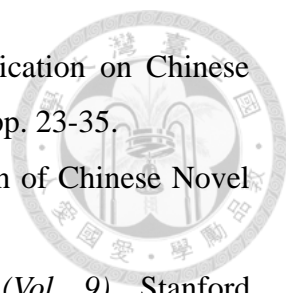
non-competition case both the new referent and the new form need to be accepted. Additionally, the testing wordlists should include more various lexical items in Huayu-speaking community in order to enhance its representativeness. It should be further classified with geographical considerations and retrieved registers of words, so the accuracy of inclusion can be further testified. We can even tailor wordlists to language speakers coming from different regions, and the stabilization of variants in different registers may be revealed.

Meanwhile, the diffused words may carve unique experiences differing from previous conceptual and semantic representation, which could be sustaining support for its being conventionalized, so follow-up observations are in need. Additionally, the characteristics of inactivated words and those reviving ones can be deeper qualitatively analyzed as reference to ensure reasons for being filtered out from operated lexicon. In addition to understanding of lexical items, different types of lexicon need to be discussed. The classification based on whether it is in language using or in language comprehension, or the discussion on age lexicon and gender lexicon may all shed lights on human cognition and advanced application.

References

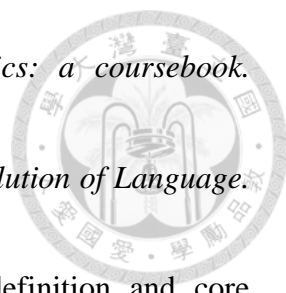
- Aitchison, & Lewis. (1996). *The mental word web: Forgetting the links*. Svartvik.
- Aitchison, J. (2001). *Language change: progress or decay?* Cambridge University Press.
- Aitchison, J. (2012). *Words in the mind : an introduction to the mental lexicon*. Chichester, West Sussex ; Malden, MA : Wiley-Blackwell.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, pp. 716-23.

- 
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*(19), pp. 716–23.
- Algeo, J. (1980). Where do all the new words come from. *American Speech*, 55(4), pp. 264-77.
- Altmann, E.G. , Zakary L. , & Whichard, Motter, A.E. (2013). Identifying Trends in Word Frequency Dynamics. *Journal of Statistical Physics*, 151(1-2), pp. 277-288.
- Altmann, E.G., Pierrehumbert, J.B., & Motter, A.E. (2011). Niche as a determinant of word fate in online groups. *PloS one*, 6(5).
- Baayen, R. H. (2009). Corpus linguistics in morphology: morphological productivity. In *Corpus linguistics. An international handbook* (pp. 900-19.).
- Barnhart. (2007). A calculus for new words. *Dictionaries*(28), pp. 132–138.
- Barnhart, C. (1978). American lexicography, 1945–1973. *American Speech*, 53(2), pp. 83-140.
- Bauer. (1983). *English Word-formation*. Cambridge University Press, Cambridge.
- Betz, W. (1949). *Deutsch und Lateinisch: Die Lehnbildungen der althochdeutschen Benediktinerregel*. Bonn: Bouvier.
- Boulanger, V. (1997). *What Makes a Coinage Successful?: The Factors Influencing the Adoption of English New Words*. University of Georgia.
- Brekle, H. (1978). Reflections on the conditions for the coining and understanding of nominal compounds. In U. Wolfgang , & W. Meid (Ed.), *Proceedings of the 12th International Congress of Linguists* (pp. 68-77). Innsbruck: Universitätsverlag Innsbruck.
- Brinton, L. (2002). Grammaticalization versus lexicalization reconsidered: on the 'late' use of temporal adverbs. In T. Fanego, L.-C. M.J. , & J. Pérez-Guerra (Ed.), *English historical syntax and morphology: selected papers from IIICEHL* (pp. 67-97). Amsterdam: Benjamins.
- Brinton, L., & Dieter , S. (1995). Functional renewal. In Andersen, *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIE* (pp. 33-47).
- Caramazza, A. (1997). How many levels of processing are there in lexical access? . *Cognitive Neuropsychology*(14), pp. 177–208.

- 
- Ceng, W.-X. (2013). Huayu baqian ci cihui fanji yanjiu [Classification on Chinese 8,000 Vocabulary]. *Teaching Chinese as Second Language*, pp. 23-35.
- Chang, P, & Ahrens, K. (2008). Towards a Model for the Prediction of Chinese Novel Verbs. *PACLIC* , (pp. 131-40).
- Chao, Y. (1976). *Aspects of Chinese sociolinguistics: essays (Vol. 9)*. Stanford University Press.
- Charles , W., & Miller, G. (1989). Contexts of antonymous adjectives. *Applied Psycholinguistics*(10), pp. 357-75.
- Chesley, P., & Baayen, R. H. (2010). Predicting new words from newer words: Lexical borrowings in French. *Linguistics*, 48(6), pp. 1343–74.
- Choueka, Y., Klein, S., & Neuwitz, E. (1983). Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal of the Association for Literary and Linguistic Computing*, 4.
- Church, K. W. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), pp. 22–29.
- Church, K., Gale, W. A., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In *Lexical Acquisition: Using On-line Resources to Build a Lexicon* (pp. 115–64). Lawrence Erlbaum.
- Committee, N. L. (1998). *Collection of Neologisms I*. Taipei: Ministry of Education.
- Cook., P. (2010). *Exploiting linguistic knowledge to infer properties of neologisms*. University of Toronto.
- Cruse. (1992). Antonymy revisited: Some thoughts on the relationship between words and concepts. In *Frames, Fields, and Contrasts* (pp. 289-306). Hillsdale, NJ: Lawrence Erlbaum associates .
- Cruse. (2001). The lexicon. In Mark Aronoff and Janie Rees-Miller. In *The Handbook of Linguistics* (pp. 238-64). Blackwell Publishers Inc., Malden, MA.
- Cruse. (2011). *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford:Oxford University Press.
- Donovan , R., & O’Neil, M. (2008). A systematic approach to the selection of neologisms for inclusion in a large monolingual dictionary. *Proceedings of the 13th Euralex International Congress* (pp. 571-579). Barcelona.
- Duckworth, D. (1977). *Zur terminologischen und systematischen Grundlage der*

- Forschung auf dem Gebiet der englisch-deutschen Interferenz.*
- Evans, V., & Melanie, G. (2006). *Cognitive Linguistics An Introduction*. Edinburgh:Edinburgh University Press.
- Evert, S. (2005). *The statistics of word cooccurrences: word pairs and collocations*.
- Farrar, S., & Langendoen, D. (2003). *A linguistic ontology for the semantic web*. Glot International.
- Fellbaum, C. (2014). Large-Scale Lexicography in the Digital Age. *International Journal of Lexicography*.
- Fellbaum, L. (1995). *Morphological Aspects of Language Processing*. Hove:Lawrence Erlbaum.
- Fenk-Oczlon. (1989). Word frequency and word order in freezes. *Linguistics*(27), pp. 517-56.
- Fernández-Domínguez, J. (2010). Productivity vs. Lexicalization: Frequency-Based Hypotheses on Word-Formation. *Poznań Studies in Contemporary Linguistics*, 46(2), pp. 193-219.
- Firth, J. (1957). A synopsis of linguistic theory 1930–55. In *In Studies in linguistic*. The Philological Society, Oxford.
- Fischer. (1998). *Lexical Change in Present Day English: A Corpus-Based Study*. Gunter Narr Verlag, Tübingen, Germany.
- Fischer, O., Rosenbach, A., & Stein, D. (2000). *Pathways of change: grammaticalization in English (Vol. 53)*. John Benjamins Publishing.
- Fromkin. (1971). The non-anomalous nature of anomalous utterances. *Language*(47), pp. 27-52.
- Geeraerts, D. (2010). *Theories of Lexical Semantics*. Oxford: Oxford University.
- Gibbs, & Gonzales. (1985). Syntactic frozenness in processing and remembering idioms. *Cognition*, pp. 243-59.
- Giuliano, V. (1965). The interpretation of word associations. In M. Stevens, V. Giuliano, & L. Heilprin (Ed.), *Proceedings of the Symposium on Statistical Association Methods For Mechanized Documentation*. 269, pp. 25-32. Washington,DC: National Bureau of Standards Miscellaneous Publication.
- Greenberg, Joseph H. (1991). The last stages of grammatical elements: Contractive and xpansive desemanticization. In Heine, & Traugott, *Approaches to*

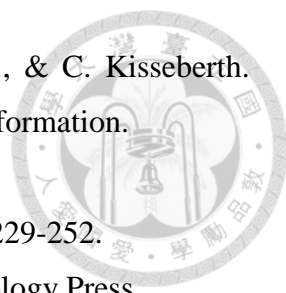
- grammaticalization* (pp. 301–314).
- Gross, D., & Miller, K. (1989). Adjectives in WordNet. *International Journal of Lexicography*(3), pp. 265-77.
- Halliday, M., & Webster, J. (2007). *Language and society (Vol. 10)*. Bloomsbury Publishing.
- Hargraves, O. (2007). Taming the wild beast. *Dictionaries*(28), pp. 139-141.
- Harley, T. (2005). *The Psychology of Language*. New York: Psychology Press.
- Harris, A., & Lyle, C. (1995). *Historical Syntax in Cross-Linguistics Perspective*. Cambridge, UK: Cambridge University Press.
- Haugen, E. (1950). The analysis of linguistic borrowing. *Language*(26), pp. 210-31.
- Hay, J., & Baayen, H. (2002). Parsing and productivity. In G. Booij, & J. van Marle, *Yearbook of morphology* (pp. 203-35). Dordrecht/Boston/London: Kluwer.
- Heine, B. (1997). *Cognitive Foundations of Grammar*. Oxford: Oxford University Press.
- Heine, B. (2003). *(De)grammaticalization*. Kate Burridge and Barry Blake.
- Heine, B., Claudi, U., & Hünnemeyer, F. (1991). *Grammaticalization: A Conceptualframework*. Chicago: The University of Chicago.
- Hickey, R. (2003). *Motives for language change*. Cambridge University Press.
- Hohenhaus, P. (2005). Lexicalization and institutionalization. In *Handbook of word-formation* (pp. 353-73). Springer Netherlands.
- Hong, J.-f., Wu, Y., & Huang, C.-R. (2005). yitizi yu yiti ci cihui yuyi chutan [Probe on Variants in Chinese Characters and Lexical Items]. *CLSW2005*.
- Hsu, F. h. (1999). *Taiwan dangdai guoyu xin ci tan wei [Exploring the Contemporary Mandarin New Words in Taiwan]*. National Taiwan Normal University.
- Hsu, F.-h. (2006). *Taiwan dangdai guoyu xin ci tan wei[Probe on Neologisms of Taiwan Modern Chinese]*.
- Huang, C.-R. (2005). Hanzi zhishi bi oda de ji ge cengmian: Zi, ci, yu ciyi guanxi gailun[Ontology of Chinese Characters in Several Perspectives: Characters, Lexical Items, and Semantic Relations]. *International Conference on Chinese Character and Globalization*. Taipei.
- Hudson, R. (1996). *Sociolinguistics*. Cambridge University Press.
- Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

- 
- Hurford, J. R., Heasley, B., & Smith, M. B. (2007). *Semantics: a coursebook*. Cambridge University Press.
- Hurford, J., Michael, S.-K., & Chris. (1998). *Approaches to the Evolution of Language*. Cambridge, UK: Cambridge University Press.
- Jarema, G., & Libben, G. . (2007). Introduction: Matters of definition and core perspectives. In G. Jarema , & G. Libben, *The Mental Lexicon: Core Perspectives* (pp. 1-7). Oxford: ELSEVIER.
- Jenkins. (1970). The 1952 Mnesota word association norms. In Postman, & Keppel, *Norms of word association* (pp. 1-38). Academic Press New York.
- Johnson-Laird. (1983). *Mental Models*. Cambridge:Cambridge University Press.
- Jucker, A. H., & Taavitsainen, I. (Eds.). (2010). *Historical pragmatics* . Walter de Gruyter.
- Keller, R. (1994). *On language change: The invisible hand in language*. Psychology Press.
- Kerremans, D. (2015). A Web of New Words: A Corpus-based Study of the Conventionalization Process of English Neologisms.
- Kerremans, D., Stegmayr, S., & Schmid, H. J. . (2011). The NeoCrawler: identifying and retrieving neologisms from the internet and monitoring ongoing change. *Current Methods in Historical Semantics*(73), p. 59.
- Kessler, B. (2001). *The significance of word lists*. Center for the Study of Language and Inf.
- Kim, H. (2007). *Xiandai hanyu xin ci yanjiu [Sutdy on Neologisms of Modern Chinese]*.
- Kjellmer, G. (2000). Potential Words. *Word*(51), pp. 205-28.
- Klein, D. E. (2001). The representation of polysemous words. *Journal of Memory and Language*(45), pp. 259–82.
- Kövecses, Z. (2000). *Metaphor and Emotion*. Cambridge: Cambridge University Press.
- Kreidler. (1998). *Introducing English Semantics*. London: Routledge.
- L’Homme, M. C. (2014). Why Lexical Semantics is Important for E-Lexicography and Why it is Equally Important to Hide its Formal Representations from Users of Dictionaries. *International Journal of Lexicography*, 27(4), pp. 360-77.
- Lakoff, G., & Kövecses, Z. (1987). The cognitive model of anger inherent in American English. In D. Holland, & N. Quinn, *Cultural Models in Language and Thought*

- (pp. 195–221). Cambridge: Cambridge University Press.
- Lakoff, George, & Mark Johnson . (1980). *Metaphors We Live By*. Chicago:University of Chicago Press.
- Langacker, R. (1987). *Foundations of Cognitive Grammar, Vol. 1: Theoretical Prerequisites*. Stanford, CA: Stanford UP.
- Lass, R. (1990). How to do things with junk: exaptation in language evolution. *Journal of Linguistics*, 26, pp. 79-102.
- Lass, R. (1997). *Historical Linguistics and Language Change*. Cambridge, UK: Cambridge University Press.
- Leech, G. (1981). *Semantics* (2nd ed.). Harmondsworth: Penguin.
- Lehmann, C. (1995[1982]). *Thoughts on Grammaticalization (originally published as Thoughts on Grammaticalization: A Programmatic Sketch, Vol. 1. University of Cologne: Arbeiten des Kölner Universalienprojekts 49)*. Munich: LINCOM EUROPA.
- Lehr, A. (1996). Kollokationen und maschinenlesbare Korpora. *volume 168 of Germanistische Linguistik*. Niemeyer, Tübingen.
- Lehrer. (2003). Understanding trendy neologisms. *Italian Journal of Linguistics*, 15(2), pp. 369–382.
- Lipka, L. (1977). Lexikalisierung, Idiomatisierung und Hypostasierung als Probleme einer synchronen Wortbildungslehre. In K. Dieter , & E. Herbert , *Perspektiven der Wortbildungsforschung Beitrage zum Wuppertaler Wortbildungskolloquium vom 9. – 10.* (pp. 155-64). Brekle. Bonn: Bouvier.
- Liu, T.-j. (2014). *PTT Corpus: Construction and Applications*.
- Love, N. (2006). *Language and history: Integrationist perspectives* . Routledge.
- Lyons. (1981). *Language, Meaning, and Context*. London:Fontana.
- Masini, F., & Huang, h.-q. (1997). *Xiandai hanyu cihui de xingcheng: Shijiu shiji hanyu wailai ci yanjiu [The Formation of Modern Chinese Vocabulary: Loan Words in the Nineteenth Century]*. Foreign Chinese Dictionary.
- Metcalf, A. (2002). *Success, Predicting New Words: the Mystery of Their Success*. Boston New York:Houghton Mifflin Company.
- Metcalf, A. (2007). The enigma of 9/11. *Dictionaries*(28), pp. 160–162.
- Milroy, L., & Gordon, M. (2008). *Sociolinguistics: Method and interpretation (Vol. 13)*.

- John Wiley & Sons.
- Moon, R. (2013). Braving Synonymy: From Data to Dictionary. *International Journal of Lexicography*, 26(3), pp. 260-278.
- Murphy, G., & Andrew, J. (1993). The conceptual basis of antonymy and synonymy in adjectives. *Journal of Memory and Language*(32), pp. 301-319.
- Murphy, M. (2013). What We Talk about When We Talk about Synonyms (And What it can tell us About Thesauruses). *International Journal of Lexicography*.
- Norde, M. (2002). The final stages of grammaticalization: Affixhood and beyond. In W. Diewald, *Typological Studies in Language* (pp. 45-81).
- Plag, I. (2006). Productivity. In *The handbook of English linguistics* (pp. 537-56.).
- Polguere, A. (2014). From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*.
- Renouf, A. (2013). A finer definition of neology in English: the life-cycle of a word. . *Corpus perspectives on patterns of lexis*, pp. 177-207.
- Rey, A. (1995). The Concept of neologism and the evolution of terminologies in individual languages. In *In Essays on Terminology*. Amsterdam: John Benjamins.
- Romagnoli, C. (2013). The Lexicographic Approach to Modern Chinese Synonyms. *International Journal of Lexicography*, 26(4), pp. 407-23.
- Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore, *Cognitive Development and the Acquisition of Language* (pp. 111–44). New York: Academic Press.
- Sabino, R. (2005). Survey Says . . . Gameday. *American Speech*(80), pp. 61–77.
- Schmid, H. J. (2005). *Englische Morphologie und Wortbildung: Eine Einführung*.
- Schmid, H.-J. (2008). New words in the mind: Concept-formation and entrenchment of neologisms. In *Anglia-Zeitschrift für englische Philologie* (pp. 1-36).
- Sheidlower. (1995). Principles for the inclusion of new words in college dictionaries. *Dictionaries*(16), pp. 33–44.
- SinclairJohn. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Speer, R., & Havasi, C. . (2012). Representing General Relational Knowledge in ConceptNet 5. *LREC* , (pp. 3679-3686).

- Starreveld, P. A. (2004). Phonological facilitation of grammatical gender retrieval. *Language and Cognitive Processes* (6), pp. 677–711.
- Tang, T.-C. (1989). *Hanyu ci fa jufa xuji* [Chinese Semantic and Syntax].
- Thomason, S., & Kaufman, T. (1988). *Language contact, creolization, and genetic*. Berkeley: University of California Press.
- Traugot, E. (2004). Exaptation and grammaticalization. In *Linguistic Studies Based on Corpora* (pp. 133-156). Tokyo: Hituzi Syobo Publishing Co.
- Traugott, E., & Bernd, H. (1991). *Approaches to Grammaticalization*. Amsterdam: Benjamins.
- Traugott, Elizabeth C, & Richard B. Dasher . (2005). *Regularity in Semantic Change*. Cambridge: Cambridge University Press.
- Ulrike, O. (2010). Using Corpus Methodology for Semantic and Pragmatic Analysis: What Can Corpora tell Us About the Linguistic Expression of Emotions? *Cognitive Linguistics*, 21(4), pp. 727–63.
- Vincent, N. (1995). Exaptation and grammaticalization. In Andersen, *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES* (pp. 433-448).
- Wang, Z.-m. (2010). Jiyu shijian kuadu de hanyu jiaoxue changyong ci biao tongji yanjiu [Statistic Information for Chinese Teaching Wordlists based on Temporal Information]. *huawen jiaoxue yu yanjiu*(4), pp. 49-55.
- Wang, W. S., Ke, J., & Minett, J. W. (2004). Computational studies of language evolution. In *Monograph Series B* (pp. 65-108).
- Warren, P. (2012). *Introducing Psycholinguistics*. Cambridge: Cambridge University Press.
- Weinreich, U. (1953). *Languages in contact: findings and problems*. The Hague: Mouton.
- Williams, G. (2003). Les collocations et l'école contextualiste britannique. In F. Grossmann, & A. Tutin, *Les Collocations: analyse et traitement* (pp. 33-44). Amsterdam: De Werelt.
- Wischer, I., & Diewald, G. (2002). *New reflections on grammaticalization (Vol. 49)*. John Benjamins Publishing.
- Yip, M. (1980). *The tonal phonology of Chinese*. Ph D Dissertation, MIT. Published.

- 
- Yip, M. (1994). Isolated Uses of Prosodic Categories. In J. Cole , & C. Kisseberth.
Stanford,California: Center for the Study of Language and Information.
- Yip, M. (2003). What Phonology has Learnt from Chinese.
- Yip, M. (2007). Tone. *The Cambridge Handbook of Phonology*, pp. 229-252.
- Yip, P. (2000). *The Chinese lexicon: a comprehensive survey*. Psychology Press.
- Yu, N. (1998). *The Contemporary Theory of Metaphor: A Perspective from Chinese*.
Amsterdam: Benjamins.
- Yu, N. (2002). Body and Emotion :Body Parts in Chinese Expression of Emotion.
Pragmatics & Cognition(10), pp. 341–67.
- Zhu, J.-n. (2000). Taiwan xiaoyuan xin ci de fazhan he dua jiaoxue de yingxiong[Influence of Neologisms from Taiwan Campus on Teaching]. *6th Global Chinese Teaching Conference*. Taipei.

Appendices



Appendix 1 Predictors Used in Current Study and Their Correspondence to Previous Models

Proposed features in current study	Features in Kjellmer (2000)	Features in Metcalf (2002): FUDGE	Features in Chang(2008)	Features in Kerremans (2015)	Delete
<u>Phonology</u>					
number of syllable	Ph1. It has phonological parallels in the language. Ph2. It is easy to pronounce. G2. Its spelling agrees with its	Unobstrusiveness			Delete: Not meet features in Chinese



	pronunciation			
<u>Morphology</u>	M3. Its derivative	“Generating new	Productive Affixes	Source of Data:
Component	affix is highly	forms (level 2)”of		Google Book N-gram
Richness of the	productive.	“Generation of Other		Corpus (GBNC)
monosyllabic verb		Forms and		
or of the elements in		Meanings”		
the dissyllabic verb				
constructions				
<u>Morphology</u>	G1. It has			
Number of	graphematic parallels			
graphematic	in the language.			
variation				
<u>Morphology</u>	M1. It has			
be encoded by	morphological			
Chinese character	parallels in the			
or not	language.			
	M2. It follows			
	morphological			
	principles.			
<u>Morphology</u>	M4. Its derivative		words should not	



Mixed originated morphemes or not affix is compatible with the stem. have morphemes of mixed origins

Syntax

Co-occurrence

“Generation of Other Forms and Meanings”

Productivity words having more than ten collocates would be scored as two; those with less than ten collocates but having more than three Word Sketch functions would be considered moderately productive and scored as one; and those with less than ten collocates and having less than or equal to three Word Sketch

H6: The early development of syntagmatic lexical networks, represented by collocations in the present study, promotes conventionalization.

functions would be scored as zero

Syntax

Parts of Speech

Semantics

Number of senses

“Variety of meanings(level 2) ”of “Generation of Other Forms and Meanings”(Metcalf 2002)

Semantics

Number of synonym

S1. It has semantic parallels in the language.
O2. It is concise

Semantic Gaps

if there are no competing synonyms, then we consider the word filling up a semantic gap.

Semantics

Number of near synonym

S1. It has semantic parallels in the language.
O2. It is concise

Semantic Gaps

if there are no competing synonyms, then we consider the word filling up a semantic gap.



Semantics

Number of antonym

Semantics

Number of holonym

Semantics

Number of

hyponym

S2. It is transparent to the layman.

Unobtrusiveness

Transparency:
we adopt identical operational definitions as in Metcalf's model, i.e., the meanings of transparent words should not be specialized and must be clearly inferable from the form.

H1: Semantic ambiguity

Delete:

This can be reflected in the dissemination across language users and number of senses for to investigate the meaning is morphological or metonymic originated is not so meaningful because based on frequency effect as well as studies on



mental lexicon once the sense of the form has been highly activated, then it becomes automation, so there is not significant activating differences in reaction time as in those cases of entrenched metaphors.

Sociolinguistics
loan words or not

O1. It has prestigious and/or exotic connotations. Unobtrusiveness

Sociolinguistics
Dissemination
across users

Number of User IDs/total frequency Diversity(variety of users and situations)

Posts is assumed to be relatively more stringent in word use

Pragmatics
Number of Involved



Conceptual Relation

Type

(ConceptNet)

Pragmatics

Number of Related

O3. It has humorous

Unobtrusiveness

Concept Words

connotations

(ConceptNet)

Frequency of Use

Frequency:
normalized ratio in
the year 1996 in order
to simulate the
prediction process

H5: The
nameworthiness of
the represented
concept or its
salience in society
promotes
conventionalization.

This aspect has been
reflected in activeness
over themes writing
style and
dissemination.

Pragmatics

Activeness in

Different Writing

Styles:

Diversity(variety of
users and situations)

Total frequency and



Slope in PTT Posts

**(Excluding
Gossiping for its
inclusiveness in
various topics)**

**Total frequency and
Slope in PTT
comments**

**(Excluding
Gossiping for its
inclusiveness in
various topics)**

Pragmatics

Activeness in

Different Themes:

**Number of
Activation Themes**

**(Total frequency
and Slope in**

Diversity(variety of
users and situations)

Posts take the lead in
directing themes, so
the information
retrieved from posts

**different theme
boards (posts))
(Including
Gossiping for its
inclusiveness in
various topics)**



Endurance of the
Concept

This has been
reflected in Constant
U

Appendix 2 Brief Summarization on Boards Used in Current Study



Board Name	Theme	Number of Posts	Tokens in Posts	Number of Comments (PBA order ⁷)	Tokens in Comments (PBA order)
LoL	Games	36,752	9,327,450	Push	Push
				489,870	827,382
				Boo	Boo
				142,323	229,608
				Arrow	Arrow
				279,887	527,352
				Total:	Total:
				912,080	1,584,342
ToS	Games	36,834	10,778,419	Push	Push
				659,366	654,276
				Boo	Boo
				153,930	121,578
				Arrow	Arrow
				341,425	518,388

⁷ PBA order means that the order of the frequency is listed as “push,””boo,””arrow.”



				Total: 1,154,721	Total: 1,294,242
PuzzleDragon	Games	17,030	4,205,373	Push 335,158 Boo 32,526 Arrow 275,892 Total: 643,576	Push 334,281 Boo 22,241 Arrow 273,699 Total: 630,221
MentTalk	Gender	29,236	9,447,661	Push 146,900 Boo 22,215 Arrow 210,405 Total: 379,520	Push 282,936 Boo 31,321 Arrow 411,736 Total: 725,993
WomenTalk	Gender	77,655	28,142,648	Push 564,540	Push 1,065,755



				Boo	Boo
				64,134	59,860
				Arrow	Arrow
				448,560	868,396
				Total:	Total:
				1,077,234	1,994,011
Boy_Girl	Gender	38,252	22,219,120	Push	Push
				93,226	181,758
				Boo	Boo
				43,587	38,900
				Arrow	Arrow
				119,895	206,862
				Total:	Total:
				256,708	427,520
Hate	Mood	162,774	26,988,668	Push	Push
				412,456	102,972
				Boo	Boo
				357,376	9,143
				Arrow	Arrow
				459,336	171,194



				Total: 1,229,168	Total: 283,309
happy	Mood	20,727	2,318,504	Push 4,422 Boo 20,395 Arrow 9,104 Total 33,921	Push 1,302 Boo 38 Arrow 741 Total 2,081
Sad	Mood	22,099	4,251,144	Push 10,611 Boo 11,310 Arrow 14,209 Total: 36,130	Push 2,937 Boo 36 Arrow 2,493 Total: 5,466
NBA	Sport	34,363	14,993,792	Push 10,948	Push 18,372



				Boo	Boo
				21,008	1,546
				Arrow	Arrow
				6,029	8,234
				Total:	Total:
				37,985	28,152
Baseball	Sport	41,286	11,028,551	Push	Push
				424,879	848,940
				Boo	Boo
				81,045	159,917
				Arrow	Arrow
				278,456	546,850
				Total:	Total:
				784,380	1,555,707
movie	Lifestyle	43,112	21,651,766	Push	Push
				202,636	175,607
				Boo	Boo
				118,592	17,700
				Arrow	Arrow
				171,649	120,230



				Total: 492,877	Total: 313,537
Food	Lifestyle	72,529	13,908,516	Push 34,547 Boo 69,715 Arrow 37,085 Total: 141347	Push 21,470 Boo 12 Arrow 11,731 Total: 33213
BuyTogether	Business	37,901	6,261,136	Push 443,000 Boo 11,748 Arrow 56,915 Total: 511,663	Push 852,412 Boo 2,075 Arrow 94,443 Total: 948,930
home_sale	Business	26,477	10,509,415	Push 73,810	Push 149,668



				Boo	Boo
				18,668	10,840
				Arrow	Arrow
				146,516	273,025
				Total:	Total:
				238,994	433,533
Stock	Business	23,159	9,406,420	Push	Push
				128,387	241,467
				Boo	Boo
				21,755	28,508
				Arrow	Arrow
				115,819	238,497
				Total	Total
				265,961	508,472
StupidClown	Story	44,547	14,330,189	ush	Push
				141,313	280,920
				Boo	Boo
				28,375	7,209
				Arrow	Arrow
				61,934	77,664



				=231619	=365793
joke	Story	44,282	5,957,892	Push	Push
				101,042	164,054
				Boo	Boo
				33,816	33,298
				Arrow	Arrow
				37,568	50,995
				Total:	Total:
				172,426	248,347
ask	ask	49,479	5,591,480	Push	Push
				42,344	77,126
				Boo	Boo
				16,561	2,820
				Arrow	Arrow
				91,099	167,816
				Total:	Total:
				150,004	247,762



Kaohsiung	Geography	54,879	10,467,668	Push	Push
				101,936	152,073
				Boo	Boo
				13,543	12,906
				Arrow	Arrow
				80,414	129,088
Total:	Total:				
				95,893	294,067
Keelung	Geography	21,470	3,837,887	Push	Push
				18,449	18,241
				Boo	Boo
				21,547	1,003
				Arrow	Arrow
				21,195	16,088
Total:	Total:				
				61,191	35,332
TaichungCont	Geography	15,617	2,983,035	Push	Push
				8,981	7,841
				Boo	Boo
				5,902	122



			Arrow	Arrow
			11,630	7,254
			Total:	Total:
			26,513	15,217
Gossiping	Gossiping	552,747	126,421,529	

Appendix 3 Constant U value for Lexical Items Before 1950



	lexical.items	_constantU_yearmonth_comment
1	是	0.329554
2	有	0.3252
3	就	0.304488
4	上	0.300824
5	還	0.296838
6	行	0.289446
7	說	0.287399
8	幫	0.284865
9	去	0.284556
10	靠	0.28287
11	為	0.279203
12	中肯	0.274776
13	按	0.274399
14	明	0.271547
15	根據	0.265427
16	關	0.264634
17	增加	0.264406
18	論	0.260204
19	需要	0.258214
20	主	0.253801
21	難得	0.253608
22	賠	0.253037
23	講	0.252759
24	依	0.251863
25	仍	0.2517

26	存在	0.250033
27	殺	0.249134
28	依照	0.246334
29	望	0.24469
30	依據	0.243787
31	進行	0.243654
32	起來	0.242004
33	據	0.241924
34	司	0.241415
35	照	0.237104
36	熱門	0.236161
37	配合	0.236136
38	衰	0.235031
39	解決	0.234676
40	鬥	0.232501
41	按照	0.231994
42	實現	0.223663
43	憑	0.221721
44	吵	0.219342
45	熱	0.218193
46	准	0.214435
47	冷門	0.214433
48	懷念	0.213032
49	可用	0.212126
50	反串	0.205794
51	熱情	0.19992
52	餓	0.199058





53	熊熊	0.197607
54	遷	0.19665
55	冷靜	0.19475
56	抽菸	0.193983
57	熱血	0.193966
58	汙染	0.193539
59	萌	0.189759
60	評估	0.18858
61	忽視	0.187261
62	豫	0.187
63	反彈	0.184916
64	冷清	0.184443
65	展示	0.181978
66	宰	0.181612
67	淡定	0.180105
68	粉	0.177767
69	回饋	0.169777
70	抽煙	0.167109
71	冷卻	0.15597
72	吸血	0.153999
73	憂傷	0.148936
74	苦惱	0.148467
75	加熱	0.144578
76	飢	0.141772
77	使然	0.139135
78	嬉	0.138439
79	穿越	0.137304



80	溫暖	0.133948
81	信服	0.131225
82	斷定	0.128743
83	暖身	0.127459
84	結拜	0.12569
85	冷場	0.12557
86	快活	0.125316
87	候補	0.124862
88	憂鬱	0.121629
89	憂心	0.120536
90	丟棄	0.118879
91	冷淡	0.116712
92	搾	0.115672
93	談論	0.114752
94	秉	0.114677
95	鄰近	0.112922
96	引入	0.111832
97	冷漠	0.106467
98	吸菸	0.10544
99	憂愁	0.104812
100	致電	0.103968
101	吸煙	0.100525
102	靜坐	0.100317
103	緊迫	0.099856
104	自滿	0.084862
105	再版	0.084223
106	憂慮	0.073977

107	逾越	0.069322
108	熱場	0.064402
109	憑藉	0.062677
110	熱身	0.059502
111	卞	0.055252
112	失序	0.053236
113	暖場	0.051142
114	選任	0.044276
115	發愁	0.036139
116	如次	0.036139
117	煩憂	0.029501
118	關於	0.029501
119	謳歌	0.020856
120	悲愁	0.014746
121	愁悶	0
122	憂惱	0
123	憂煩	0
124	並軌	0
125	首由	0
126	茲誌	0
127	繼由	0
128	荒無人煙	0





Appendix 4 Constant U value for Lexical Items After 1950

	lexical.items	constantU_yearmonth_comment
1	違規	0.250621
2	本日	0.243929
3	測試	0.241627
4	搞笑	0.22459
5	認同	0.221773
6	調高	0.217755
7	搞定	0.216255
8	考量	0.215162
9	抓包	0.206056
10	加減	0.20386
11	吐槽	0.202901
12	榨	0.202638
13	變身	0.20015
14	卡位	0.199528
15	蝦米	0.193661
16	上網	0.192646
17	預估	0.182994
18	跳槽	0.177702
19	解套	0.174464
20	牽拖	0.171905



21	閃人	0.170896
22	帶動	0.170386
23	菜鳥	0.169444
24	互動	0.167576
25	促銷	0.162175
26	耍帥	0.161794
27	嗆聲	0.152005
28	破功	0.150462
29	有助於	0.150126
30	惦惦	0.150002
31	仲介	0.14317
32	槓龜	0.143083
33	哈啦	0.141496
34	跳票	0.140218
35	有型	0.139178
36	死忠	0.137543
37	外掛	0.137495
38	嚇嚇叫	0.136193
39	暗爽	0.136008
40	抓狂	0.13424
41	全職	0.133752
42	焗	0.133653
43	落實	0.133486
44	定今	0.133226
45	歹戲拖棚	0.12966
46	瞎掰	0.129542
47	臭屁	0.128658



48	旅遊	0.128089
49	放鴿子	0.126679
50	冷血	0.126221
51	搞怪	0.125996
52	雞婆	0.124519
53	阿達	0.124174
54	續攤	0.122283
55	融資	0.119672
56	大尾	0.119181
57	假仙	0.119103
58	瘦身	0.118842
59	持股	0.118721
60	收驚	0.117991
61	對盤	0.117963
62	發飆	0.116659
63	幹架	0.113649
64	分租	0.112766
65	顧人怨	0.111528
66	打拚	0.111013
67	落跑	0.110412
68	秀逗	0.110202
69	撇清	0.109612
70	投信	0.107578
71	生猛	0.107205
72	脫窗	0.106162
73	辦桌	0.10545
74	上櫃	0.102003



75	速配	0.101298
76	膨風	0.10083
77	創投	0.100504
78	建構	0.100003
79	鬱卒	0.094209
80	打通關	0.094039
81	牽絲	0.091267
82	做怪	0.089412
83	精煉	0.088704
84	沒皮條	0.088266
85	條直	0.086457
86	晃點	0.085436
87	鬥陣	0.085329
88	閃神	0.085004
89	血拼	0.084997
90	善變	0.084172
91	比拼	0.083777
92	撿角	0.082521
93	踢館	0.081022
94	卒仔	0.079819
95	研判	0.078258
96	穿幫	0.076078
97	強強滾	0.075394
98	作伙	0.073922
99	塑身	0.072911
100	吃螺絲	0.072723
101	鴨霸	0.072723



102	俗擱大碗	0.070888
103	逗陣	0.067023
104	了不了	0.066082
105	建置	0.065402
106	間接接吻	0.062677
107	暗槓	0.062677
108	甲意	0.062677
109	蓋高尚	0.060905
110	亮光	0.060302
111	踢鐵板	0.05908
112	開新板	0.057197
113	一把罩	0.053978
114	釘孤支	0.051142
115	呷意	0.04896
116	搓圓仔湯	0.048711
117	耍酷	0.046676
118	燒滾滾	0.046676
119	趴帶	0.046676
120	插一腳	0.044276
121	網路上身	0.044276
122	相輸	0.041913
123	瞎拼	0.041739
124	全民開講	0.039039
125	俗俗賣	0.039039
126	釘孤枝	0.039039
127	打破	0.036139
128	專電	0.036139

129	搞飛機	0.036139
130	叩應	0.036139
131	知影	0.036139
132	破病	0.036139
133	征收	0.029501
134	嘎嘎叫	0.029501
135	大車拚	0.029501
136	秘雕	0.029501
137	莫宰羊	0.020856
138	漂染	0.020856
139	易貨	0.020856
140	哈草	0.020856
141	摸蜆	0.020856
142	大俗賣	0.020856
143	英英美代 子	0.014746
144	犯愁	0
145	愁苦	0
146	焦心	0
147	哈一支	0
148	耍炫	0
149	倒豎姆指	0
150	健胸	0
151	理財	0
152	連結	0
153	連線	0
154	釣妹妹	0





155	嗑網	0
156	慢性自殺	0
157	敲桿	0
158	稽征	0
159	講刀巴話	0
160	颯舞	0
161	自助旅行	0
162	酷斃了	0
163	卯死了	0
164	全身美白	0
165	忍未條	0
166	車拚	0
167	虎爛	0
168	相招逗陣	0
169	臭蓋	0
170	莫法度	0
171	喇雷	0
172	搶鬧	0
173	一元捶捶	0
174	老神在在	0
175	阿哩不達	0
176	閉淑	0
177	無三小路 用	0
178	損龜	0
179	篤爛	0
180	龜毛	0



Appendix 5 Constant U value for Diffused Lexical Items

	lexical.items	constantU_yearmonth_comment
1	丐丐	0.294443
2	科科	0.272053
3	劣退	0.270267
4	低調	0.269966
5	END	0.265144
6	神	0.257006
7	黑	0.25684
8	八卦	0.254749
9	搜尋	0.251239
10	神人	0.2489
11	不解釋	0.24184
12	顆顆	0.232269
13	頗厂	0.231561



14	機車	0.224454
15	開燈	0.224297
16	筆記	0.222584
17	躺著也中槍	0.222429
18	贊	0.222325
19	GG	0.221227
20	人肉	0.220239
21	腿	0.219792
22	BJ4	0.219313
23	囧	0.215757
24	厂厂	0.214975
25	暈	0.212564
26	沒壞	0.205843
27	踹共	0.205651
28	OP	0.199915
29	低調推	0.194522
30	頗呵	0.194353
31	打臉	0.188494
32	高調	0.187185
33	高富帥	0.184262
34	給力	0.179641
35	R.I.P.	0.179492
36	測風向	0.175235
37	根本呵呵	0.165406
38	頗喝	0.164196
39	幫高調	0.161484
40	關燈	0.161479



41	台肯	0.160532
42	富奸	0.158348
43	CD	0.147237
44	根本厂厂	0.147223
45	神馬	0.143667
46	帶風向	0.14323
47	Lag	0.137426
48	力挺	0.133529
49	草泥馬	0.119844
50	無厘頭	0.117838
51	娘炮	0.116786
52	坑爹	0.114926
53	雷人	0.114238
54	站台	0.113952
55	打醬油	0.112138
56	累格	0.105852
57	牛逼	0.104091
58	低調噓	0.101637
59	稀飯	0.101615
60	蛋疼	0.093939
61	飲茶	0.077539
62	有木有	0.076133
63	人肉搜索	0.075394
64	屌炸天	0.071336
65	凸槌	0.067023
66	碉堡了	0.051142
67	華肯	0.036139

68	接地氣	0.020856
69	富奸化	0
70	拍磚	0

