王伯雅　　國立台灣大學語言學研究所
論文題目(中文)：
詞彙穩定的秘密—對各語言學面向的質性與量化分析
論文題目(英文)：
Secrets of Lexical Conventionalization:
A Quantitative and Qualitative Exploratory Analysis on Linguistic Factors
指導教授：謝舒凱

論文摘要

Language variation and change have been widely investigated since they are encapsulated phenomenon involving many linguistic factors. However, scant attention has been paid to understand factors contributing to how lexical items are adopted into lexicon from both quantitative and qualitative perspectives. Thus, the focal point of this thesis is the *lexicalization*, which is regarded as diachronic processes subject to normal constraints on language change (Brinton and Traugott, 2005). The term lexicalization in previous studies, sometimes equated with semantic change in general, can refer to the newly coding of conceptual categories in the synchronic sense, or the process of adopting into the lexicon in the diachronic sense. Both perspectives on the lexicalization reveal the complexity of this multivariate phenomenon, and suggest a proper treatment at local/synchronic and global/diachronic level. To better capture the dynamic nature and underlying cognitive mechanism of lexicalization in Chinese, this paper proposes to present the quantitative profile of a set of target lexical items based on large-scaled web corpus (i.e., google book ngram corpus and PTT corpus) , and providing cognitive-functional linguistic explanation as well.

Previous related works in the field of historical linguistics, lexical semantics, and computational linguistics have shown insights in understanding lexical semantic changes from different perspectives. However, little research has been done on adopting both quantitative and qualitative methods to delineate the picture of fluctuation of lexical items. Besides, the generality of included target words, their temporal information and other linguistic aspects should all be considered to have deeper understanding of factors contributing to conventionalization of a word. Therefore, this study aims to provide quantitative profiling and qualitative analysis within proposed three life stages of lexical items (diffusion, conventionalization, and inactivation, cf. Kerremans, 2012), focusing on target words from different temporal points, and employing linear models with twenty one linguistic variables from six linguistic aspects (phonology, morphology, semantics, syntax, pragmatics, and

sociolinguistics).

In regard to quantitative profiling, we have obtained about one hundred million data from 2000 to 2014, including posts on twenty popular boards as well as all of the comments in these posts. Multifarious topics are included: games, gender issues, emotions, economics etc. Linear regression, logistic regression, and prediction model are the three dimensions probed in this study to profile the quantitative characteristics. Linear regression model is built to understand highlighted linguistic aspects for words from different temporal points. The result indicates that pragmatic information can best account behavioral performance of *words over a century*(e.g. “上,”“去,” “有”), while syntactical one best captures *words born after 1950*, those who were once diffused words sixty years ago, but now fluctuate differently in use(e.g. “抓包,”“認同,” “違規”). This implies that words live longer may be correlated with rich experiential and pragmatic using world knowledge, but for those who are newly coined, their structurally syntactic compatibility plays vital role in deciding their future fluctuation in use.

Given that *diffused words* (e.g. “ㄎㄎ,”“低調,” “劣退”)are similar to words existing over centuries in their distribution of revised constant U (monthly average frequency divided by standard deviation of total frequency), logistic regression model is constructed to sketch differences between words over a century and diffused words. 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 five statistically significant variables that distinguish diffused words and words existing over centuries. On the other hand, words coined after 1950 and diffused words show similarities in their linguistic characteristics. That is, words coined after 1950 are once diffused words, so their later fluctuations are able to suggest possible future fluctuated conditions present diffused words may meet. Thus, prediction model based on training data from words after 1950 is built to foretell potential life of diffused words. It shows that "number of types co-occurring before target words" is statistically valued in presage. To further testify, words that have existed over hundreds of years, and recent diffused words are taken as test data. The accuracy of the test result reaches 0.6335.

As for qualitative analysis, two issues are discussed: competitions among words from the same synset as well as sketched linguistic characteristics for words from different temporal points. To the first issue, by analyzing competitions among words from the same WordNet *synset*, it is concluded that “structural compatibility” and “involved conceptual relations” may be the key for one lexical item to winning over the other synonymous member. When it comes to the second issue, words coming from different temporal points show disparities in their activeness of being used in

comments and posts in social media like PTT. Diffused words are more actively used in comments. This implies that they are closely interrelated with feedback-oriented oral style and diffused through interaction.

The findings we have achieved in both quantitative profiling and qualitative analysis can further be applied to construct resources of lexicography. In general, pragmatically stable in use, syntactic compatibility and semantic numbers of senses are suggested to be taken as standards for expanding inclusion of words in dictionary. A preliminary pilot study on updates of Huayu 8,000 Chinese words is conducted based on above mentioned criteria. The resulted updated wordlist proves that the inclusion is comprehensive, for it contains words that are popularly used variants (“抽菸” instead of “抽煙”), lexical items that are more stable semantic representations (“吸煙” instead of “抽菸”), and related vocabularies lexicalized from the same conceptual experiences (e.g. “冷淡,” “冷血,” “熱情,” “熱門,” “熱身,” “暖身”).

In brief, though there are still many future directions for further studying, present work has contributed to propose elements that influence lexical items to be adopted into lexicon. To begin with, it is quantitatively proved that pragmatic world knowledge and structurally syntactic compatibility play statistically different roles to words in different temporal points. Besides, five statistically significant linguistic variables are anchored to distinguish diffused words and words existing over centuries. Third, “number of types co-occurring before target words” is a key factor in predicting latent fluctuation of present diffused words based on the testing result. On the other hand, results of qualitative analysis also provide imperative insights that structural compatibility and involved conceptual relations may be the keys for one lexical item to winning over the other synonymous member. Meanwhile, diffused words are more actively used in comments. This phenomenon corresponds to their characteristics of correlating with feedback-oriented oral style and of being diffused through interaction. In addition, the application of these found results on updating wordlists also corroborates that the features extracted from this study are workable to understand factors contributing to influence conventionalization of lexical items.