摘 要

汉语研究中对长短被字句的争论由来已久，本研究从跨语言变体的角度出发，基于“中文十亿字标注语料库”，对汉语“被”字构式 NP2（即紧随“被”字后的名词性成分，一般为施事）的隐现进行了多因素分析，总结出了影响 NP2 隐现的重要因素以及不同汉语变体间 NP2 隐现的差异。

本研究从三种语言变体（大陆汉语、台湾汉语和新加坡汉语）的语料中抽取并清理得到 2918 个目标句，进行人工标注：因变量为 NP2 的隐现；自变量共 16 个，包括 1 个语言外部变量（汉语变体），以及与句法和语义相关的 15 个语言内部变量。以“被”字构式的核心动词为随机效应，我们建立了混合效应逻辑斯蒂回归模型，筛选出对 NP2 隐现有显著影响的因素并观察与变体间有显著交互效应的变量；其次，我们基于三种汉语变体各自的语料，分别建立随机森林和决策树模型，得到各变体影响 NP2 隐现的显著因素及其重要性排序。

模型结果显示，“被”字构式 NP2 的出现和消隐具有各自不同的典型用法。当谓语为不及物、NP2 无生，被动事件描述的是 NP1 在被动作用下产生物理或生理的变化时，NP2 倾向于出现；当 NP1 为不定的具体或抽象名词，“被”字构式出现在定语位置或从句中，被动句为否定句时，NP2 倾向于消隐。以上特征可以从语言经济性原则、尾重原则、“被+V+N”的词汇化以及被动句自身特点对 NP2 隐现的要求这四个方面得到解释。新加坡汉语、大陆汉语和台湾汉语中“被”字构式 NP2 隐现的共时差异，反映出语言内部约束在变体间具有差异性以及不同的文化社会背景对语言使用产生影响。基于新加坡汉语的欧化、台湾汉语中保留的其他被动标记（如“遭”和“获”）以及大陆汉语口语化的独立性发展，我们可以从跨语言变体的共时差异中进一步总结并预测汉语历时变化的特征。

本研究的创新点在于：首先，结合大型语料库，利用多因素方法，我们对“被”字构式 NP2 的隐现有了更细颗粒度的描写；其次，本研究跨变体角度的融入也是具有开创性的；最后，结合概率语法框架和认知社会语言学理论，我们对“被”字构式 NP2 隐现所对应的典型特征以及变体差异进行了分析和归因，弥补了前人研究的不足。

关键词：“被”字构式 NP2 隐现；汉语变体；概率语法；多因素分析

Abstract

There has been a long debate about the long vs. the short *BEI* passive sentences in Chinese. This study is a multifactorial analysis of NP2 (the agent, typically) occurrence in Chinese *BEI* construction from a cross-variety perspective. We retrieved one-month data of mainland Chinese, Taiwan Chinese and Singapore Chinese respectively from the "Tagged Chinese Gigaword Corpus" and all three shares nearly the same words in the same period of time. We got passive sentences marked by *BEI* automatically using AntConc and further checked them manually under the principle of interchangeability. After that, we randomly selected 2918 observations (30% of valid observations from these three varieties) to annotate. Our dependent variable is the occurrence of NP2, while the independent variables include 15 language-internal factors and a language-external factor (lectal variation). Taking verbs in *BEI* constructions as the random effect, we built a mixed-effect binary logistic regression to obtain significant factors and interactional effects between lectal variation and any language-internal factors. At last, we built random forests and conditional inference trees for three varieties respectively to obtain their own significant factors as well as their importance rankings of all predictors.

The results show that *BEI* constructions with or without NP2 have totally different prototypical usages. For *BEI* construction with NP2, the predicate tends to be intransitive, and NP2 are more likely to be inanimate with the passive event describing a physical change of NP1. Whereas, for *BEI* construction without NP2, with an indefinite concrete or non-concrete NP1, it is more likely to appear in clauses or in negative sentences. All those prototypical usages are able to be explained by the principle of language economy, the principle of end-weight, the lexicalization of "*BEI* + V + N" construction as well as the universally internal demands of the passive sentence itself. Among those three lectal varieties, *BEI* constructions with or without NP2 haven't shown significant difference, but they all prefer to choose *BEI* construction without NP2. However, according to the significant interactions between lectal varieties and the semantic categories as well as the topics of passives, and the different significant factors selected in three varieties, we concluded that the language-internal constraint is of diversity across language varieties, and the cultural and social background plays an important role on the language use. Furthermore, the synchronic lectal variation was explained by the Europeanized Singapore Chinese, the

conservation of other passive markers (e.g., *ZAO* and *HUO*) in Taiwan Chinese and the independent development of mainland Chinese towards colloquialization. Therefore, from the synchronic lectal variations, we are able to conclude and predict the diachronic change and development of Chinese.

There are three main innovations of our research. Firstly, as a corpus-based multifactorial analysis, this research obtains a fine-grained description of the NP2 occurrence of Chinese *BEI* passive construction. Secondly, the combination of cross-variety perspective also made our research being groundbreaking. Thirdly, working within the framework if Probabilistic Grammar and Cognitive Sociolinguistics, we are able to further analyze and explain the prototypical usages of Chinese *BEI* passive construction with or without NP2 and its lectal variant differences, which fills the gap of previous research.

Keywords: NP2 occurrence in Chinese *BEI* construction; lectal variation; Probabilistic Grammar; multifactorial analysis