The *Journal of Chinese Linguistics* vol.50, no.2 (June 2022): 443−474

EARLY EVIDENCE OF THE PARETO PRINCIPLE IN GRAMMATICAL DISTRIBUTION: CAUSATIVE SITUATIONS IN CHINESE CONVERSATIONAL DISCOURSE

Danjie Su
University of Arkansas, Fayetteville

ABSTRACT

This study is an initial report on Pareto distribution (the 80/20 rule) of grammatical constructions; namely, about 20% of the types of grammatical constructions for causative situations account for about 80% of the uses in conversation. I use a data-driven approach to investigate the grammatical constructions that Chinese L1 speakers choose in spontaneous talk show conversations to describe causative situations. I identify two specific Pareto distributional patterns. 1) The distribution of all 22 constructions for causative situations constitutes a Pareto ABC diagram with the A-class (*ba*-; unmarked passive; *rang*-; *bei*-; resultative; *gei*-) containing 27.3% of the types but accounting for 88.8% of all the 1,497 uses. 2) Most uses of a grammatical construction come from a small set of subtypes: The full *ba*- accounts for 87.9% of all *ba*- uses; the reduced *bei*- accounts for 86.8%; 37.5% of *rang*- subtypes account for 84.2%. These patterns can be

Acknowledgements This article is a substantial revision of a chapter of my dissertation (UCLA, 2017). I greatly appreciate the insightful input from Hongyin Tao, Charles Goodwin, Shoichi Iwasaki, and Sung-Ock S. Sohn. I thank the Office of the Vice-Chancellor for Research and Innovation, World Languages, Literatures & Cultures, and Asian Studies at the University of Arkansas for funding and Phoenix Television for data. I am also grateful to the reviewers for their constructive and helpful suggestions. I alone take responsibility for any shortcomings of this article.

Danjie Su (苏丹洁) [danjiesu@uark.edu]; Department of World Languages, Literatures & Cultures, Kimpel Hall 425, University of Arkansas, Fayetteville, AR 72701, USA;
https://orcid.org/0000-0002-7444-577X

explained by the Lens concept. I conclude that a few constructions account for most grammatical choices of L1 Chinese speakers in conversation. Understanding these grammatical distributions in natural discourse can improve the efficiency and efficacy of language teaching and Natural Language Processing (NLP).

KEYWORDS
Grammatical construction **F**requency **L**ens **C**hoice **M**andarin

1. INTRODUCTION

In linguistics, the famous Zipf's word frequency law (Zipf 1935) has inspired much research on word distribution in natural discourse (Piantadosi 2014). Zipf's law reveals that "the frequency of a word is inversely proportional to its statistical rank" (Fagan and Gençay 2010) in a natural corpus. Against the backdrop of much research on word distributions, this study turns to the distribution of grammatical constructions. This study examines the distribution of all the grammatical constructions that speakers use for causative situations in conversation—a topic that no existing publications have focused on (section 2.2). My research finds no Zipfian inverse correlation of frequency and rank but a strikingly simple pattern that reflects the Pareto principle (the 80/20 rule): About 20% of the types of grammatical constructions account for about 80% of their actual uses in conversational discourse.

The Pareto principle is widely observed in many fields outside linguistics. Its essence is that about 80% of an effect comes from approximately 20% of the causes. Economist Vilfredo Pareto observed that approximately 80% of the land in Italy was owned by 20% of the population (Pareto 1896–1897). Joseph Juran (Juran and Godfrey 1998) termed this distribution the Pareto principle and applied it to quality management in organizational operation. This 80/20 rule has been reported in many social activities. In libraries, 80% of library circulations involve 20% of holdings (Buckland 1975), and 24.7% of public library patrons borrow the majority (75.3%) of the collections (Yang and Shieh 2019). Digital databases such as Elsevier SD and Wiley Blackwell consistently exhibited the Pareto principle in full-text downloads (Zhu and Xiang 2016).

ZHU, Qiandong, and Huimin Xiang. 2016. Differences of Pareto principle performance in e-resource download distribution: An empirical study. *The Electronic Library* 34(5): 846–855.

ZIPF, George K. 1935. *The Psycho-Biology of Language*. Boston: Houghton Mifflin.

语法呈帕累托分布的早期证据：现代汉语自然会话中致使情景的语法构式分布

苏丹洁
阿肯色大学

摘要

本研究是关于自然会话中语法构式的帕累托(Pareto)分布(二八法则)的第一份报告——大约 20%的语法构式类型占表述致使情景的所有实际用例的 80%。基于脱口秀自然会话语料，本文使用数据驱动的方法穷尽式地探究汉语母语者选择何种语法构式表述会话中的致使情景。本文关于帕累托分布的具体发现是：(一)会话中表述致使情景的所有22种汉语语法构式的分布反映了帕累托原理及其ABC等级分布。A级的构式类型数量为22种构式类型的27.3%，却占到所有1,497条用例的 88.8%。A 级包括的最高频构式依次是：把字句、无标记被动句、让字句、被字句、结果补语、给字句。B 级的构式类型数量同样占27.3%，却仅占所有用例的 8.9%。C 级的构式类型数量占了近一半(45.5%)，却只占所有用例的 2.3%。(二)语法构式的大多数用例来自个别子类型：完整版把字句占所有把字句用例的 87.9%；减短版被字句占所有被字句用例的 86.8%；37.5%的让字句类型占所有让字句用例的 84.2%。Lens 理论可以解释这些分布规律。本文结论是，汉语母语者在自然会话中选用少数构式类型来表述绝大部分致使情景。该发现进一步揭示了自然话语中语法构式的分布，这对语言教学和自然语言处理具有直接参考价值。

关键词

语法构式 频率 滤镜 选择 现代汉语