UC Merced

Title

Diachronic Entropy Rate in Language Evolution: A Case Study of 2500 Years of Historical Chinese

Permalink

Journal

ISSN

Authors

Sun, Hao
Jin, Yanwei

Publication Date

2021

Peer reviewed

Diachronic Entropy Rate in Language Evolution: A Case Study of 2500 Years of Historical Chinese

Hao Sun
Pearson, San Jose, California, United States

Yanwei Jin
University at Buffalo, Buffalo, New York, United States

Abstract

Information theory (Shannon, 1948) plays an important role in psycholinguistic and linguistic theories (Genzel & Charniak, 2002; Hale, 2003; Levy, 2008). Here, we examine how entropy rate, a measure of information content encoded in each individual word, changes diachronically in Chinese. We conduct a computational study on the four main development stages of Chinese, Old Chinese, Middle Chinese, Early Modern Chinese and Modern Chinese. We approximate entropy rate of each century by adopting a diachronic trigram language model with interpolated Kneser-Ney smoothing technique (Chen & Goodman, 1999), which is trained on multiple comprehensive data sets selected according to Chinese philology studies (Wang, 1980; Gao & Jing, 2005) covering over 2,500 years of corpus data. Our modeling results show that entropy rate, on average, increases 0.026 for each century. Within each major stage, historical Chinese demonstrates a steady rise in entropy rate, suggesting a vocabulary increase whereas entropy rate tends to fluctuate more in transitional stages, around the 10th century and the 15th century, lending support to the hypothesis that grammar competition in language contact is one of the driving forces behind major changes in diachronic Chinese. Our study demonstrates the interaction between psycholinguistic pressures and the evolution of linguistic systems.