Finding probabilistic context-free grammar in Chinese writing system

Hao Sun
Astound.AI, Menlo Park, California, United States

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

Abstract

Writing systems play a very important role in human languages, but the mathematical nature of writing systems remains understudied. Here, we conduct a case study of an open-class writing system Chinese characters, which consists of a set of expandable basic units, in contrast to most other writing systems whose basic units form closed sets, or closed-class systems. We demonstrate that probabilistic context-free grammars underlie the representation of Chinese writing, by formalizing Chinese characters as a grammar with character shapes, as nonterminal rules, and components. as terminal nodes. Rule probabilities are estimated from a character treebank of the most frequent 3500 characters. Exploratory analysis reveals Zipfian distributions of both shapes and components. Our experiments also demonstrate that Chinese writing system shows generative powers similar to PCFG, with 78% of the noncharacters generated from our grammar judged acceptable, which suggests fundamental differences between open-class and closed-class writing systems.