基于字词计量的汉语语言使用能力探究

祁晶 孔江平 吴西愉
北京大学

字和词是汉语语言表达的基本单位，对其进行计量研究可以帮助我们更好地掌握汉语的使用特点，从而为词汇学、词典学、对外汉语教学以及自然语言处理的相关研究提供参考。有研究者关注这种计量背后体现的人的语言认知特征，提出了“词涯八千”的著名推断，但其所谓的“词”是词素，即字，所例证的材料也并不全面。针对这种情况，本研究希望以跨越古今的 88 位作家的 626 部作品为研究对象，探究个体在使用语言时的字和词种类数的特点，从而反映出个体使用语言的基本能力。从共时和历时角度分析字词汇，从而探究在共时和历时两个层面字词汇使用的特征。根据对字词种类的计量结果，我们发现“字涯”的确是存在的，对于个人表达者，其上限在 5000 左右，而“词涯”的上限并不能确定。对于个体在语言中的“字涯”及词种数在历史上变化的考察，相关结果表明其使用能力在历时上并没有显著增加。这提示我们，自有文字来，个人使用字词单位的能力很可能并没有太大的增长。通过对共时和历时层面通用字词汇的探索，我们发现在同一个时期，一个作家与通用字汇的沟通度可以到 97%-99%，词汇的沟通度要低一些，大概在 60%到 70%之间。在历时变化上，通用词汇的变动要大于字汇，这可能与字汇在古代就已基本达到上限有关。历时上通用词汇的词长存在显著的多音节化趋势，二音节词逐渐占据主流，这提示我们思考，“字涯”的存在可能是历史上词长变化的一个动因。

关键词：计量语言学，词频分析，共时与历时研究

Exploring Chinese Language Using Ability —— Based on Quantitative Study of Characters and Words

Characters and words are the basic units of Chinese language expressions. A quantitative study can help us capture the characteristics of Chinese language usage, thus providing a reference for the studies on vocabulary, dictionary, Chinese teaching and natural language processing. Some researchers have focused on the cognitive characteristics of human language usage behind this measurement. Zheng had proposed a famous "8,000 words boundary" hypothesis, but the materials he illustrated were not sufficient. Therefore, this study hopes to investigate the features of Chinese characters and words used by individuals, thus reflecting their basic ability to use language. The material includes 626 works written by 88 authors spanning from the ages to the present. Based on the results of the measurement of character types, we found that the "character limit" does exist, with an upper limit of about 5,000 for personal expressions. But the boundary of words cannot be determined based on our research data. The statistical analysis of the "character limit" of individuals and the number of word varieties in language usage is not sufficient to support the assertion that their ability has increased over time. By exploring the universal vocabulary at the synchronic level, we find that a writer can use the universal characters to express 97%-99% of the information. The number is lower for universal words, at 60%-70%. At the diachronic level, variation in universal words is greater than that in universal characters. At the same time, there is a significant tendency of disyllabification of the universal lexicon over time. The results led us

to consider that the existence of the " character limit " may have been a motivation for the change in word length throughout history.

Keywords: quantitative linguistics, word frequency analysis, diachronic and synchronic studies