摘要

本文研究中国手语语料库中的高频词及其标注问题。主要围绕两个问题展开：一是对高频手语词的分布情况进行研究，并进行跨语言对比，发现中国手语高频词与其它手语、有声语言相比之下的共性和个性。二是对手语高频词本身的结构进行语音和词法层面的分析，结合手语语言学的知识更好地处理语料库中手语词的标注问题。

我们发现中国手语语料库前 500 高频词涵盖了近 80%的词频分布，对前 100 高频词的分析中发现：相比于汉语，中国手语中的实词占比更大，虚词中代词占超过一半比重，这也体现为手语的一种共性。我们把中国手语与美国手语、土耳其手语进行了比较，发现三种完全独立的手语在高频词分布上有相当比重的类似之处，而同一种文化下的手语和口语之间却相差较大，再次证明手语是一门独立的语言。另外，我们分析了高频手语词的语音和词法特征，发现语言经济性原则同样在手语高频词语音特征分布上起作用，即高频的手语词语音上更简单，有更多语音变异发生。在词法结构方面，我们分析了词缀等构词方式，发现了中国手语高频词总体上是单音节多词素，以及中国手语有较高比重的类标记复合词。最后，我们对手语词的标注问题做了一些探索，设计并提出了一些标注模式和流程。

关键词：中国手语；手语语料库；高频词；词法；标注

Abstract

This paper studies the high-frequency words in the Chinese Sign Language(CSL) corpus. It mainly focuses on two issues: one is to study the distribution of high-frequency signs, and implement cross-language and cross-modal comparisons to discover what is common across languages, and what is particular to sign languages especially CSL. The second is to explore the phonetic and lexical structure of the high-frequency signs. We hope to combine the knowledge of sign language linguistics to better deal with the problem of sign language annotation in the corpus.

We found that the first 500 high frequency words cover nearly 80% of the total lexicon. The analysis of the first 100 words showed that compared with spoken language, sign language has a larger proportion of content words, pronouns account for more than half of the function words, which is also shared across sign languages. Comparing with American Sign Language and Turkish Sign Language, we found a considerable share of similarities in the distribution of high-frequency words among the three completely independent sign languages. In short, sign languages are more similar in the distribution of high-frequency words, while sign language and spoken language under the same culture are quite different. This proved once again that sign language is an independent language. We analyzed the phonological and lexical features of high-frequency CSL signs. We found that the principle of economy also worked on the phonetic distribution of high-frequency signs, that is, high-frequency signs are phonetically simpler and have more phonetic variants. When we analyzed the morphological structure and word formation such as affixes in CSL signs, we found that high-frequency words are generally monosyllabic and multi-morphological and CSL have high proportion of classifier compound words. Finally, we focused on the annotation of sign language words, design and put forward some annotation patterns and processes.

Keywords: Chinese Sign Language; Sign Language Corpus; High Frequency Words; Morphology; Annotation