汉语语言产生数据库发表

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/26218.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

汉语语言产生数据库发表。说话，看似毫不费力，但其实是人类最复杂的认知活动之一。在过去几十年里，研究人员试图澄清言语表达（又称为语言产生）背后的认知架构和动态加工过程。近日，中国科学院心理研究所助理研究员冯臣、研究员屈青青与合作者在Nature旗下的《科学数据》（Scientific Data）发表汉语语言产生数据库。

已有的研究显示，世界上大约有7000种语言，不同语言的词汇、语音、字形系统大相径庭。现有的实证研究与理论框架主要基于印欧语系。与印欧语言不同，汉语有独特之处。

首先，汉语属于汉藏语系，是一种声调语言，同一个音节的不同音调表示不同的词汇与意义，大多数印欧语言则不使用声调来区分词义。其次，印欧语言采用字母拼写系统，而汉语采用非字母书写系统。另外，汉语的语音与字形的对应关系较为复杂，例如，声音shu可以对应书梳树或薯等多个不同的字形。因此，汉语研究对回答跨语言普遍性和特异性具有独特价值和贡献。

虽然一些研究已经表明，音位是印欧语言语音编码的主要加工单元，而音节是汉语语音编码的主要加工单元，但目前大部分关于语言产生的研究侧重于印欧语言，而对于汉语这一非印欧语言的研究则相对有限。

冯臣和屈青青等人发现，这些研究主要关注于口语产生过程，对于书写等字形产生的关注则明显不足。同时，许多研究依赖于较小的样本量和有限的实验刺激，这可能导致统计效力不足和研究结果的可重复性问题。构建汉语语言产生数据库的需求与日俱增。

此次发表的汉语语言产生数据库记录了667名被试在7种不同语言产生任务中的约20万个试次的反应时间，以及实验材料的多个语言学变量，如词频、字频、音节频率、习得年龄等。

据介绍，该数据库为研究汉语口语和书写产生的语言加工提供了丰富数据资源，为探索汉语产生的普遍性与特异性提供了数据基础。此外，该数据库对于开发汉语语言产生的人工智能模型具有独特价值与贡献，有望推动语言认知科学与人工智能领域的交叉与共同进步。（来源：中国科学报 沈春蕾）

相关论文信息：https://doi.org/10.1038/s41597-024-03022-8

作者：冯臣等 来源：《科学数据》

更多 科学进展 请访问 https://www.iikx.com/news/progress/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发