心理所创建中文阅读眼动数据库

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/19334.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！

现代认知科学表明，眼睛的运动和大脑内部的认知活动存在紧密联系。在心理学领域，眼动追踪技术为揭示人类高级认知过程的心理机制提供了重要途径。在过去的20年中，大量实验研究采用眼动追踪技术考察了中文阅读的认知机制，并取得了丰硕成果。然而，大多数实验研究受到被试量和实验刺激数量的限制，很难满足当前大数据技术和人工智能的发展。因此，建构包含大样本的中文阅读眼动数据库的需求与日俱增。

基于过去十余年的研究，中国科学院心理研究所行为科学重点实验室李兴珊团队研究人员建立了大规模的中文阅读眼动数据库Chinese Eye-Movement Database。该数据库包含来自57项中文句子阅读实验的眼动数据（包含1718名被试、8015个中文句子、近140万个注视点），计算了8551个中文词的九项眼动指标（图1）。统计分析显示，该数据库可以复现以往研究中经典的词频与词长效应，即读者对较低频或较长的词加工更困难，从而产生更多的回视和更长的注视时间。

该数据库具有广泛的应用前景。在中文阅读的认知机制研究中，研究人员可直接利用该数据库检验相关的理论假设，节约经济和时间成本；同时，该数据库可以为建立中文阅读计算模型提供基准数据，帮助其进行参数寻优。在跨语言研究领域，该数据库可与其他语言中的同类数据库进行对比，考察不同语言阅读机制的一致性和特异性。在人工智能领域，自然语言处理的模型大量使用了与注意相关的机制（如为不同词汇分配不同的权重或激活状态），而眼动数据则为这种注意的分配提供了直接参考；大量研究表明，将眼动数据纳入自然语言处理模型，能够有效提升模型的任务表现（如词性标注、句法分析、文本理解等）。因此，数据库将为优化中文自然语言处理模型提供重要的数据资源。该数据库中报告的词汇的眼动指标可作为反映词汇阅读加工难度的指标，帮助研究者更好地控制和操纵实验研究中阅读材料的难度，并有助于为不同阅读能力的读者匹配合适的阅读材料。

综上所述，该数据库将为中文阅读认知机制的大数据研究提供重要支撑，促进该领域的发展，也将为人工智能领域的模型开发与训练提供数据基础，促进人工智能与认知科学的融合发展。

相关成果已在线发表于*Scientific Data*。数据库所涉及的全部原始注视点数据、实验材料，以及数据分析代码已全部通过Open Science Framework共享。

论文链接

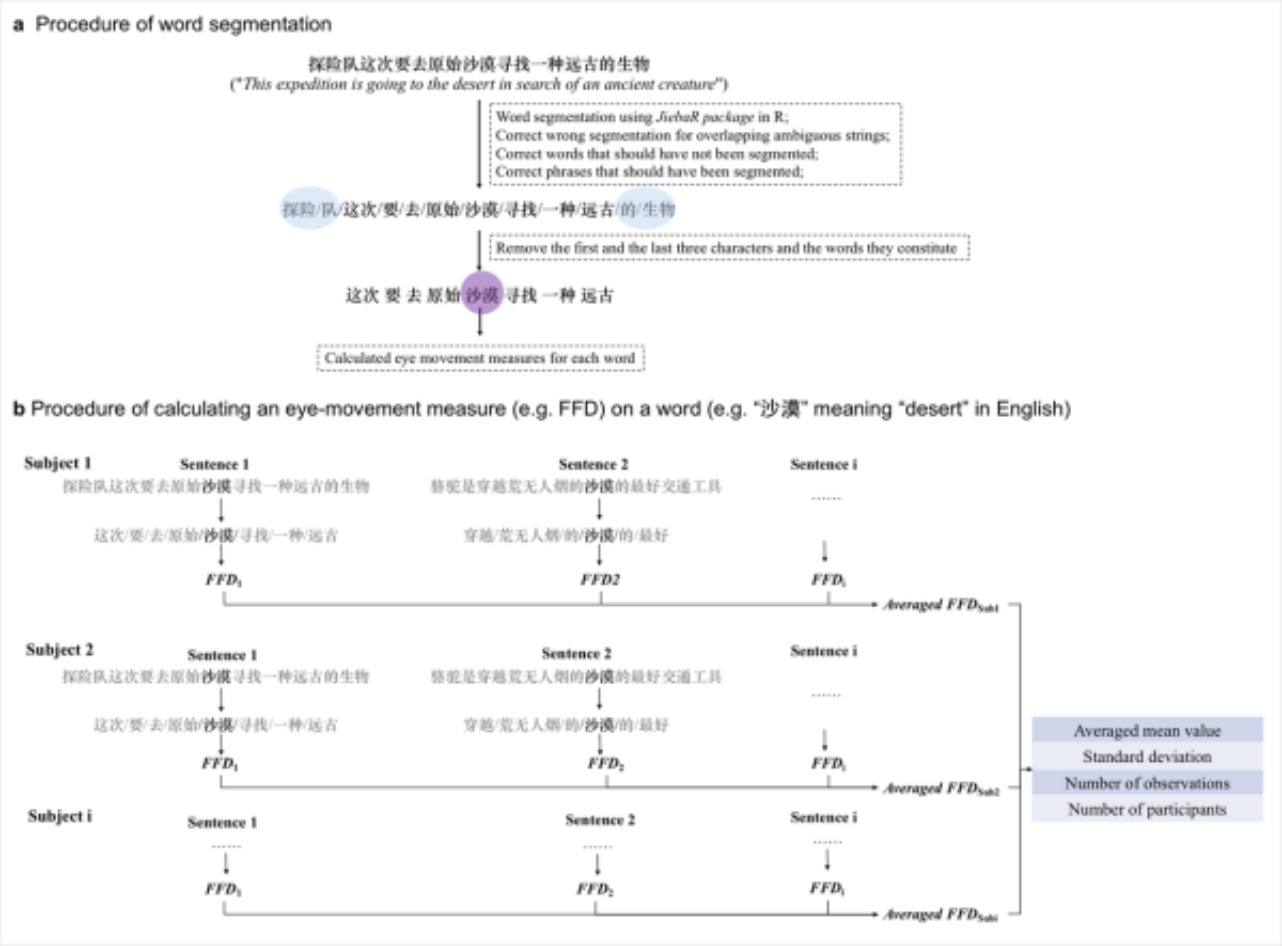


图1 词切分与眼动指标计算流程图。a.单个句子中的词切分流程；b.为单个词汇（如“ 沙漠 ”）上的某项眼动指标，如“ 首次注视时间 ”（First Fixation Duration, FFD）的计算流程。

研究团队单位：心理研究所

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