

阅读眼动追踪语料库的构建与应用研究综述

王晓明 赵歆波

西北工业大学计算机学院空天地海一体化大数据应用技术国家工程实验室 西安 710072

(wxmgo@163.com)



摘要 阅读文字时眼球的运动反映了人类的认知过程。阅读眼动数据是认知心理学、应用语言学、计算机科学等领域中重要的基础数据,而我国在阅读眼动研究的基础数据方面较欠缺。针对这一现状,首先介绍了阅读眼动追踪语料库产生的背景以及国内外的相关文献;然后从影响阅读眼动的低水平视觉因素和高水平视觉因素角度介绍了阅读眼动追踪语料库的内容及所使用的各项眼动指标,如单一注视时间、首次注视时间、凝视时间、总注视时间、回视出次数、回视入次数等,并分析了使用语料库研究法进行阅读眼动研究相比传统阅读眼动研究具有的3个优势;最后从语料库眼动指标变量、语料规模、语料内容、语料语种、被试规模、被试特征、采集设备等方面介绍了国外已经建成的较有影响力的若干阅读眼动追踪语料库,以供阅读眼动研究者参考。在眼动追踪语料库应用研究方面,对认知心理学、应用语言学和计算机科学等相关领域已开展的主要研究进行述评,重点介绍了在计算机科学的眼动可计算模型、自然语言处理、模式识别3个领域中基于阅读眼动追踪语料库开展的典型研究。在中文阅读眼动追踪语料库的构建与应用研究方面,介绍了我国相关研究的开展现状,分析了我国在眼动基础数据方面欠缺的原因,并从国家、科研机构、科研工作者3个层面提出了解决此问题的对策和建议。

关键词: 眼动追踪;眼动数据;阅读眼动;语料库;人工智能;计算语言学

中图法分类号 TP391

Survey of Construction and Application of Reading Eye-tracking Corpus

WANG Xiao-ming and ZHAO Xin-bo

National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China

Abstract The eye movements in reading are a reflection of the human cognitive process. Reading eye movement data is an important basic data in fields such as cognitive psychology, applied linguistics and computer science, while China is lack of the basic study data in this field. In view of this situation, this paper first introduced the background of reading eye-tracking corpus and the related literatures at home and abroad. Then, it presented the contents and indexes of eye movement in reading eye-tracking corpus including single fixation duration, the first fixation duration, gaze duration, total fixation duration, regression in count, regression out count from low level visual factors and high level visual factors, and analyzed three advantages of using corpus research method for reading eye movement compared to the traditional reading eye movement experiments. At last, some of influential and completed reading eye-tracking corpora were elaborated from the perspectives of the index variables, corpus size, corpus content, corpus language, scale of participants, characteristics of participants and data acquire equipment. It is expected to provide some reference for people who engage in reading eye movement. In the applied research of eye-tracking corpus, this paper reviewed the major researches in cognitive psychology, applied linguistics, computer science and related fields. Based on eye-tracking corpus, the representative studies carried out by computer science in eye movement computational model, natural language processing and pattern recognition were introduced with emphasis. Besides, the studies in eye tracking corpus construction and application in China were covered. This paper reviewed the current situation of relevant studies, analyzed the reasons for the lack of basic data, and proposed the solutions and suggestions from the point of view of the state, scientific research institutions and scientific workers respectively.

到稿日期:2019-08-05 返修日期:2019-11-12 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:国家自然科学基金(61231016,61871326);教育部人文社会科学研究一般项目(18YJCZH180);陕西省社会科学基金年度项目(2019M001)

This work was supported by the National Natural Science Foundation of China (61231016,61871326), Humanities and Social Science Fund of Ministry of Education of China (18YJCZH180) and Social Science Foundation of Shaanxi Province (2019M001).

通信作者:赵歆波(xbozhao@nwpu.edu.cn)

Keywords Eye-tracking, Eye movement data, Reading eye movement, Corpus, Artificial intelligence, Computational linguistics

1 引言

眼动追踪语料库(Eye-tracking Corpus)是以计算机为介质来承载语言知识的基础资料。语料库中存放的是在语言的实际使用中出现过的语言材料,以及被试阅读这些语言材料的过程中由眼动仪采集的眼动数据,这些语料和眼动数据经过加工后形成有用的眼动数据资源。眼动追踪语料库在阅读行为研究、语言处理研究、计算机科学等领域都有重要的应用价值。

眼动追踪语料库产生的背景与阅读行为研究有关。阅读是人类通过语言文字来获取信息、认识世界、发展思维并获得知识体验的活动,它是从视觉材料中获取信息的过程^[1]。20世纪80年代以后,阅读眼动研究像雨后春笋般涌现,相关的研究涉及语言单位的方方面面,例如字、词、句的识别,阅读的知觉广度,句子阅读,语篇阅读,第二语言习得等^[2],这些基于眼动实验的语言研究推动了有关认知过程理论的发展。传统的阅读眼动实验是为测试一个或多个特定的假设而精心设计的,其选择特定的被试,包含从数据采集到统计分析一整套冗长繁琐的实验操作^[3-5]。该实验对实验设置提出了较高的要求,实验者不仅需要购买或租用眼动仪等设备,还要学习掌握眼动仪的使用和数据统计分析方法;除此之外,还要选择阅读材料,组织被试者开展实验等。整个眼动实验的代价较大,且数据的准确性和客观性得不到保证。随着更复杂的阅读可计算模型的出现以及人工智能技术的飞速发展,人们对更大规模的自然状态下的文本阅读数据的需求变得更加迫切。在此背景下,阅读眼动追踪语料库应运而生。阅读眼动追踪语料库是一种由专业人士构建的多模态语料库,包含文字(语料)、文字的语言学属性,以及若干被试阅读这些文字时的眼动记录。

国外已经开展了眼动追踪语料库的相关研究,涉及眼动追踪语料库的构建以及应用等。在 Google Scholar 中,以“eye tracking data”为关键词,共搜到文献 607 篇;以“fixation data”为关键词,共搜到文献 253 篇;以“eye tracking corpus”或“eye tracking corpora”为关键词,共搜到文献 22 篇。可见,基于眼动数据开展的研究比较多,但把眼动数据作为语料资源或者基于眼动语料库进行的研究还较少。这种情况在国内

学界也同样存在,在 CNKI 中以“眼动数据”为关键词,共搜到文献 49 篇。但是,在 CNKI 中还未找到以“眼动语料库”或“眼动追踪语料库”为主题的文献。

基于国内外相关研究的现状,本文从语料库眼动指标变量、语料规模、语料内容、语料语种、被试规模、被试特征、采集设备等方面介绍已经建成的较有影响力的若干阅读眼动追踪语料库,并对认知心理学、应用语言学和计算机科学等相关领域已开展的典型应用研究进行述评。

2 阅读眼动追踪语料库的内容与优势

在阅读过程中,读者通常觉得眼睛是平滑地从左向右移动,但实际上,读者的眼睛运动轨迹由一系列的注视(Fixation)和眼跳(Saccade)组成^[6]。这种行为是由人类视网膜的生理结构特点所决定的,视网膜中央凹(Fovea)集中了绝大部分视神经细胞,视像只有落在这个区域才可以被人类“看清楚”;而不幸的是,中央凹仅相当于 5 度的视角^[7]。因此,读者需要通过连续的眼跳改变注视点,使下一步要注视的内容落在视网膜中央凹区域,以看清楚文本中的必要信息,进而理解阅读内容。眼动追踪帮助研究人员使用眼动仪以最小的干扰和代价确定被试者在阅读时关注某个单词的位置和次数,以及他们在阅读时从一个单词到另一个单词的眼动序列,提供一种良好的、实时的关于阅读过程中认知加工过程的一系列指标。图 1 给出成人读者阅读时的眼睛运动轨迹,黑点代表注视点(Fixation)位置,旁边的数字代表每个注视点的持续时间(单位为 ms),实线箭头指向眼跳(Saccade)方向,虚线箭头指向回视(Regression)方向。

在传统的阅读眼动研究中,研究者往往要根据研究的需要确定分析的目标区域,即“兴趣区”(Area of Interest, AOI)。兴趣区可以是一个单词,也可以是一个较大的区域,如一个短语或一个句子。研究者对只有一个字或词的较小兴趣区进行分析时,更关注语言特征对词汇通达的影响;对较大兴趣区进行分析时则可能更关注词语间的关系分析以及句法结构的分析^[8]。为了满足不同的研究需要,阅读眼动追踪语料库的“兴趣区”通常设置得较精细,至少是基于单词,更精细的甚至基于字符。当需要对较大兴趣区进行研究时,可以采用合并多个小兴趣区的方式。

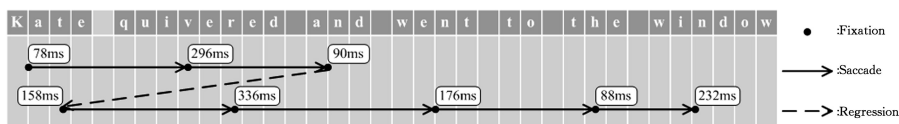


图 1 成人阅读者的眼动轨迹
Fig. 1 Eye track of adult reader

在阅读期间,眼睛运动由低水平视觉因素(如词长、笔画数等)和高水平视觉因素(或认知因素,如可预测性等)之间的复杂相互作用来控制^[6]。眼动追踪语料库中的指标变量就体现了这种双重影响因素。

阅读眼动追踪语料库的内容除了包括语料和文字的语言学信息(词性、词长等)之外,还包括被试者阅读语料时注视点在各个兴趣区上的记录(被试者的眼动指标通常有单一注视时间、首次注视时间、凝视时间、总注视时间、注视次数、回视

入次数、回视出次数、眼跳时间、眼跳距离、瞳孔直径等)。阅读眼动追踪语料库普遍采用以下眼动指标。

1)单一注视时间(Single Fixation Duration):在最初的从左到右的句子阅读中,兴趣区内有且只有一次注视时的注视时间。单一注视时间受词汇特征的影响较大,如词频、词长、词语的可预测性等^[8]。

2)首次注视时间(First Fixation Duration):被试第一次注视某兴趣区的持续时间。一般认为,首次注视时间代表早期的认知加工过程,是实验研究中经常使用的一个眼动指标^[9]。

3)凝视时间(Gaze Duration):被试第一次注视某兴趣区到离开该兴趣区之前的所有注视点的时间总和。凝视时间是反映初始加工的指标。

4)总注视时间(Total Fixation Duration):被试对某个兴趣区所有注视时间的总和,包括首次注视时间、凝视时间和注视点移动到其它区域后再返回该区域进行重新阅读的时间。总注视时间是一个晚期注视时间指标。

5)回视出次数(Regression Out Count):被试从注视某个兴趣区时开始,从该区域发生回视的次数。回视出次数反映了被试在该区域发现问题的情况。

6)回视入次数(Regression In Count):被试回视落入某个兴趣区的次数。语境预测性低的词有更多的回视入次数。

综上,眼动追踪语料库不仅选择了尽可能普通的语料和

尽可能广泛的被试群体,而且体现了阅读眼动的低水平视觉因素和高水平视觉因素。因此,使用阅读眼动追踪语料库开展研究至少有 3 方面优势。

1)研究样本更客观。传统实验设置中常常选择一组特定的被试,并将被试结果数据集限定在关键维度上,这样仅能显示语言行为的一部分;而眼动追踪语料库基于对大量被试者的观察,数据的统计学意义更具客观性。

2)结果数据更准确。在涉及连续词汇变量的研究方面,如词频的影响等,该语料库可以从更广的尺度进行评估计算。此外,由专业人员来做数据清洗、数据异常值和无效值的处理,会使结果数据更加准确。

3)研究代价更小。传统的研究方法使得设计新的眼动实验相当耗时,并且需要昂贵的眼动仪设备;而基于语料库的研究无需研究者设计新的实验和收集新的数据就能够验证多个假设,并定量分析语言行为和评价语言模型,为验证、修正和反驳现有理论假设提供有力的证据^[10]。

3 国外重要的眼动追踪语料库

国外已经建成的较有影响力的阅读眼动追踪语料库一般都包括语料内容、语料词长、注视持续时间、注视序列、注视位置等信息,部分语料库有语料词性、可预测性统计等信息。表1列出国外已建成的较有影响力的眼动追踪语料库的基本情况。不同语料库在语种、语料、被试、设备的选择上不尽相同。

表 1 国外已建成的较有影响力的眼动追踪语料库的基本情况
Table 1 Basic situation of more influential eye tracking corpus built abroad

语料库名称	语料语种	发布机构	发布时间/年	被试规模/人	语料内容	语料规模	眼动仪采样频率/Hz
邓迪(Dundee)	英语、法语	英国邓迪大学	2003	20	英文《独立报》和法文《世界报》	2379 个英语句子(56212 个英语单词)	1000
波茨坦(Potsdam)	德语	德国波茨坦大学	2004	33	德语短句	144 个句子(1138 个单词)	250
波茨坦-阿拉哈巴德印地语语料库(Potsdam-Allahabad Hindi Eye Tracking Corpus)	印地语	印度理工学院	2015	30	《Hindi-Urdu 树库》中涉及电影、娱乐和体育等主题	153 个句子(2610 个单词)	500
普罗沃(Provo)	英语	美国杨百翰大学	2017	84	在线新闻、科普杂志和通俗小说	55 篇短文(2689 个单词)	1000
盖科(GECO)	英语、荷兰语	比利时根特大学	2017	33	侦探小说《泰尔斯庄园奇案》	54364 个英语单词和59716 个荷兰语单词	1000
祖科(ZuCo)	英语	瑞士苏黎世联邦理工学院	2018	12	《斯坦福情感树库》和《维基百科关系抽取语料库》	1107 个句子(21629 个单词)	500

邓迪(Dundee)语料库^[16]是一个英、法双语眼动语料库,是英国邓迪大学的一项研究成果,由 Alan Kennedy 等于 2003 年发布。它是世界上最著名的眼动追踪语料库,目前已被公认为心理语言学中的重要资源^[12]。邓迪语料库包含了 10 位以英语为母语的被试以及 10 位以法语为母语的被试在阅读文本时的眼动数据。阅读材料为英文 *The Independent* 和法文 *Le Monde* 上的文章,分别包含 20 篇文章,每篇文章都由所有被试者阅读。英语语料库由 2379 个句子、56212 个单词组成。实验使用 Dr. Bouis Oculometer 眼动仪记录阅读眼动数据,对右眼进行采样(采样频率为 1000 Hz)。

Dundee Corpus 的英文部分在 2009 年被标注了词性

(Part-of-speech, POS)信息^[13]。这一层标注促进了新的心理语言学研究,例如使用分层短语结构和顺序结构模型测试读者模型的研究。Barrett 等^[14]于 2015 年发布了一项名为“The Dundee Treebank”的资源。该研究在 POS 注释上添加一层依赖语法,通过用手动指定的语法替换自动解析来促进该领域的进一步研究。

波茨坦(Potsdam)语料库^[15]是一个德语眼动语料库,由德国波茨坦大学的 Kliegl 等于 2004 年发布。该眼动语料库除了包含被试的阅读眼动数据外,还包括语料的可预测性规范——使用完形填空的方式获取了 272 名德国母语者提供的对缺失单词预测的准确性统计。进行可预测性统计的被试者

中包括 116 名高中生(17~19 岁)、76 名大学生(19~38 岁)、80 名老年人(66~80 岁)。阅读眼动数据采集自 33 位大学生(平均 21.9 岁,年龄区间为 19~28 岁)和 32 位年龄较大的成人(平均 69.9 岁,年龄区间为 65~83 岁)。阅读材料是作者精心设计的包含各种语法结构的德语短句,一共有 144 个句子(1138 个单词),每个句长 5~11 个词不等,平均每句 7.9 个单词。实验所用的仪器是 SR EyeLink System(SMI)眼动仪,采样频率为 250 Hz。

波茨坦-阿拉哈巴德印地语语料库(Potsdam-Allahabad Hindi Eye Tracking Corpus)^[16]是一个印地语眼动语料库,由印度理工学院的 Husain、德国波茨坦大学的 Vasishth 和印度阿拉哈巴德大学的 Srinivasan 于 2015 年发布。被试包括阿拉哈巴德大学的 30 名研究生和本科生,他们均能够流利地阅读乌尔都语和印地语。阅读材料选自于 Hindi-Urdu tree-bank^[17]中涉及电影、娱乐和体育等主题的 153 个句子(2 610 个单词)。所选择的语料都不是孤立的句子,而是由几个句子组成的简短叙述。实验采用 SMI iView X HED 眼动仪,以 500 Hz 的频率进行采样。

普罗沃(Provo)语料库^[18]是一个英语眼动语料库,由美国杨百翰大学的 Luke 等于 2017 年发布。该眼动语料库由两部分组成:可预测性规范和阅读眼动数据。可预测性规范的数据来自杨百翰大学的 470 名被试(包括 267 名女性和 203 名男性),其年龄为 18~50 岁(平均 21 岁)。被试阅读了包括在线新闻、科普杂志和通俗小说等题材在内的 55 篇短文。这些短文的平均长度为 50 个单词(范围为 39~62),平均包含 2.5 个句子(范围为 1~5),每个句子的平均长度为 13.3 个单词(范围为 3~52)。所有的被试均阅读了共含 2 689 个单词的 55 篇短文,最终形成了语料的可预测性规范——对缺失单词预测的准确性统计。阅读眼动数据来自杨百翰大学的另外 84 名被试,所有被试均为英语母语者,他们均未参加过可预测性规范数据的收集过程。眼动数据采集所用的设备是 SR Research EyeLink 1000 Plus 眼动仪,以 1 000 Hz 采样频率记录右眼眼动数据。

盖科(Ghent Eye-Tracking Corpus, GECO)语料库^[19]是一个英、荷双语眼动语料库,由比利时根特大学的 Cop 等于 2017 年发布。它的最大特点是:双语读者阅读一部完整小说的眼动追踪数据。参与构建该眼动语料库的被试包括 19 名来自根特大学的英、荷双语者和 14 名来自南安普敦大学的英语单语者。双语者的平均年龄为 21.2 岁(范围为 18~24),单语者的平均年龄为 21.8 岁(范围为 18~36)。阅读材料为英国女作家阿加莎·克里斯蒂(Agatha Christie, 1890—1976 年)的侦探小说《泰尔斯庄园奇案》(英语为 *The Mysterious Affair at Styles*, 荷兰语为 *De zaak Styles*),整部小说共 13 章。双语者的母语为荷兰语,第二语言为英语。双语者被分为两组,第一组的 10 人用母语阅读 1—7 章,用第二语言阅读 8—13 章;第二组的 9 人用第二语言阅读 1—7 章,用母语阅读 8—13 章。最终收集了 59 716 个荷兰语单词和 54 364 个英语单词的阅读眼动数据。实验使用塔式安装的 EyeLink

1 000 系统记录眼动数据,采样频率为 1 000 Hz。

祖科(Zurich Cognitive Language Processing Corpus, Zu-Co)语料库^[20]是一个英语眼动语料库,由瑞士苏黎世联邦理工学院的 Hollenstein 等于 2018 年发布。这是一个结合脑电图(Electroencephalography, EEG)和阅读自然语句的被试的眼动追踪记录的数据集,包括 12 名成年母语英语人士的高密度脑电图和眼动追踪数据,这些被试每人阅读 4~6h 的英语文本,这些文本包括 1 107 个句子(21 629 个单词)。实验设置了两项正常阅读任务和一项特定阅读任务,记录被试阅读给定语料时的脑电和眼动追踪数据,形成一个数据集。作者认为这个数据集可以成为自然语言处理(Natural Language Processing, NLP)的重要资源。

除此之外,还有一些其他的眼动追踪语料库。例如,在视觉词识别领域中有影响力的心理语言语料库——由 Balota 等在 2007 年开展的英语词典项目(English Lexicon Project, ELP)^[21],收集了 816 名被试者阅读词汇的眼动数据。此外, Ferrand 等在 2010 年开展了法语相关研究(French Lexicon Project, FLP)^[22];Keuleers 等在 2010 年开展了荷兰语相关研究(Dutch Lexicon Project, DLP)^[23];Keuleers 等在 2012 年开展了英式英语相关研究(British Lexicon Project, BLP)^[24]。

4 眼动追踪语料库的应用

多年以来,眼动追踪语料库在认知心理学、应用语言学、计算机科学领域得到了广泛的应用,并推动了语言认知过程和自然语言处理相关理论的发展。

4.1 眼动追踪语料库在认知心理学的应用

在认知心理学方面,随着心理学研究理念和研究方法的变化,实证和多学科合作正成为心理学研究发展的两大趋势^[25]。眼动法是进行阅读研究时使用较多的一种实验方法,这种方法将材料全部呈现在计算机屏幕上,被试只需要正常阅读句子即可,眼动仪和计算机连接在一起,能准确地追踪和记录被试的眼动轨迹^[26]。这种方法比较接近于自然条件下的阅读,可以最大限度地使读者的阅读不受干扰地顺畅进行。通过这种方法,研究者可以分析和研究被试阅读过程中对语言的认知加工过程^[27]。大量的研究表明,人类眼动行为存在与假设的随机分布不同的重尾分布特征,眼动可以反映视觉信息的选择模式,具有一定的规律性,而这些规律有助于揭示人类对外界事物认知加工过程的心理机制^[9]。眼动语料库已被用于评价各种心理学假设和理论,如词频效应^[28]、字长效应^[29]、邻域效应^[30]、副中央凹-中央凹效应^[31]、文章标记效应^[32]、顺序效应^[33]、词汇决策任务^[34]、评估单词识别的复杂计算模型^[35]、词频和可预测性之间的相互作用^[36]。这些研究阐述了眼动追踪语料库可用于研究不同处理粒度(如单词和句子级别)的变量以及它们之间的相互作用,并从时间维度解读一些特定的阅读效应。

4.2 眼动追踪语料库在应用语言学的应用

在应用语言学方面,眼动追踪语料库被用来评价各种语言模型。使用眼动跟踪语料库测试语言模型,获得的是对部

分或整个特定模型的普遍性测试,测试结果更接近自然状态。Demberg 等^[37]用 Dundee 语料库的数据研究了主语/宾语句子的不对称性,并受这些结果的启发,建立了句法处理模型。Mitchell 等^[38]使用 Dundee 眼动语料库来研究句子阅读中的预测。Frank 等^[39]以及 Fossum 等^[40]都使用 Dundee 眼动追踪语料库来评估他们的语言模型,并研究了语言模型与句子处理中分层机制作用的关系。Kuperman 等^[41]使用 ELP 的元数据和 Dundee 语料库研究了词汇决策时间与阅读眼动数据的关联机制。

4.3 眼动追踪语料库在计算机科学的应用

在计算机科学方面,眼动追踪语料库作为一种数据资源,主要被应用在以数据科学和统计科学为基础的人工智能领域。

1)眼动可计算模型。为了描述阅读时的眼球运动规律,研究者提出了各种眼动模型,如 E-Z 读者模型^[42]等。早期,眼动模型的各个参数拟合了经验数据集,它们对眼动行为的建模通常基于被试群体中的平均值。然而,除了阅读材料本身的特性会影响读者在阅读过程中的眼跳模式外,读者自身的特性也会对眼跳模式产生一定的影响。其中,读者的阅读水平是一个很重要的影响因素,具体表现为:相较于初学者和阅读困难者,熟练阅读者的眼跳模式更有规律性^[43-45],即眼动模式在个体之间是显著不同的。因此,阅读的个体化差异问题成为了眼动模型构建领域中一个富有挑战性的研究方向。一些学者从人工智能的角度研究了眼动模式,与之前的研究相比,这些研究较少地与心理学假设联系在一起,其使用阅读过程中读者的眼球运动数据作为训练数据来预测该读者阅读其他文本时的注视点。Nilsson 等^[46]第一次提出用统计学习方法来进行阅读眼动研究,使用逻辑回归分类器和非常有限的一组特征对未见文本进行阅读注视词预测。实验表明,在预测结果可以超过多数基线,并能够证实读者阅读眼动的词频效应。随后,Nilsson 等^[47]又提出了一种基于数据驱动的阅读眼动控制模型。在该模型中,发动眼跳计划的过程被看作一个在线处理函数,眼跳位置由从眼动追踪语料库中获取的训练数据对模型进行训练而得到。Hara 等^[48]使用 Dundee 眼动追踪语料库的数据训练条件随机场模型,以预测每个单词是否被被试注视。实验结果表明,在使用词汇特征和屏幕位置特征的情况下,该模型对于每个被试具有 73%~84% 的预测准确度,预测注视词分布和实际注视词分布之间的总相似度为 0.9462,这说明了从眼动追踪数据捕获一般阅读策略的可能性。Matties 等^[49]使用线性条件随机场模型对跨读者阅读眼动进行建模,得到了逻辑回归更好的效果。Wang 等^[50-52]提出了一种利用深度神经网络预测阅读注视词的眼动模型,该模型将 Provo 眼动追踪料库中读者的部分眼动数据作为训练数据来预测该读者阅读其他文本时的注视词。实验结果表明,该模型使用较少的特征就能够获得较高的准确率。

2)自然语言处理。为了训练自然语言处理程序,通常需要大型标记数据集。例如,若要训练一个情感计算系统来预

测一个句子的情感极性(即积极、消极、中立),则需要大量被标记的句子。通常,人们必须阅读这些训练句子,并为每个句子指定一种情感,这显然会非常耗时。一些研究以人类阅读时的生理活动数据(眼动或脑电)来代替这项劳动密集型的任务。也就是说,可以直接从生理活动的原始数据中发现并提取文本理解和注释的相关信息。举例来说,意见和情绪来自阅读文本的读者,这反映在他们的大脑活动中,会影响他们的阅读行为。因此,可以利用机器学习技术从记录的生理活动数据中解码这些信息,并用其替代或补充繁琐的人工标注。先前的研究已经证明眼动信息优化了自然语言处理任务,例如词性标注^[53]、情感分析^[54]和词嵌入评估^[55]。一项典型的研究是 Hollenstein 等^[48]使用 ZuKo 眼动追踪语料库对不同的自然语言处理任务进行实验,包括文本挖掘信息的提取、实体和关系的发现,以及语义任务等。

3)模式识别。生物特征识别技术已被广泛运用到电子商务、电子产品、网络安全等领域。眼动作为人类的一种行为特征,具有随身性、安全性、广泛性、可采集性等特点^[56],因此,可以通过对比人类的眼动轨迹特征来达到识别用户的目的。Kasproski 等^[57]发表的研究是第一个将眼球运动作为生物识别的研究,他们在 9 个被试的眼动追踪语料库上借用语音识别中常用的技术进行眼动生物识别,获得了平均 1% 的假正例率(False Positive Rate,FPR)和 23% 的假反例率(False Negative Rate,FNR)。Holland 等^[58]研究复杂眼球运动模式(Complex Eye Movement Patterns,CEM-P),在 32 名被试者的眼动追踪语料库上获得了 27% 的等误率(Equal Error Rate,EER)。Rigas 等^[59]将基于图形的匹配技术(类似于人脸识别)应用于眼球运动位置匹配,在一个由 15 个被试者组成的眼动追踪语料库上,所提出的技术实现了 30% 的等误差率。Cantoni 等^[60]提出了一种被称为 GANT 的注视分析技术,对于眼动追踪语料库中的每个被试,使用不同点的注视时间和回视次数构建眼动模型,并利用密度图的 Frobenius 范数寻找两个记录之间的相似性。

5 中文阅读眼动追踪语料库的现状分析

中国心理学会理事长、天津师范大学教育部人文社科重点研究基地主任白学军教授^[61]于 2017 年指出我国眼动研究在基础数据方面较为欠缺,如语言研究方面缺乏基础的语料库,亟待完善。

北京师范大学的 Yan 和德国波茨坦大学的 Kliegl 等^[62]于 2010 年构建了一个中文阅读眼动语料库,来自北京师范大学的 30 名中文母语的学生参与了眼动追踪语料库的构建。被试阅读的阅读材料是 Beijing Sentence Corpus(BSC)^[63]中的 150 个句子,这些句子选自《人民日报》,实验所使用的设备为 EyeLink II system,采样频率为 500 Hz。这是目前已知的中文阅读眼动追踪语料库,但该语料库未被公开获取;而基于中文阅读眼动追踪语料库的研究更是少之又少。

综上,我国在阅读眼动追踪语料库研究方面与国外还有明显差距,本文认为有如下 4 方面原因。

1)我国阅读眼动研究起步较晚。国外早在1879年就有关于阅读眼动方面的研究成果发表,而我国直到20世纪80年代后才有学者开始从事阅读眼动研究。

2)阅读眼动语料库的相关学科缺乏交叉融合平台。阅读眼动语料库涵盖心理学、语言学、计算机科学等学科,而我国心理学学科主要集中在师范类高校,语言学主要集中在外语类院校,计算机科学学科主要集中在工科类院校,总体上缺乏相关学科之间交叉融合的平台。

3)研究者缺乏构建数据资源的动力。我国科研评价体制导致研究者普遍重论文、专著成果,轻数据资源建设成果。数据资源的建设耗时耗力,加上我国知识产权保护意识不够,研究者的工作难以获得直接回报。

4)中文的特殊性导致不能直接使用基于拼音文字的阅读眼动资源。中文书写相比拼音文字存在一定的特殊性,如中文词与词之间不存在空格,因而不能直接使用基于拼音文字的阅读眼动追踪语料库进行中文阅读的研究,这直接导致了此类应用研究较缺乏。

基于以上分析,针对我国阅读眼动追踪语料库研究较为缺乏的现状,本文提出如下对策与建议。

1)国家应从战略角度重视语言基础数据的建设,加大数据资源建设投入和知识产权保护的力度,构建多元化的科研成果评价机制;特别是在人工智能技术飞速发展的今天,形成了“数据为王”的局面,语言基础数据的建设将直接影响国家科技实力的整体水平。

2)各科研机构 and 高校要重视相关学科的交叉研究,积极落实国家出台的科研成果多元化评价政策以及各项知识产权保护政策,积极引导研究人员开展语言数据资源相关的交叉学科研究。

3)研究者要重视阅读眼动追踪语料库构建的工程属性。语料库建设是一个平台工程,尽管其宗旨是服务于心理学、语言学等研究,但作为计算机化的工程项目,它具有软件的众多特质;特别是阅读眼动追踪语料库的建设更是对研究者的综合知识提出了较高的要求,研究者应充分借鉴软件工程、测试学、统计学现有的成熟技术,做好语料采集、文本加工标注、数据降噪等处理工作,以提升语料库的应用价值。

随着国家对科研投入力度的加大和各领域对阅读眼动研究的深入,阅读眼动基础数据问题已引起了学界的广泛重视,有关学者已经对此立项研究,如教育部社科项目《中文阅读眼动追踪语料库构建及眼跳目标动态预测研究》等。

结束语 本文是国内首个介绍阅读眼动追踪语料库的综述性研究,通过对现有的眼动追踪语料以及在其基础上开展的应用进行分析和介绍,不难得出眼动追踪语料库的以下特性:1)非常适合研究语言的广泛性现象——只要某些句法结构或具有某些词汇特征的单词在语料库中足够频繁地出现,就可以使用眼动语料库开展研究;2)具有显著的无限限制性数据——由于数据的采集并非基于特定的目的,因而得出的研究结论更具普遍性;3)为研究人员提供了廉价、易获取的数据,无需耗时和使用昂贵的设备去不断设计新实验或收集新

数据。我国在阅读眼动基础数据方面较欠缺,在语言研究方面缺乏基础的中文阅读眼动语料库,语言处理研究与国外还有明显差距,这个问题已引起了学界的广泛重视,相信随着国家对科研投入力度的加大和各领域对阅读眼动研究的深入,这个问题不久将会得到解决。

参 考 文 献

- [1] LI J Y, PU Y T. A Review of Domestic Studies on L2 Reading Strategies in the Past Decade [J]. Foreign Language Education, 2017, 38(3): 62-67.
- [2] RAYNER K. Eye movements in reading and information processing: 20 years of research [J]. Psychological Bulletin, 1998, 124(3): 372-422.
- [3] COLTHEART M, RASTLE K, PERRY C, et al. DRC: A dual route cascaded model of visual word recognition and reading aloud [J]. Psychological Review, 2001, 108(1): 204-256.
- [4] DILKINA K, MCCLELLAND J L, PLAUT D C. Are there mental lexicons? The role of semantics in lexical decision [J]. Brain Research, 2010, 1365(3): 66-81.
- [5] HARM M W, SEIDENBERG M S. Computing the Meanings of Words in Reading: Cooperative Division of Labor Between Visual and Phonological Processes [J]. Psychological Review, 2004, 111(3): 662-720.
- [6] MENG H X. The selection mechanism of saccade target in Chinese reading [M]. Guangzhou: World Book Publishing Guangdong Co., Ltd., 2016: 30-108.
- [7] LIU X L. Visual neurophysiology [M]. Beijing: People's Medical Publishing House, 2011: 1-42.
- [8] YAN G L, XIONG J P, ZANG C L, et al. Review of Eye-movement Measures in Reading Research [J]. Advances in Psychological Science, 2013, 21(4): 589-605.
- [9] WANG F J, TIAN M, HUANG Y P, et al. Classification Model of Visual Attention Based on Eye Movement Data [J]. Computer Science, 2016, 43(1): 85-88, 115.
- [10] MOUSIKOU P, SADAT J, LUCAS R, et al. Moving beyond the monosyllable in models of skilled reading: Mega-study of disyllabic nonword reading [J]. Journal of Memory and Language, 2017, 93(3): 169-192.
- [11] KENNEDY A, HILL R, PYNTE J. The Dundee Corpus [C]// Proceedings of the 12th European Conference on Eye Movement. Dundee, Scotland: Elsevier, 2003: 13-23.
- [12] University of Dundee. Dundee Corpus update could help reveal the secrets of reading [EB/OL]. (2018-04-13). <https://www.dundee.ac.uk/social-sciences/news/2018/article/dundee-corpus-update-could-help-reveal-the-secrets-of-reading.php>.
- [13] FRANK S. Surprisal-based comparison between a symbolic and a connectionist model of sentence processing [J]. Proceedings of the annual meeting of the Cognitive Science Society, 2009, 31(31): 1139-1144.
- [14] BARRETT M, AGIĆ Ž, SØGAARD A. The Dundee Treebank [C]// Proceedings of the 14th International Workshop on Tree-

- banks and Linguistic Theories. 2015.
- [15] KLIEGL R, GRABNER E, ROLFS M, et al. Length, frequency, and predictability effects of words on eye movements in reading [J]. *European Journal of Cognitive Psychology*, 2004, 16(1/2): 262-284.
 - [16] HUSAIN S, VASISHTH S, SRINIVASAN N. Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus [J]. *Journal of Eye Movement Research*, 2014, 8(2): 1-13.
 - [17] BHATT R, NARASIMHAN B, PALMER M, et al. A multi-representational and multi-layered treebank for hindi/urdu [C]//*Proceedings of the Third Linguistic Annotation Workshop*. Association for Computational Linguistics, 2009: 186-189.
 - [18] LUKE S G, CHRISTIANSON K. The Provo Corpus: A large eye-tracking corpus with predictability norms [J]. *Behavior Research Methods*, 2018, 50(2): 826-833.
 - [19] COP U, DIRIX N, DRIEGHE D, et al. Presenting GECO: An eye tracking corpus of monolingual and bilingual sentence reading [J]. *Behavior Research Methods*, 2016, 49(2): 1-14.
 - [20] HOLLENSTEIN N, ROTSZTEJN J, TROENDLE M, et al. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading [J]. *Scientific Data*, 2018, 5: 180291-180291.
 - [21] BALOTA D A, YAP M J, HUTCHISON K A, et al. The English lexicon project [J]. *Behavior Research Methods*, 2007, 39(3): 445-459.
 - [22] FERRAND L, NEW B, BRYLSBAERT M, et al. The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords [J]. *Behavior Research Methods*, 2010, 42(2): 488-496.
 - [23] KEULEERS E, DIEPENDAELE K, BRYLSBAERT M. Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords [J]. *Frontiers in Psychology*, 2010, 13(1): 174-183.
 - [24] KEULEERS E, LACEY P, RASTLE K, et al. The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words [J]. *Behavior research methods*, 2012, 44(1): 287-304.
 - [25] SHI F. The Macrohistory, Mesohistory and Microhistory of Evolutionary Linguistics [J]. *Nankai Journal (Philosophy, Literature and Social Science Edition)*, 2018, 33(4): 65-71.
 - [26] GUO X Y, LI L, GENG H J. Eye-movement Analysis of Visual Similarity Perception on Synthesized Texture Images [J]. *Computer Science*, 2018, 45(8): 223-228.
 - [27] YU M. Eye Movement Research on Syntactic Ambiguity Processing in Modern Chinese [M]. Tianjin: Nankai University Press, 2014: 21-25.
 - [28] KUPERMAN V, VAN DYKE J A. Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers [J]. *Journal of Experimental Psychology: Human Perception and Performance*, 2013, 39(3): 802-813.
 - [29] YAP M J, BALOTA D A. Visual word recognition of multisyllabic words [J]. *Journal of Memory and Language*, 2009, 60(4): 502-529.
 - [30] WHITNEY C. Location, location, location: How it affects the neighborhood (effect) [J]. *Brain and Language*, 2011, 118(3): 90-104.
 - [31] KENNEDY A, PYNTE J. Parafoveal-on-foveal effects in normal reading [J]. *Vision Research*, 2005, 45(2): 153-168.
 - [32] PYNTE J, KENNEDY A. The influence of punctuation and word class on distributed processing in normal reading [J]. *Vision Research*, 2007, 47(9): 1215-1227.
 - [33] KENNEDY A, PYNTE J. The consequences of violations to reading order: An eye movement analysis [J]. *Vision Research*, 2008, 48(21): 2309-2320.
 - [34] DIEPENDAELE K, BRYLSBAERT M, NERI P. How noisy is lexical decision? [J]. *Frontiers in psychology*, 2012, 13(3): 348-353.
 - [35] NORRIS D, KINOSHITA S. Orthographic processing is universal; it's what you do with it that's different [J]. *Behavioral and Brain Sciences*, 2012, 35(5): 296-297.
 - [36] KENNEDY A, PYNTE J, MURRAY W S, et al. Frequency and predictability effects in the Dundee Corpus: An eye movement analysis [J]. *The Quarterly Journal of Experimental Psychology*, 2013, 66(3): 601-618.
 - [37] DEMBERG V, KELLER F. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity [J]. *Cognition*, 2008, 109(2): 193-210.
 - [38] MITCHELL J, LAPATA M, DEMBERG V, et al. Syntactic and semantic factors in processing difficulty: An integrated measure [C]//*Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010: 196-206.
 - [39] FRANK S L, BOD R. Insensitivity of the human sentence-processing system to hierarchical structure [J]. *Psychological Science*, 2011, 22(6): 829-834.
 - [40] FOSSUM V, LEVY R. Sequential vs. hierarchical syntactic models of human incremental sentence processing [C]//*Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, 2012: 61-69.
 - [41] KUPERMAN V, DRIEGHE D, KEULEERS E, et al. How strongly do word reading times and lexical decision times correlate? Combining data from eye movement corpora and megastudies [J]. *The Quarterly Journal of Experimental Psychology*, 2013, 66(3): 563-580.
 - [42] REICHLE E D, RAYNER K, POLLATSEK A. The E-Z reader model of eye-movement control in reading: comparisons to other models [J]. *Behavioral & Brain Sciences*, 2003, 26(4): 477-526.
 - [43] DUCROT S, LÉTÉ B, SPRENGER-CHAROLLES L, et al. The optimal viewing position effect in beginning and dyslexic readers [J]. *Current Psychology Letters, Behaviour, Brain & Cognition*, 2003, 33(10): 23-33.
 - [44] JOSEPH H S S L, LIVERSEGE S P, BLYTHE H I, et al. Word length and landing position effects during reading in children and adults [J]. *Vision Research*, 2009, 49(16): 2078-2086.
 - [45] VITU F, MCCONKIE G W, KERR P, et al. Fixation location

- effects on fixation durations during reading: An inverted optimal viewing position effect[J]. *Vision Research*, 2001, 41 (25/26): 3513-3533.
- [46] NILSSON M, NIVRE J. Learning where to look: modeling eye movements in reading[C]// *Thirteenth Conference on Computational Natural Language Learning*. Boulder, Colorado: Association for Computational Linguistics, 2009: 93-101.
- [47] NILSSON M, NIVRE J. Towards a data-driven model of eye movement control in reading[C]// *The Workshop on Cognitive Modeling & Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, 2010: 63-71.
- [48] HARA T, MOCHIIHASHI D, KANO Y, et al. Predicting Word Fixations in Text with a CRF Model for Capturing General Reading Strategies among Readers[C]// *The Workshop on Eye-Tracking & Natural Language Processing*. Mumbai: Association for Computational Linguistics, 2012: 55-70.
- [49] MATTIES F, SØGAARD A. With blinkers on: robust prediction of eye movements across readers[C]// *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington: Association for Computational Linguistics, 2013: 803-807.
- [50] WANG X, ZHAO X, XIA M. The Prediction Model of Saccade Target Based on LSTM-CRF for Chinese Reading[C]// *International Conference on Brain Inspired Cognitive Systems*. Cham: Springer, 2018: 44-53.
- [51] WANG X M, ZHAO X B, REN J C. A New Type of Eye Movement Model Based on Recurrent Neural Networks for Simulating the Gaze Behavior of Human Reading [J]. *Complexity*, 2019, 2019: 1-12.
- [52] WANG X M, ZHAO X B. Eye movement prediction of individuals while reading based on deep neural networks [J]. *Journal of Tsinghua University (Science and Technology)*, 2019, 59(6): 468-475.
- [53] BARRETT M, BINGEL J, KELLER F, et al. Weakly supervised part-of-speech tagging using eye-tracking data [C] // *54th Annual Meeting of the Association for Computational Linguistics*, (ACL 2016). 2016: 579-584.
- [54] MISHRA A, KANOJIA D, NAGAR S, et al. Leveraging Cognitive Features for Sentiment Analysis[C]// *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. 2016.
- [55] SØGAARD A. Evaluating word embeddings with fMRI and eye-tracking[C]// *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. 2016: 116-121.
- [56] WANG Z M, ZHANG S, HE Y. Selective Ensemble Learning Human Activity Recognition Model Based on Diversity Measurement Cluster[J]. *Computer Science*. 2018, 45(1): 307-312.
- [57] KASPROWSKI P, OBER J. Eye movements in biometrics[C]// *International Workshop on Biometric Authentication*. Berlin: Springer, 2004: 248-258.
- [58] HOLLAND C, KOMOGORTSEV O V. Biometric identification via eye movement scanpaths in reading[C]// *2011 International Joint Conference on Biometrics (IJCB)*. IEEE, 2011: 1-8.
- [59] RIGAS I, ECONOMOU G, FOTOPOULOS S. Biometric identification based on the eye movements and graph matching techniques[J]. *Pattern Recognition Letters*, 2012, 33(6): 786-792.
- [60] CANTONI V, GALDI C, NAPPI M, et al. GANT: Gaze analysis technique for human identification [J]. *Pattern Recognition*, 2015, 48(4): 1027-1038.
- [61] Laboratory of Bilingual Cognition and Development of Guangdong University of Foreign Studies. Professor Bai Xue-jun talks about "Several Basic Characteristics of Chinese Eye Movement Research"[EB/OL]. (2017-02-26) [2018-12-03]. <http://bcd-lab.gdufs.edu.cn/info/1017/1463.htm>.
- [62] YAN M, KLIEGL R, RICHTER E M, et al. Flexible saccade-target selection in Chinese reading [J]. *The Quarterly Journal of Experimental Psychology*, 2010, 63(4): 705-725.
- [63] YU B, ZHANG W, JING Q, et al. STM capacity for Chinese and English language materials [J]. *Memory & Cognition*, 1985, 13(3): 202-207.



WANG Xiao-ming, born in 1982, Ph.D, lecturer, is member of China Computer Federation (CCF). His main research interests include reading eye movement modeling and artificial intelligence.



ZHAO Xin-bo, born in 1970, Ph.D, professor, Ph.D supervisor. His main research interests include image processing, computer vision, pattern recognition and artificial intelligence.