心理科学进展 2011, Vol. 19, No. 4, 459–470
Advances in Psychological Science DOI: 10.3724/SP.J.1042.2011.00459

·主编特邀(Editor-In-Chief Invited)·

编者按：

作为最重要的语言单位之一，词在语言处理的过程中具有非常重要的作用。然而，与字母文字不同，中文词和词之间没有空格分隔。因此中文读者在阅读时就需要采用一定的机制将词切分开来。没有空格的条件下，中文读者是如何进行词的切分的？这个问题长期以来没有引起人们足够的重视。李兴珊博士在这个领域做了一系列实验研究和建模研究，并取得了一些初步成果。本文系统地综述了近期在中文词切分领域的一些成果，并提出了研究展望。希望这个综述可以增进我们对中文阅读的认知机理的理解，也能引起大家对中文词切分认知机理研究的重视。另外，本文也综述了一些关于词的心理现实性的研究结果，相信这些结果对语言学界认识词在语言理解中的作用起到一些借鉴作用。

(本文责任编辑：杨玉芳)

中文阅读中词切分的认知机理述评*

李兴珊[1] 刘萍萍[1,2] 马国杰[1,2]

([1]中国科学院心理研究所行为科学重点实验室，北京 100101)
([2]中国科学院研究生院， 北京 100049)

摘 要 大量的认知科学研究表明，词在阅读认知加工过程中起着非常重要的作用。因此在阅读中一个重要的过程就是把词从文本中切分出来。有别于英文等拼音文字，中文文本的词之间没有空格分隔。在没有空格辅助的条件下，中文阅读者是如何进行词的切分呢？本文主要综述了近期认知心理学和计算机科学领域针对该问题的研究进展，主要包括：1)介绍了一些词作为整体进行加工的心理学证据；2)报告了词切分的认知机理方面的研究现状和一个中文词切分和识别的模型；3)简要回顾了计算机科学中的词切分研究，并指出了与心理学中词切分研究的区别和联系；4)提出了一些尚待解决的问题和研究展望，这些问题的提出和解决将可能促进中文词切分的认知机理全面理解。

关键词 词切分；词识别；词的边界效应；中文阅读
分类号 B842.5

1 引言

阅读的认知机理是西方认知心理学、发展心理学以及现代认知神经科学的重要研究内容，具有悠久的研究历史以及众多的研究文献。西方的研究主要集中在英文等拼音文字，对中文的研究相对较少。中文与英文在很多方面都截然不同。例如，英文是拼音文字而中文是表意文字，英文词之间有空格而中文词之间没有空格。正因为如此，英文阅读的研究结论和理论模型不能直接应用于中文阅读。中文是中华民族的宝贵财富，是全球近四分之一人口的母语，更是凝聚全球华人的一个重要纽带。如何理解中文阅读的认知机理并使之更好地为中文读者服务，是认知心理学需要研究的一个重要问题。

词是语言中最小的能够独立运用的有音有义的单位(黄伯荣，廖序东，2007)。鉴于词在阅读中的重要作用，要理解语言中的句子乃至段落篇章，首先需要对词进行加工识别。要对词进行加工又需要把词从文本中切分出来，此过程简称为

收稿日期：2011-01-27
* 国家自然基金面上项目(Y0JJ292C01)和中国科学院知识创新工程重要方向项目(Y0CX222Y01)资助。
通讯作者：李兴珊, E-mail: lixs@psych.ac.cn

“词切分”。因此，词的切分在信息处理和阅读认知加工过程中起着重要作用(Li, Rayner, & Cave, 2009; Rayner, 1998, 2009; 郑昭明, 1981)，是词的加工过程中的首要的一个环节(Packard, 2000)。

大部分拼音文字的书写系统都是利用空间线索来标记词的边界。例如，英文中词和词之间的空格，可以将一连串的字母分隔成一个个的词。这些空格在英文读者阅读时进行词的切分起着十分重要的作用。已有研究表明，英文中的“空格”极大地促进了词汇识别和加工速率，删除或掩蔽词间空格会严重干扰读者正常的文字理解，致使阅读效率下降 30%~50%左右(Perea & Acha, 2009; Rayner, Fischer, & Pollatsek, 1998; Winskel, Radach, & Luksanneeyanawin, 2009)。另外，现代认知科学的一个重要工具是利用建模技术来模拟复杂的认知过程。英文阅读的模型往往都假设词在模型中起着非常重要的作用，例如，E-Z 阅读者模型(E-Z Reader model, Reichle, Erik, Pollatsek, Fisher, & Rayner, 1998; Reichle, Warren, & McConnell, 2009)、SWIFT 模型(Saccade-generation with inhibition by foveal targets, Engbert, Longtin, & Kliegl, 2002; Nuthmann & Engbert, 2009)和 IA 模型(Interactive Activation model, McClelland & Rumelhart, 1981)等。这些模型都认为，读者利用词之间的空格，在视觉感知阶段就把词切分开了，因此不需要更高级加工过程的参与。鉴于词和词之间没有明显的空格界限标记是中文阅读的特点，已有的英文阅读模型也并不适用于解释中文阅读的认知机理。因此理解词的切分机理也是理解中文阅读认知机理的一个关键问题。

词切分认知机理的研究和解决，不仅仅是揭示中文阅读机理的一个关键环节，还是现代信息科学发展的要求。由于现代信息技术的发展，大量的信息以文字的形式爆炸式增长。如何利用计算机对这些信息进行分类、检索等自动处理，是计算机科学面临的巨大问题。然而，目前利用计算机进行中文信息的自动处理还有很大难度。其中一个重要的难点在于词的切分。词的切分是计算机理解文本的基础。目前计算机自动分词的准确率还不能十分令人满意 (90%左右，许嘉璐，傅永和, 2006)。与计算机相比，中文读者在词的切分过程中，尤其是在切分准确度上，具有天然的优势。因此，词切分认知机理的研究和解决，可以为计算机科学的中文信息处理提供科学依据。

另外，理解词的切分机理也有可能为中文阅读技术上的革新提供理论支撑。现代信息技术的突飞猛进为文字呈现方式的革新提供了可能。计算机、互联网、电子图书已经逐步走进人们的生活。怎样更好地利用这些现代媒体呈现中文文本，从而提高人们的阅读效率已经显得越来越重要。其中的一个问题就是：改进中文阅读材料的呈现方式(比如在中文阅读的词之间增加空格)是否会提高人们的阅读效率？理解中文词切分的认知机理可以为这些革新提供理论支持。

因此，理解中文词切分的认知机理是摆在认知心理学、认知神经科学面前的一个重要问题。本文主要综述了认知科学领域的词切分认知机理方面的研究。首先总结一些在中文阅读中词作为整体进行加工的心理学证据，其次概括词切分的认知机理方面的研究成果，然后介绍一个中文词切分和识别的模型，之后简要综述计算机科学中的词切分研究，最后提出一些尚待解决的问题，从而为完全揭示中文词切分的认知机理提供一定的思路。

应该注意到，因为中文词之间没有标志，所以词的概念都是比英文中词的概念要复杂很多。首先，词和短语的界限还不是很明确。比如，语言学界对离合词(如“提高”)究竟是词还是词组还存在争论(刘泽先, 1955; 王力, 1982)。其次，不同人对词的边界的划分也存在不一致性(Hoosain, 1991)。比如，很多人把“美丽的”划分为一个词，而另一些人把它划分为“美丽”和“的”两个词。中文词的概念的复杂性为研究词的切分增添了挑战。然而，也正因为词的概念的复杂性，我们研究词的切分的认知机理才更有意义。

2 中文阅读中词作为整体加工的心理学证据

中文阅读的基本单位是词还是字？目前，这个问题在语言学领域仍然存在分歧(Hoosain, 1991)。但是，很多心理学的研究表明，中文词是作为一个整体进行加工的。本文将从下面四个角度展开论述。

2.1 词优效应

Cattell 早在 1886 年就发现，一个字母在单词中比在无意义的字母串中更容易被辨认。例如，相比在无意义的字母串“owrd”中，被试更准确快

速地在真词“word”中识别出字母“d”。这种识别差异的现象，称为词优效应(word superiority effect)。后续研究表明，词优效应并不是因为猜测引起的(Reicher, 1969)。词优效应表明“词”在阅读中是作为一个整体进行加工的。类似英文中的发现，中文阅读中同样存在词优效应。郑昭明(1981)曾经在字词快速呈现的条件下，要求被试检测一个字在 2 个字中的呈现位置，这 2 个字有时组成一个词(真词条件)，有时不能组成词(非词条件)。结果发现，中文读者在真词条件中识别字的正确率显著高于非词条件。中文中的词优效应表明刺激中的两个汉字并不是被独立地进行处理。相反，当两个字组成一个词时，这两个汉字是作为一个整体进行加工的。

2.2 词的加工与注意分布

Li 和 Logan (2008)发现中文词的加工影响着视觉注意的分布。在他们的研究中，被试看到以注视点为中心的两排两列共四个汉字，同一排或者同一列的两个汉字组成一个双字词。结果表明，被试的视觉注意在组成一个词的两个字之间的转移，比在不能组成一个词的两个字之间的转移速度更快。这个结果与基于物体的视觉注意方面的研究结果类似(Egly, Driver, & Rafal, 1994)。不同的是，传统的基于物体的视觉注意中的物体一般由低水平的视觉信息(比如两个方框)来定义。而这个研究采用的是中文词材料。这些结果表明，对这些由词构成的“物体”的加工，影响了被试的视觉注意分布。因而, Li 和 Logan 提出，与物体的加工类似。中文词也是作为一个整体来进行加工的。

2.3 词的属性对眼动模式的影响

大量英文阅读的研究表明，词汇的属性影响英文读者的眼动模式(Rayner, 1998, 2009; White, 2008)。与此类似，中文阅读中词的频率、预测性和复杂性等属性，也影响着读者对于这些具有不同属性词的注视时间和注视位置(Rayner, Li, Juhasz, & Yan, 2005; Yan, Tian, Bai, & Rayner, 2006)。Yan 等人(2006)发现中文阅读中的字频和词频都会影响被试的注视时间，但是字频效应受到词频效应的调节，即在词频较低时字频效应比较明显，但是在高频词中字频效应则变弱。Rayner 等人(2005)则发现中文阅读中词的预测性影响着被试对于关键词的注视时间和跳读概率。相对于较低预测性的词，高预测性的词更容易被跳读，并且被读者注视的时间较短。因而，中文阅读中词的属性能够影响着读者的眼动控制，成为支持中文阅读中以词为整体加工的又一个有利证据。

2.4 字间空格与词间空格

最近一些研究通过考察空格在中文阅读中的作用，同样支持了词作为整体加工的观点。Bai, Yan, Liversedge, Zang 和 Rayner (2008)在中文阅读文本中加入不同方式的空格，发现对被试阅读效率的影响程度不同。词之间加入空格时与正常文本条件下的阅读效率没有显著性差异。但是，在字与字之间加入空格时，阅读效率显著下降。这说明，词在中文阅读过程中是作为一个整体进行加工的，当在字间插入空格时，这种自下而上的视觉信息破坏了词的整体性，致使阅读效率下降。

综上所述，中文词在阅读中是作为整体进行加工的，具有一定的心理现实性。因而，词切分是中文阅读中一个必然的过程。

3　探讨中文词切分机理的实验研究

3.1 空格对中文词切分的影响

本文已经在引言部分简要论述了空格在英文词切分过程中的重要作用。在英文中，空格可以为读者提供词边界、词长等信息。那么，对一些表意文字，如中文、日文、泰文等，如果在这些文本中插入空格，又会对阅读产生怎样的影响呢？

近 30 年来，多位研究者通过不同的实验方法来考察空格在中文阅读中的作用。1974 年，刘英茂、叶重新、王联慧和张迎桂等人较早地通过在词与词之间插入空格的方法，来考察词单位对整个句子阅读效率的影响。结果发现，词间空格条件下的阅读时间显著长于无空格条件，词间空格干扰了被试的正常阅读。刘英茂等人分析，一方面，可能是因为词间空格改变了文本的呈现方式，扰乱了读者的阅读习惯，致使阅读效率下降；另一方面，文本通过 Gerbrands 两视野的速示器呈现，仪器本身的局限性可能不能直接反映句子的加工过程。

之后，一些研究者通过眼动技术，考察了空格对于中文句子阅读的影响，却发现了与刘英茂等(1974)不一样的实验结论，即词间空格并没有干扰正常阅读。中文读者在阅读词间空格和正常文本的中文句子时，整体阅读时间没有显著性差

异(Bai et al, 2008; Inholff, 刘伟民, 王坚, 符德江, 1997)。但是, Bai 等人(2008)的研究发现, 字间空格显著干扰了正常阅读。这些实验结论的不一致性, 除了因为实验技术的差异以外, 还有一个可能的原因是, 刘英茂等人(1974)选取的实验材料中, 词间空格条件下的每个句子仅包括 7 个字, 却组成了 6 个词, 即词单位的标准可能与被试切分词的标准不一致(彭瑞元, 陈振宇, 2004)。另外一种推测是, 刘英茂等人的研究结论是字间空格而不是词间空格干扰了被试的正常阅读。中文阅读材料中每个字间增加空格会干扰被试的正常阅读, 词间空格并没有显著地促进阅读, 一方面可能是空格破坏了中文读者传统的阅读习惯, 另一方面可能是字间或词间空格的间距增加了句子的物理长度, 从而抵消了空格标记词边界的促进作用(Bai et al., 2008; 刘英茂等, 1974)。

虽然上述的研究表明词之间增加空格并不能提高人们的阅读效率, 但是, 一些研究者提出, 在歧义或者难度较高的阅读材料中, 空格却能起到一定的促进作用(Hsu & Huang, 2000; Inhoff et al., 1997)。Hsu 和 Huang (2000)同样考察词间空格对中文阅读的影响, 发现空格促进了被试对中文阅读的理解。他们采用反应时记录的方法, 选择了具有一定歧义难度的文本材料, 比如“花生长在屋后的田里”。结果表明, 相比于传统的无空格文本, 词间空格减少了阅读时间, 提高了句子理解的准确度。最近, Ren 和 Yang (2010)通过眼动技术, 分别考察词、短语和从句后面的逗号与中文句子阅读中语法边界的关系。结果发现, 相比于没有逗号的条件, 逗号减少了被试对关键词和整体句子的注视时间, 表明逗号促进了词的识别过程。在某种程度上, 逗号和空格在标记词边界的作用上, 具有相似的功能, 均被一些研究者认为是词切分的边界线索。

另外, 一些研究者选择在中国学习的留学生以及中文初学者为被试, 来考察空格在中文阅读中的作用(白学军, 张涛, 田丽娟, 梁菲菲, 王天林, 2010; 高珊, 2006; 沈德立等, 2010)。高珊(2006)选择反应时和纸笔测验的方式, 考察词边界信息对留学生中文阅读影响。研究表明, 对于初级或者中级汉语水平的母语为拼音文字的欧美留学生来说, 词间空格促进了阅读加工, 而非词空格干扰了阅读理解。白学军等人(2010)通过眼动技术考察空格对美国留学生中文阅读的影响, 也得到类似的发现。但是, 词间空格并没有促进母语为表意文字的日韩等国留学生的中文阅读(高珊, 2006), 也没有促进汉语初学者的中国小学生们的阅读(沈德立等, 2010)。

上文主要总结了空格在中文阅读中的实验研究以及中文词切分的部分认知机理, 发现在一般情况下, 中文读者在阅读中即使有空格信息可以应用, 基本上也没有依靠空格进行词的切分, 而是利用了其他信息(比如 :自身的知识经验)进行中文词的切分。但是, 在一些较难或者有歧义的句子材料中, 空格起了一定的促进作用。值得思考的是, 母语为拼音文字的中文初学者, 更容易受到空格等低水平视觉信息的影响, 采用这种自下而上的加工方式, 即有效利用空格信息进行词切分; 但是母语为表意文字的中文初学者, 比如小学生们或者日韩留学生, 并没有显著地受到空格的影响, 同中文读者表现出相似的模式。

和中文文本相似的词间没有空格的语言还包括泰文、日文等。日文包括日本汉字、平假名和片假名, 其中日本汉字为表意文字, 但是平假名和片假名为拼音文字。一些研究者考察了空格在泰文和日文中的作用, 结果发现, 词间加空格促进了泰文被试的阅读速度(Kohsom & Gobet, 1997)以及词汇识别, 但是并不影响眼动控制和词切分, 这说明泰文的阅读加工也是以词为单位进行的(Winskel et al., 2009)。日文阅读研究表明, 在平假名—日本汉字的混合文本中插入空格, 并没有促进作用, 但是在平假名文本中插入空格, 可以提高阅读效率(Sainio, Hyona, Bingushi, & Bertram, 2007)。泰文、日文和中文同为无空格的文本, 但是在插入空格的条件下, 读者的阅读理解却有不同的成绩表现, 这可能与文字的书写文体有一定相关性。汉字和日本汉字都属于表意文字的词素音节文字, 但是泰文属于拼音文字, 和西方文字有很大相似性。拼音文字和表意文字有可能在词切分的认知机理方面具有很大差异性。

3.2 单向切分与多重激活

视野中往往会同时出现几个词, 这些词是以怎样的顺序进行加工的? Inhoff 和 Wu (2005)探讨了两种可能的加工策略。一种为单向切分假设(unidirectional parsing hypothesis), 认为简体中文文本中字词的加工是严格按照序列方式, 从左向

右逐渐进行的; 另一种为多重激活假设(multiple activation hypothesis), 认为在一个知觉广度范围内的所有汉字可能组合成的词汇都会被激活, 而不受到方向性的限制。

Inhoff 和 Wu (2005)选择了具有空间歧义的 4 个连续关键字的句子作为阅读材料, 探讨中文读者如何确定词的边界信息, 以及区分上述两种词切分机理假设的正确性。他们比较了读者在两种条件下的眼动控制模式, 一种是在歧义条件下, 4 个连续关键字能够组成三个词, 例如, “专科学生”可以被切分为“专科”、“科学”和“学生”; 在控制条件中, 4 个连续关键字只能组成 2 个词, 中间 2 个字不能组成一个词, 例如, “专科毕业”, 只能被切分为“专科”和“毕业”。结果发现, 被试在歧义条件中对于 4 个连续关键字的凝视时间、整体注视时间以及 4 个关键字中间 2 个字的注视时间都显著长于控制条件。他们认为, 实验结果支持了多重激活假设, 表明在知觉广度范围内所有字词的切分处理是同时激活的, 并没有严格按照从左向右的序列加工方式进行(Inhoff & Wu, 2005)。

3.3 词切分对眼动落点位置的影响

英文阅读的眼动研究表明, 当眼睛注视位置在词的中央时, 相对于注视在词首或词尾, 读者对于该词的重新注视概率比较低, 并且对于该词的识别速率比较快(O' Regan & Jacobs, 1992)。此处的眼动落点位置称为最佳注视位置(optimal viewing position, 简称 OVP)。但是, 首次注视的偏好注视位置(preferred viewing location, 简称 PVL), 大约在词首和词中央之间, OVP 的左侧(Rayner, 1979)。不少研究表明, 英文中作为词边界标记的空格, 有助于读者确定眼跳目标。如果去掉空格或者用字母、数字等替代空格, 均会干扰阅读, 并且影响读者的眼动模式(Perea & Acha, 2009; Pollatsek & Rayner, 1982; Rayner et al., 1998; Winskel et al., 2009)。例如, Rayner 等人(1998)发现, 去掉空格将导致读者延长词的注视时间, 增加重复注视的概率。Winskel 等人(2009)同样发现, 当阅读没有空格的英文时, PVL 落点接近词首位置, 而不是在有空格条件下的词中间偏左的位置。

那么在无空格文本中, 如泰文、日文、中文等, 插入空格或词切分将会对眼动落点位置产生怎样的影响呢？Winskel 等人(2009)考察了空格对泰文阅读的眼动模式的影响, 发现有空格和无空格条件下, 被试的首次注视位置没有显著性差异。与此类似, 一些研究者以日文为阅读材料, 考察了日本读者的落点位置分布 (Kajii, Nazir, & Osaka, 2001; Sainio et al., 2007)。Kajii 等人(2001)采用传统的无空格日本文本为材料, 发现读者的 PVL 倾向于落在词首, 而不是词的中央, 并且注视位置的分布依赖于字的属性。相比于日文中的平假名和片假名, 读者更多的注视日本汉字。但是, 在全部是平假名的材料中, 字的位置并没有影响注视位置的分布。Sainio 等人(2007)在此研究的基础上, 在日文中加入空格, 考察读者的眼动模式。相对于无空格条件, 发现在平假名之间插入空格, PVL 逐渐迁移到词的中央, 同英文空格的研究类似。但是, 在日本汉字和平假名的混合文本中, PVL 并没有受到空格的影响。

中文阅读中的眼动落点选择比较复杂。早期的中文眼动阅读发现 PVL 曲线比较平缓, 即首次注视落点的位置平均分布在字的各个位置 (Tsai & McConkie, 2003; Yang & McConkie, 1999)。最近, Yan, Kliegl, Richter, Nuthmann 和 Shu (2010)等人对中文阅读中的单字词、双字词、三字词和四字词的眼动落点分布进行分析。他们发现, 当一个词只有一个注视点时, 被试倾向于注视词的中心; 而当一个词被注视多次时, 首个注视点倾向于落在词首位置。他们认为, 是因为在副中央窝存在一个词切分过程。如果在副中央窝区域能够完成切分过程, 读者的下一次注视落点位置则倾向于在词的中央; 反之, 落点位置在词首。根据这个假说, 中文阅读者在眼睛注视在一个词之前可能利用一些副中央窝中获取的信息进行词的切分, 并用切分的结果引导眼动落点的选择(Yan et al., 2010)。这个假说虽然可以解释 Yan 等人观察到的实验数据, 然而遗憾的是, 这个解释却不是唯一可以解释数据的假说。另一种可能的解释是：因为眼跳刚好落在一个词的中央位置, 整个词都可以被识别出来, 因此同一个词上也就不需要再一次注视。而如果眼跳刚好落在一个词的前面, 则需要注视在这个词的其他字后才能完成整个词的识别。

Li, Liu 和 Rayner (in press)选择不同词长(2 字条件和 4 字条件)的关键词, 嵌入到相同的句子框架中, 考察中文读者的眼跳落点位置。假如在某些情况下中文阅读中的眼动落点会瞄准词的中

间位置(如同 Yan 等人的假说所预测的)，则被试在 2 字词和 4 字词条件下 PVL 曲线也应该是不同的。因为 4 字词的中间位置比 2 字词的词中间位置偏右，4 字词条件下的 PVL 曲线的峰值也应该向右偏移。然而, Li 等人的结果并没有验证这个结论。2 字条件和 4 字条件下的 PVL 曲线是几乎相同的。这说明，目前并没有证据表明，中文阅读中的眼动控制像英文中那样，以词的中间位置作为眼跳的目标。

概括之，中文阅读中词切分和眼动落点位置的关系仍然需要进一步的探讨。以往研究表明，词的边界信息影响着眼动中注视的落点位置，研究者可以通过注视的落点位置来推测词切分的认知机理。

3.4 词的边界效应

为了探索中文词切分的机理，Li 等人(2009)提出并检验了两种中文词切分机理的理论假设，分别为前馈假设(feed-forward hypothesis)和整体假设(holistic hypothesis)。前馈假设是指通过视觉加工单元获得的视觉信息，传送到字的识别单元，不同的字相互独立地进行加工，识别出的字前馈到词的识别单元，在词的识别阶段整合成词并完成词的识别。前馈假设认为中文词的认知过程只有自下而上的前馈，而没有自上而下的反馈。而整体假设是指汉字的视觉信息加工、字的识别和词的识别等各个单元之间相互作用，共同合作来影响词的切分和识别。假如前馈假设正确，词的属性如词的边界不会影响字的识别；而整体假设做出了相反的预测。

该研究中，被试看到 4 个汉字，这 4 个字在一种条件下是一个词(1 词条件)，在另一种条件下是两个词(2 词条件)，即前 2 个字组成一个词，后 2 个字组成另外一个词。被试的任务是尽可能多地报告看到的字。因为呈现时间非常短，被试往往不能报告所有的字。结果发现，被试往往能够快速地识别出 1 词条件中所有的字，但是在 2 词条件中，被试仅能识别出前 2 个字，却很难认出后 2 个字。此现象被称为“词的边界效应” (word boundary effect)。为了排除被试在 1 词条件中看到前 2 个字而猜测出整个词的可能性，Li 等在另一个实验中又增加了半词条件。半词条件是指 4 个字中前 2 个字是一个 4 字词的前半部分，然而后 2 个字却不能与前两个字组成任何有意义的词汇。假如词的边界效应是因为被试根据前两个字猜出整个四字词，那么被试在这种条件下也应该报告完整的 4 字词。然而，实验结果发现被试一般不能够报告出 4 字词(图 1，四种条件下字的识别准确率)。这样就排除了被试的主观猜测对边界效应的影响。

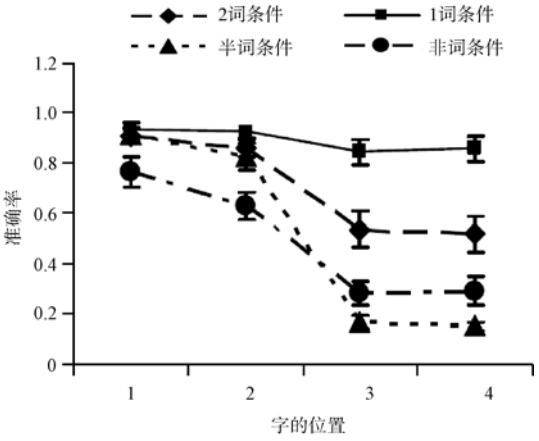

图 1 4 种条件下字识别的准确率。横坐标代表字的 4 个不同位置，纵坐标代表字识别的准确率。(引自：Li et al., 2009)

这些研究结果表明，整体假设更符合中文词切分的认知机理。词的切分和识别过程相互影响，是一个自下而上和自上而下交互进行的信息加工方式，而并不是简单的自下而上的前馈过程。

4 词切分和识别模型

一些研究者在实验论证和理论假设的基础上，尝试着通过构建模型，来解释词切分过程。Hoosain (1991)认为词的切分过程需要读者自身的语言知识、语境加工等来协助才能完成一个个词的切分。这些一个个的词单元再组合成多种多样的句子。按照这种方式，一个个独立的汉字识别的平行加工过程隶属于词的识别，被试从自上而下和自下而上过程中获取足够多的信息，才能完成词的切分。Perfetti 和 Tan (1999)采用花园路径句子材料，发现中文读者更倾向于将系列汉字串切分成由 2 个汉字组成的单元，而不是将每一个汉字都看作一个词。因而，他们提出读者在阅读中文句子时，倾向于采用两字结合的策略(a two-character assembly strategy)。

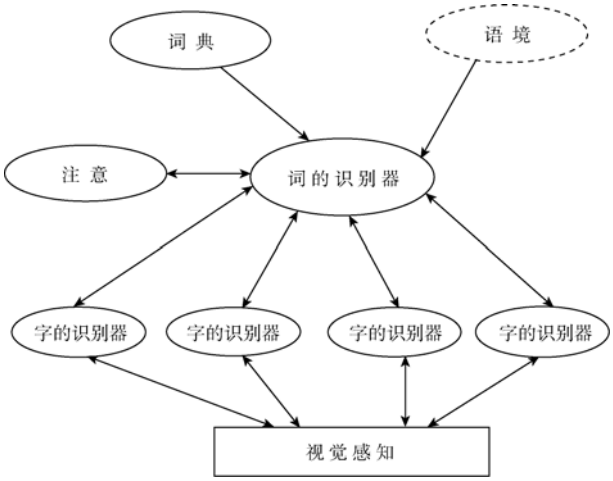

图 2　词切分和识别模型 (引自：Li et al., 2009)

Li 等人(2009)借鉴了英文词汇识别的理论，模拟了中文词的识别和切分过程，构建了一个中文词切分和识别的计算模型(图 2)，并且很好地拟合了实验观测到的数据。该模型借鉴了 McClelland 和 Rumelhart (1981)提出的交互激活模型的一些假设。该模型认为英文词汇识别包括多层次加工单元，即视觉感知单元，字母加工单元和词汇加工单元。首先，字的一些视觉特征被视觉系统感知到，然后传送到词的加工单元，并激活相对应的词，被激活的词又反过来影响字的识别。在词的识别单元，相关的词相互竞争。通过多次循环，词语识别单元中将会有一个唯一的词条胜出。当某一个词胜出后，词就被识别出来，也就完成了词的切分。该模型假设各个层次间存在交互作用，因而相应位置的字的识别与视觉加工也会受到影响。属于该词的字的加工单元以及相应的视觉处理单元会被促进，而不属于该词的单元会被抑制。在 Li 等人(2009)研究中的 2 词条件下，由于视觉注意等因素的影响，第一个词往往更容易在竞争中胜出，从而会对属于该词的字的识别有促进作用，因而属于第一个词的字容易被识别出来。与此对应，由于属于第二个词的字不容易在竞争中取胜，因此相应的字的识别只能接受自下而上的信息，所以识别速度会相对较慢。这样，就会在字的识别与视觉加工水平呈现出词的边界效应。

除了交互假设外，该模型做出了如下重要假设：首先，词的识别与切分是一个统一的过程，二者是不可区分的。只有当词识别出来时，词才被切分开来。第二，落在视野中字的加工是并行进行的。但是，这些字的识别效率受到视觉注意的影响。离注视点距离越远，识别效率越低。第三，词的识别是一个串行的过程。在词的识别单元，在一次竞争过程中只能有一个词胜出。第四，当一个词被识别出后，刚刚识别出的词和字所对应的单元被抑制，然后再开始下一轮的竞争，从而开始下一个词的识别过程。

Li 等人(2009)的模型也可以解释 Inhoff 和 Wu (2005)的研究结果。如上所述，Inhoff 和 Wu (2005)发现读者阅读时，在“专科学生”等歧义短语上的注视时间长于“专科毕业”等非歧义短语。Li 等人的模型认为字的加工是并行的，所以视野中出现的字所能组成的词之间会相互竞争。读者在加工“专科学生”等歧义短语时，“专科”，“科学”和“学生”等词就会被激活。而当看到“专科毕业”时，只有“专科”和“毕业”等词被激活。因此，“专科学生”比“专科毕业”在词的识别单元激活更多的词条。在词的识别单元中，有更多词条参与竞争，意味着需要花费更长的时间决出胜利者。因此，被试在注视“专科学生”比在“专科毕业”上花费了更长的时间。

据我们所知，Li 等人(2009)提出的模型首次尝试通过计算机仿真，来模拟和解释中文词切分和识别的认知机理。虽然这个中文词切分和识别

模型还需要一定的改进和完善，但是，它仍然能够为解释中文词切分的机理问题提供一种思路，从而促进更深入地理解中文阅读的认知机理。

5 词切分在计算机科学中的研究

词切分在计算机领域中被称为分词，就是将连续的字序列按照一定的规范重新组合成词序列的过程，这个步骤是计算机科学中自然语言处理的基础。中文自然语言处理的对象是越来越庞大的语料库，如何快速而便捷地完成自动文摘、智能检索、信息监控成为日趋重要的课题。此外，词切分的研究对于机器翻译、语音识别、智能输入、数据挖掘、人机对话和机器测评等领域都具有重要的应用价值。随着计算机处理速度的加快，以及相应人工智能的发展，词切分的应用前景将更加广阔。这种广阔的应用前景，在一定程度上推动着词切分的研究，使得这一研究领域在近二十多年来未曾降温。

5.1 计算机科学中词切分的基本方法

计算机科学中的分词系统主要有以下三种：基于词典的分词系统、基于人工智能的分词系统以及基于统计的分词系统。

基于词典的分词系统的基本思想是：事先建立一部完善的辞典，尽可能包含所有可能出现的词汇，计算机进行词切分的活动，就是对文本信息进行扫描，与现有词典进行匹配的过程，当匹配成功时就完成了词汇的切分。这类切分系统包括三个要素：一是词典，二是扫描方式，三是匹配法则。词典要素中最重要的是词典机制，即词典是如何组织的。现有的词典机制包括二分查找的索引表、TRIE 索引树、PATRICIA 搜索树以及哈希(Hash)机制，这类不同的词典机制的制定都是为了提高词切分的速度和准确率。对于文本扫描方式，最先被运用的是序列扫描，包括正向匹配法、逆向匹配法以及双向扫描法。吴胜远(1996)曾介绍了一种基于并行处理的方法，这种方法建立在多级内码(Multilevel Machine Code, MMC)理论的基础上，使分词的速度有所提高。对于匹配原则，最大匹配和最小匹配两种原则较为常用，最大匹配确保词典中存在的较长词串例如“中国科学院”被切分为一个词序列，而最小匹配确保了单位句子切分出的词数最少。

基于人工智能的分词系统主要包括专家系统分词法和神经网络分词法。专家系统分词法模拟了人脑的功能，把分词看作推理判断的过程，人们事先将分词机制，例如语法、语义、构词法等信息，存储到专家系统的推理机中，推理机运用分词机制判断并输出词汇。而神经网络分词法包括输入层、内隐层和输出层，这种方法模拟了人脑的结构以及分布处理的工作方式，构造出神经元之间的连接机制，并将通过样本学习获得的分词规范内隐地存储在神经网络的内隐层中，该网络模型具有很强的在线学习能力，可以通过学习和训练改变其内部权值，以达到正确的分词效果。

基于统计的分词方法也叫做无词典分词法，它的主要思路是利用字与字之间的互信息来实现的。互信息就是文本库中汉字相邻出现的概率，在文本库中，两个汉字相邻概率越高，就越有可能共同构成一个词汇。这种方法只需要对文本中字符的互信息进行统计，不需要借助于词典，在解决歧义词和未登录词方面比基于词典的切分更具有一定优势。

5.2 计算机科学中词切分的现状

汉语词汇复杂多变，很难从中概括出一条统一的规律适合于所有词汇的切分。基于词汇的切分困境在于很难组建一个完善的词典，并构造出一个完善的切分策略。基于人工智能的切分目前尚处于发展阶段，它的困境是如何获取完备的规则库，包含所有的切分策略，或者如何找到一个完善的样本，让计算机在短时间内习得误差最小的切分方式。基于统计的切分策略所面临的困境是算法问题，很难完全依赖于一种简单的算法就能解决所有切分问题。三类切分系统虽然在一定程度上解决了大部分词汇切分问题，但由于汉语组词的不确定性(Gao, Li, Wu, & Huang, 2006)，使得依赖于机器的切分在精确度上仍然存在一定缺陷。

在当前的切分系统中，我们很难找到简单地依赖于一种切分方法的模型。单一切分模型所带来的问题通过模型整合得到了部分解决，例如，词典与统计算法的结合以及词典与智能系统的结合。目前基于词典与统计结合的模型，主要包括高频优先法、扩充转移网络法(ATN)、约束矩阵法以及分词—句法一体化法，这些模型在解决现实切分问题上比单纯依赖匹配或统计有很大优势。

词典与智能系统的结合也有了一定的发展。最近, 李华、陈硕和练睿婷(2010)提出了神经网络和匹配融合的中文分词研究, 在精确度上比单一系统有较多的提高。

此外, 基于字的切分与基于词的切分结合也是一个重要趋势。黄昌宁和赵海(2007)指出, 自动分词方法在 2002 年之前基本上是基于词(或词典)的, 而在 2002 年之后, 基于字的模型逐渐占据主导地位。Xue 和 Shen (2003)运用一种基于字的切分—最大墒模型, 取得了可喜的成绩。随后, 微软公司运用条件随机场(CRF)模型(Gao et al., 2006; Zhao, Huang, & Li, 2006)研制的 MSRA 分词系统在 Bakeoff 评测中达到或接近了历年评测的最佳水平。然而, 基于字的切分优势不能掩蔽本身的缺陷, 两种方法的融合也是势在必行。宋彦、蔡东风、张桂平和赵海 (2009)提出了基于字词联合解码的中文分词方法, 发挥了基于字的切分和基于词的切分两种模型的优势, 有效改善了单一模型的性能。

智能化是未来时代的主题, 如何使词汇的切分实现真正的智能化, 是未来词汇切分研究的重要课题。机器智能化切分的目标是达到专家切分的高精确度, 而这项工程的完成还需要一定的时间。

5.3 与心理学中词切分研究的异同

计算机科学与心理科学中都在探讨词切分的问题, 两个领域对词切分的研究有一定相似之处, 这两个领域都研究语言信息的加工问题, 聚焦点都是词切分。然而, 两者又存在很大不同。前者研究计算机的自动分词方法, 而后者探讨人脑的词切分机制。此外, 两者研究目标不同, 前者是为了实现一种技术, 使得计算机能够更快速地处理语言信息, 从而在不同领域实现其应用价值。后者是为了探讨大脑对语言信息的处理机制, 即词切分在大脑中究竟是一个怎样的认知过程。

在一定程度上, 计算机科学与心理科学中的词切分研究是相互影响的。计算机中的词切分内容为心理学词切分的研究提供了一些启示, 例如, 歧义词切分问题、理解与切分的加工顺序等问题。心理学的研究成果也能为计算机中词切分研究提供一定借鉴, 譬如对词的可预测性、心理词典的研究等, 这些都能在计算机智能化切分上体现出应用价值。因而, 二者具有一定的关联性, 并且相互影响, 可以相互借鉴, 但绝不能相互替代。

6 需要解决的问题及研究展望

中文词切分的研究仍然存在一些尚待解决的问题。本章节将简要陈述这些问题, 希望为未来的研究提供一些思路。

6.1 词的识别与词的切分的时序关系

中文阅读中, 词的切分和词的识别有没有先后顺序？一方面, 英文的词的识别模型一般是指对已经切分完的词进行识别, 所以这些模型要求先进行词的切分。而另一方面, 词的切分需要利用与词的知识相关联的高水平信息。所以词的识别与切分哪个在先哪个在后的问题, 有点类似于先有鸡还是先有蛋的问题。Li 等人(2009)的模型假设词的切分和词的识别是一个不可分的统一过程, 二者同时完成。然而, 这个假设还需要实验研究进一步的考证。

6.2 语法与语义等高水平信息的加工对词切分的影响

在词的切分中, 有些歧义字段仅依靠词汇信息, 是不能完全消除歧义的。比如“花生长在屋后的田里”, 既可以理解为“花\生长\在\屋\后的\田里”, 也可以切分为“花生\长\在\屋\后的\田里”。这种词切分的不一致性也会造成不同的理解方式(Li et al., 2009; 彭瑞元, 陈振宇, 2004)。仅依靠词汇的信息或仅依靠这个句子的信息, 都无法确定哪种切分方法或理解方式是正确的。这种句子的正确切分必须依赖于前后文的语境信息。中文词的切分过程中如何使用语法、语义等信息, 是理解中文词切分机理的一个关键问题。

6.3 词的加工过程中不同层次的相互影响

词的识别模型经常假设, 词的加工分为几个层次来进行的。Li 等人(2009)的词切分和识别模型也做了类似的假设, 认为词的切分是一个字的识别和词的识别交互进行的过程。但是, 不同层次之间是如何交互的？这还是一个需要进行实验研究的问题。词的切分过程是否能够影响视觉处理层次？Li 和 Logan (2008)的研究提供了一定的证据, 表明中文词的感知与普通物体的感知过程具有一定的相似性, 词的感知可以影响视觉注意的分布。那么, 在阅读过程中, 词的边界是否也能够影响视觉注意的分布呢？

6.4 词在大脑中的编码方式

不同的中文阅读者对词的边界认识存在着很大的不一致性(Hoosain, 1991)。例如，人们在切分“重要的”时候，一般会切分成“/重要的/”或“/重要/的/”两种形式；在切分短语如“新闻媒体”时，有些人认为是一个词，而另一些人认为是两个词(“新闻”和“媒体”各是一个词)。英文中同样存在这种类型的复合词，比如“softball”或“soft-footed” (Juhasz, Inhoff, & Rayner, 2005)。出现这种现象的原因究竟是什么呢？一个可能的原因是人们对词的表征方式存在不一致性。很多研究者提出了心理词典的概念(Aitchison, 2003)。心理词典对词切分究竟起什么作用呢？人的阅读习得过程及阅读经验，在词的表征产生过程中又起到什么样的作用呢？另外，研究字和词、词和词组在大脑中的编码方式和处理机制的异同，也是词的切分机理研究的一个重要内容。

由此可见，关于中文词的切分机理的研究还有很多问题需要解决。随着研究者们对词的切分机理研究的推进,也将可能提出更多更深入的问题。解决这些问题并提出和解决新的问题的过程，将会极大地促进我们对词切分的认知机理的全面理解。

参考文献

白学军，张涛，田丽娟，梁菲菲，王天林. (2010). 词切分对美国留学生汉语阅读影响的眼动研究. *心理研究, 3*, 25–30.

高珊. (2006). *词边界信息对留学生汉语阅读的影响*. 硕士论文：北京语言大学.

黄伯荣，廖序东(编). (2007). *现代汉语*. 北京：高等教育出版社.

黄昌宁，赵海. (2007). 中文分词十年回顾. *中文信息学报, 21*, 8–19.

Inholff, A., 刘伟民，王坚，符德江. (1997). 汉语句子阅读中的眼动与空间信息的运用. In 彭聃龄，舒华，陈烜之(编), *汉语认知研究* (pp. 296–312). 济南：山东教育出版社.

李华，陈硕，练睿婷. (2010). 神经网络和匹配融合的中文分词研究. *心智与计算, 4*, 117–127.

刘英茂，叶重新，王联慧，张迎桂. (1974). 词单位对阅读效率的影响. *中华心理学刊, 16*, 25–32.

刘泽先. (1955). 用连写来规定词儿. 见：*《中国语文》*编辑部编.*《汉语的语儿和拼写法》*第一集 (pp.92–95). 北京：中华书局.

彭瑞元，陈振宇. (2004). "偶语易安、奇字难适"：探讨中文读者断词不一致之原因. *中华心理学刊, 46*, 49–55.

沈德立，白学军，臧传丽，闫国利，冯本才，范晓红. (2010). 词切分对初学者句子阅读影响的眼动研究. *心理学报, 42*, 159–172.

宋彦，蔡东风，张桂平，赵海. (2009). 一种基于字词联合解码的中文分词方法. *软件学报, 20*, 2366–2375.

王力. (1982). *汉语语法纲要*. 上海：上海教育出版社。

吴胜远. (1996). 一种汉语分词方法. *计算机研究与发展, 33*, 306–311.

许嘉璐，傅永和(编). (2006). *中文信息处理现代汉语词汇研究*. 广州：广东教育出版社.

郑昭明. (1981). 汉字认知的历程. *中华心理学刊, 23*, 137–153.

Aitchison, J. (2003). *Words in the mind: An introduction to the mental lexicon*. Victoria, Australia: Blackwell publishing Ltd.

Bai, X., Yan, G., Liversedge, S. P., Zang, C., & Rayner, K. (2008). Reading spaced and unspaced Chinese text: evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 1277–1287.

Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual-attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General, 123*, 161–177.

Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research, 42*, 621–636.

Gao, J., Li, M., Wu, A., & Huang, C. (2006). Chinese words segmentation and named entity recognition: A pragmatic approach. *Computational Linguistics, 31*, 531–574.

Hoosain, R. (1991). *Psycholinguistic implications for linguistic relativity: A case study of Chinese*. Hillsdale, New Jersey.: Lawrence Erlbaum Associates, Inc.

Hsu, S.-H., & Huang, K.-C. (2000). Interword spacing in Chinese text layout. *Perceptual and Motor Skills, 91*, 355–365.

Inhoff, A., & Wu, C. (2005). Eye movements and the identification of spatially ambiguous words during Chinese sentence reading. *Memory & Cognition, 33*, 1345–1356.

Juhasz, B. J., Inhoff, A. W., & Rayner, K. (2005). The role of interword spaces in the processing of English compound words. *Language and Cognitive Processes, 20*, 291–316.

Kajii, N., Nazir, T. A., & Osaka, N. (2001). Eye movement control in reading unspaced text: the case of the Japanese script. *Vision Research, 41*, 2503–2510.

Kohsom, C., & Gobet, F. (1997). Adding spaces to Thai and English: Effects on Reading. In: *the Proceedings of the*

19th Annual Meeting of the Cognitive Science Society (pp.388–393).

Li, X. S., Liu, P. P., & Rayner, K. (in press). Eye movement guidance in Chinese reading: Is there a preferred viewing location? *Vision Research*.

Li, X. S., & Logan, G. (2008). Object–based attention in Chinese readers of Chinese words: Beyond Gestalt principles. *Psychonomic Bulletin & Review, 15*, 945–949.

Li, X. S., Rayner, K., & Cave, K. R. (2009). On the segmentation of Chinese words during reading. *Cognitive Psychology, 58*, 525–552.

McClelland, J. L., & Rumelhart, D. E. (1981). An Interactive Activation Model of Context Effects in Letter Perception: Part 1. An Account of Basic Findings. *Psychological Review, 88*, 375–407.

Nuthmann, A., & Engbert, R. (2009). Mindless reading revisited: An analysis based on the SWIFT model of eye-movement control. *Vision Research, 49*, 322–336.

O' Regan, J. K., & Jacobs, A. M. (1992). Optimal viewing position effect in word recognition: A challenge to current theory. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 185–197.

Packard, J. L. (2000). *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge, England.: Cambridge University Press.

Perea, M., & Acha, J. (2009). Space information is important for reading. *Vision Research, 49*, 1994–2000.

Perfetti, C. A., & Tan, L. H. (1999). The constituency model of Chinese word identification. In Jian. Wang, Albrecht W. Inhoff & Hsuan-Chih. Chen (Eds.), *Reading Chinese Script* (pp. 115-134). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Pollatsek, A., & Rayner, K. (1982). Eye movement control in reading: The role of word boundaries. *Journal of Experimental Psychology: Human Perception and Performance, 8*, 817–833.

Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. *Perception, 8*, 21–30.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372–422.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62*, 1457–1506.

Rayner, K., Fischer, M. H., & Pollatsek, A. (1998). Unspaced text interferes with both word identification and eye movement control. *Vision Research, 38*, 1129–1144.

Rayner, K., Li, X. S., Juhasz, B. J. & Yan, G. (2005). The effect of word predictability on the eye movements of Chinese readers. *Psychonomic Bulletin & Review, 12*, 1089–1093.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81*, 275–280.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review, 105*, 125–157.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E–Z reader to model effects of higher–level language processing on eye movements during reading. *Psychonomic Bulletin & Review, 16*, 1–21.

Ren, G., & Yang, Y. (2010). Syntactic boundaries and comma placement during silent reading of Chinese text: evidence from eye movements. *Journal of Research in Reading, 33*, 168–177.

Sainio, M., Hyona, J., Bingushi, K., & Bertram, R. (2007). The role of interword spacing in reading Japanese: An eye movement study. *Vision Research, 47*, 2575–2584.

Tsai, J. L., & McConkie, G. W. (2003). Where do Chinese readers send their eyes? In J. Hyona, R. Radach. & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 159–176). Oxford: UK: Elsevier.

White, S. J. (2008). Eye movement control during reading: effects of word frequency and orthographic familiarity. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 205–223.

Winskel, H., Radach, R., & Luksanneeyanawin, S. (2009). Eye movements when reading spaced and unspaced Thai and English: A comparison of Thai–English bilinguals and English monolinguals. *Journal of Memory and Language, 61*, 339–351.

Xue, N., & Shen, L. (2003). Chinese word segmentation as LMR tagging. In : *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, Taiber, Taiwan (pp.176–179).

Yan, G., Tian, H., Bai, X., & Rayner, K. (2006). The effect of word and character frequency on the eye movements of Chinese readers. *British Journal of Psychology, 97*, 259–268.

Yan, M., Kliegl, R., Richter, E. M., Nuthmann, A., & Shu, H. (2010). Flexible saccade–target selection in Chinese reading. *The Quarterly Journal of Experimental Psychology, 63*, 705–725.

Yang, H.-M., & McConkie, G. W. (1999). Reading Chinese: some basic eye–movement characteristics. In Jian. Wang, Albrecht W. Inhoff & Hsuan–Chih. Chen (Eds.), *Reading Chinese Script* (pp. 207–222). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Zhao, H., Huang, C.-N., & Li, M. (2006). An improved

Chinese word segmentation system with conditional random field. In: *the Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, Sydney (pp.162–165).

Advances in Cognitive Mechanisms of Word Segmentation During Chinese Reading

LI Xing-Shan[1]; LIU Ping-Ping[1,2]; MA Guo-Jie[1,2]

([1] *Key laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing* 100101, *China*)

([2] *Graduate University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: A number of cognitive studies have indicated that words play a critical role in reading. Hence, word segmentation is an important procedure in reading. Unlike alphabetic writing systems such as English, there are no spaces between words. Without spaces, how do Chinese readers segment words? In this article, recent progresses in the following topics on Chinese word segmentation are reviewed: 1) Evidences that Chinese characters belonging to a word are processed as a unit; 2) Some recent psychological studies on Chinese word segmentation and some of the models; 3) Word segmentation studies in computer sciences; 4) Future directions on this topic.

Key words: word segmentation; word recognition; word boundary effect; Chinese reading