《当代语言学》第9卷2007年第3期193－207页，北京

小学儿童词汇获得的自组织模型*

邢红兵 北京语言大学　舒华 北京师范大学　李平 美国里士满大学

提要 本研究利用已有的数据库及相关研究成果，实现了汉字字形和字音的表征，并运用自组织特征映射模型的生理合理性和心理合理性，对小学儿童词汇获得的心理机制进行了模拟。模型成功地模拟出了形声字命名的年级效应、频率效应、规则性效应及其交互作用。模拟结果与儿童行为实验的发现基本一致。这一模拟结果也与数据库的统计分析结果相一致，表明了儿童语言发展规律与输入学习材料的统计属性的一致性。

关键词 联结主义理论　词汇获得　自组织模型　形声字　规则性效应

1. 引言

在过去的三十年间，对英语读音的心理加工机制的解释主要有两大观点，一种是基于规则的解释，即有一组"规则"可以处理英语中的规则词读音，另外一个机制处理不规则的例外词，"双通道"理论是这种解释的主要代表。联结主义理论(connectionism)则认为词汇识别是通过大量简单加工单元的合作和竞争交互作用实现的(Plaut, *et al.* 1996; Harm and Seidenberg 1999)。外界环境的频率信息导致单元之间连接强度的调整，从而影响网络的计算。这种单一机制，既可以处理规则词，又可以处理例外词。自1980年代初以来，联结主义的并行分布式信息加工(parallel distributed processing, PDP)模型已用"分布表征"(distributed representation)成功模拟了英语词汇识别和命名中的各种现象，其中包括规则性效应、一致性效应以及规则性、一致性与频率的交互作用的"在线"(on-line)加工。

汉语的书写系统是由汉字组成的，一般认为，汉字对应的是语素，而不是表音符号。很多心理学家曾认为汉字读音是类似图形的命名，学习汉字的过程是一个死记硬背的过程，语音技巧在汉字学习中起较小的作用。然而，过去二十多年的大量研究表明，汉字正字法和语音的关系也是有规律的。汉字中的一个亚词汇基本单位——声旁提供了汉字读音的信息(Peng, *et al.* 1994; 舒华等 1998; 张亚旭等 2003)，在整字加工过程中声旁的字形得到分解，并且激活了声旁的语音信息(Zhou and Marslen-Wilson 1999a, 1999b; 武宁宁、舒华 1999; 杨珲等 2000)。然而，和英语类似，汉字的形和音的关系也是不完全规则的，有大量的不规则形声字存在。成人的汉字形声字命名实验中发现了类似英语的频率效应、规则性效应、一致性效应及其交互作用。在儿童阅读研究中发现了声旁规则性和一致性意识的发展(Wu, *et al.* 1999)，而且发展特点与小学课本汉字分布特征有很高的相关性(舒华等 1998)。

对汉字识别和命名的计算机模拟研究到目前为止还相当少，陈鹰和彭聃龄(1994)曾建

* 本研究受到国家自然科学基金(30470574, 60534080)和全国教育科学"十五"规划重点课题(DBA010171)的资助。

构了一个由42个语音输出单元、200个隐单元、420个字形单元组成的PDP模型，在汉字字形的表征上采用构成成分和结构关系两个角度对汉字进行表征。模型采用BP算法，测试了规则字和例外字在不同训练期的误差情况，发现频度和规则性存在交互作用，高频字不存在差异，低频字存在明显差异。以往的模型还没有能真正模拟知识获得的过程，主要存在以下几个方面的不足。首先，对字形的表征还只是以区别不同的汉字为主要目的，不是让模型获得有关字形的知识；其次，对模型的训练方法多是采用有指导的学习算法，这与自然语言获得的机制不一致；最后，对模型的测试只是测试模型已经训练过的项目，测试的是训练后的结果，训练的模型还不具备这方面的知识。

本研究首先对小学语文教材中使用形声字的情况进行统计分析，并利用已有的"汉字字形属性数据库"(简称"汉字属性库")①，实现了对汉字字形的表征。自组织特征映射模型(self-organizing feature map)是典型的自组织系统，因而也有人称其为"自组织模型"(SOM, self-organizing map)，该模型采用无指导的学习算法(Kohonen 1989, 1995)，与自然语言获得过程有一定程度的相似性。我们选择了自组织模型成功模拟了儿童的汉字获得过程(Xing, Shu and Li 2004)，通过对训练的二年级、四年级和六年级模型的考察，显示了模型在命名形声字过程中的特点，模拟结果与儿童的行为实验结果基本一致，与数据库分析结果也很一致。模型对没有学习过的生字的命名所表现出的特点表明：模型确实获得了这方面的知识并能扩展到生字的命名任务中。

2. 基于小学汉字数据库的统计研究

本研究首先建立了由5套小学语文的课文文本构成的"小学生语文课本课文语料库"，(简称"小学语文课文库")总字数814 178。我们首先生成各套教材的汉字表，并统计各套教材各册汉字的使用次数，然后再将5套教材的字表合并，统计汉字的总频度，并且标注每个汉字在各套教材的使用次数，最后形成"小学语文教材课文汉字数据库"(简称"小学汉字库")。

2.1 "小学汉字库"的标注

我们对"小学汉字库"的类型进行了标注，标注方法基本沿用舒华等(1998)的方法，对部分数据的分类方法进行调整。具体的标注体系如下：

(1) 基本数据。主要包括汉字字形、汉字的读音，如果是形声字则标注该形声字的声旁字形和声旁读音，例如"清"是形声字，我们标注了它的读音，同时还标注了它的声旁"青"以及声旁的读音qing1。

(2) 声旁类型。标注了声旁读音与整字读音之间的关系。主要包括：1)整字与声旁读音完全相同，例如"清"和"青"；2)整字与声旁的声调不同，例如"请"和"青"；3)整字与声旁的韵母相同，例如"晴"和"青"；4)整字与声旁的声母相同，例如"沙"和"少"；5)整字与声旁读音完全不同，例如"的"和"勺"；6)非形声字，例如"休"；7)整字或声旁是多音字，例如"便"作为声旁所构成的形声字"鞭"，"便"是多音字；8)涉及古字，不易评定的形声字。

① 指国家语言文字工作委员会规范项目"信息处理用GB 13000.1字符集汉字部件规范"课题组建立的"GB 13000.1字符集"汉字部件拆分数据库，包括全部20 902个汉字的拆分序列等信息。

(3) 声旁一致性。是指在一定范围内，同一声旁构成的全部形声字的读音和声旁读音是不是完全一致，该数据库主要区分下列情况：1)含该字声旁的所有字(小学范围)读音相同；2)含该字声旁的其他字读音有多种；3)含该字声旁的字只有一个(小学范围)。

(4) 声旁是字。是指形声字的声旁在小学阶段所学的汉字中是不是能够独立成字，包括：1)该字的声旁独立成字；2)该字的声旁不是独立字。

(5) 声旁位置。是指声旁在形声字中所处的位置，主要包括以下七种：1)声旁在形声字的右边；2)声旁在形声字的左边；3)声旁在形声字的上部；4)声旁在形声字的下部；5)声旁在形声字内部；6)声旁在形声字外部；7)其他情况，例如半包围结构等。

(6) 声旁位置是否固定。是指在小学汉字范围内全部由该声旁构成的形声字中，该声旁位置是总是固定的，还是可以变化的，包括：1)声旁位置固定；2)声旁位置可变。

2.2 "小学汉字库"中形声字的分布

我们选择了五套教材中的一套北京地区使用的教材(简称"北京教材")的汉字作为统计对象，对教材中形声字的分布和表音情况进行了分析。我们将各年级中新出现的形声字叫做新形声字。"北京教材"的用字数是3 306个，其中形声字2 475个，占总字数的75%，非形声字只有831个。首先我们统计了各年级所学的形声字的累计数占该年级全部汉字的比例和新形声字的数量以及占全部生字的比例。表1显示了各个年级的形声字的数量以及占该年级所学总字数的比例。

年级	一	二	三	四	五	六
总数量	400	878	1464	1840	2189	2475
比例	.60	.64	.69	.72	.74	.75
新形声字	400	479	586	376	349	286
比例	.60	.67	.77	.85	.85	.86

表1 各年级形声字累计数和新形声字数

从统计的结果来看，小学阶段新形声字的数量在三年级以前有一定的增长，到了三年级以后，新形声字的数量逐渐下降。总体上看，新形声字的数量随着年级的升高有下降的趋势，但是，小学阶段各个年级形声字总数占总字数的累计比例随着年级的升高增加很快，小学生各个年级学习的新形声字的比例也随年级的升高而增加。在变化趋势上，在三年级以前，形声字的累计数量及占总字数的比例上升比较快，三年级以后变化相对平稳。这两个增加的趋势表明了小学生随着年级的升高，形声字会起到越来越重要的作用。各年级形声字总数和新形声字数量的差距越来越大，可见各年级重复使用的形声字也越来越多。

我们进一步按照每个汉字的使用次数的多少将"北京教材"的全部汉字分成五个频度等级(简称"频级")。这5个频级是：高频(大于等于50次)、次高频(20到49次)、中频(10到19次)、次低频(3到9次)和低频(小于等于2次)。我们对形声字和非形声字在各个频级的分布情况进行了统计，表2显示了各个频级的形声字和非形声字的比例。

从表2的数据可以看出，形声字和非形声字的比例在不同频度等级中的变化非常明显，

高频字中形声字只占51%。而在低频字中形声字的比例约占将近90%，这说明形声字在不同频度等级的分布是不均衡的。

	高频	次高频	中频	次低频	低频
形声字	.51	.65	.75	.85	.89
非形声字	.49	.35	.25	.13	.11

表2　各个频级形声字和非形声字比例

我们还统计了形声字的家族分布。形声字的家族是指同一声旁构成的一组形声字的集合，如果声旁独立成字，而且已经学习过，那么这个家族也包括该声旁。例如声旁"皇"以及它构成的"煌、凰、蝗、惶"等形声字就构成了一个家族。在"北京教材"中，形声字家族成员数最少的是1个，最多的是17个，各个年级的最大成员数也不同，一年级是7个，六年级是17个。表3显示了各个年级的形声字家族总数以及平均成员数。

年级	一	二	三	四	五	六
家族数	123	287	478	566	632	687
最大成员数	7	9	12	12	14	17
平均成员数	2.49	2.97	3.23	3.49	3.76	3.94

表3　各年级形声字家族数及平均成员数(Xing, Shu, and Li 2004:13, Table 4)

结果显示，随着年级的升高，小学生各个年级使用的形声字家族数逐渐增大，家庭成员数逐渐增加。家族数和家族成员数的变化应该会对儿童形声字的习得产生很重要的影响，但是这种影响并不是简单的叠加效应，因为形声字的表音特点、形声字的家族一致性等都会对小学生形声字的加工产生影响。

2.3 "小学汉字库"中形声字的表音特点

我们将形声字按照表音的情况分为三类：规则字、半规则字和不规则字，规则字是指整字和声旁的读音声母、韵母和声调都完全相同，半规则字是指声旁和整字读音部分相同，不规则字是指除此以外的其他形声字。我们首先分析了各个年级的形声字的表音情况。表4显示了各个年级新形声字和总字数中规则字、半规则字和不规则字占形声字的比例。

年级		一	二	三	四	五	六
新形声字	规则字	.14	.26	.22	.29	.30	.34
	半规则字	.43	.39	.45	.42	.39	.35
	不规则字	.43	.35	.33	.29	.31	.31
累计形声字	规则字	.14	.21	.21	.23	.24	.25
	半规则字	.43	.40	.42	.42	.42	.41
	不规则字	.43	.39	.37	.35	.34	.34

表4　各年级新形声字和总形声字规则性情况

从上表的数据可以看出，无论是新形声字还是累计形声字，各个年级的规则字都是呈现上升的趋势，这说明随着年级的升高，他们所学习的新形声字的表音功能加强，六年级达到

全部新学习的形声字的34%。新规则字要比累计规则字增长快。而从各年级累计形声字来看，二年级以后虽然是增加的趋势，但是变化的幅度不大。在新形声字中，各个年级的半规则字的比例有逐渐下降的趋势，而在累计字中基本保持一定的比例。不规则字都是随着年级的升高而呈现下降的趋势。

我们进一步分析了形声字的规则性和字频的关系。表5显示了规则字、半规则字和不规则字在各个频级累计字数中所占的比例。

	高频	次高频	中频	次低频	低频
规则字	.13	.21	.23	.27	.31
半规则字	.41	.41	.44	.43	.38
不规则字	.46	.39	.33	.30	.31

表5　各个频级形声字的表音情况

结果显示，在小学阶段，规则字随着频度的降低逐渐增加，半规则字则随着频度的降低而呈现升高的趋势，而不规则字在各个频级的变化不大。根据这个结果我们可以得出下面的结论：在小学阶段，形声字的表音性和频度存在交互作用，低频字的规则性比高频字更强。

3. 自组织模型

依据大脑对信号处理的特点，Kohonen（1982，1989，1995）提出了一种神经网络模型——"自组织特征映射模型"(SOFM)。自组织过程实际上就是一种无指导的学习，自组织网络从一组表征数据中获取有意义的特征或者一些内在的规律性，所以更加接近生物神经系统。最早采用自组织模型研究语言的是Ritter和Kohonen(1989)，他们将一些动物的语义特征组织在特征网图上，经过2000次的自组织训练以后，比如野生食肉动物和鸟类被分别组织在网图的不同区域，形成不同的组，在同一组中，相同的动物更加靠近。Miikkulainen(1993，1997)的研究在利用SOFM模型研究自然语言处理领域起到重要的作用。他提出了一个关于记忆和自然语言处理的综合模型，其中的一个分支模型DISLEX，就是词汇加工的模型。李平等(Li 2003；Li and Farkas 2002；Li，Farkas and MacWhinney 2004；Li，Zhao and MacWhinney 2007)在此基础上对SOM在语言习得上的应用作了大量的研究。他们发现如果儿童在理解语言时分析成人话语中词与词的共现关系及其频率，可以获得词的语义及语法关系(Li，Burgess and Lund 2000)。从已有的研究结果来看，自组织理论是适合语言获得研究的。

3.1 模型的结构

本模拟研究是将DISLEX进行改造以后直接用于汉语的研究。DISLEX(见图1)包括两个主要部分：在不同的输入输出特征中的词的形态(lexical symbol，包括词形和语音)词典；词汇语义词典。词的形态和语义记忆是通过特征映射图(feature maps，简称"网图")来实现的。其中一个网图是输入和输出词的形态，另一个网图表示语义词典。网图中每个单位表示一个词(包括形态和语义)，通过下面两种方法来表示：(1)每个单位有一个内部参数向量，也叫权重向量，这个向量存储词的分布表征。(2)每个单位就是一个词在网图中的本地表征。网

图采用二维网络来安排分布特征，以便词之间的相似性变得更清晰。

我们对 DISLEX 的用途作了扩充，将汉字字形和字音之间建立了联系，用以模拟汉字形音之间的关系。

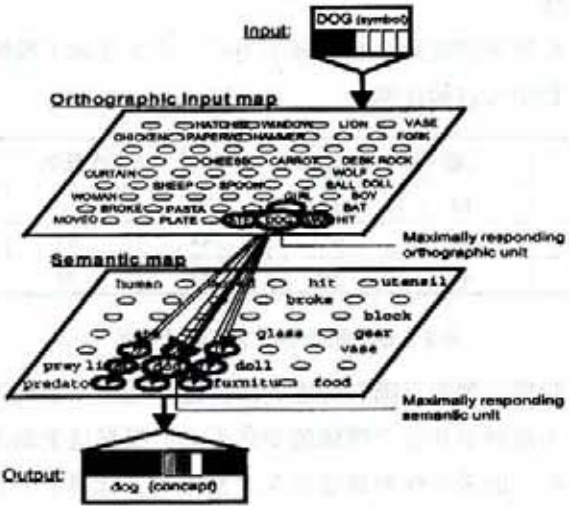

图 1　输入输出激活情况(Miikkulainen 1997)
(图中以英文 dog 为例，介绍了语音到语义的输出过程，在语义词典中，
语义 dog 和它邻近的概念都被激活，但是 dog 的激活程度最高)

3.2 汉字的表征

3.2.1 字音表征

		1	2	3	4	5	6	7
响亮度	描述	元音	浊音	清音				
	特征值	0.100	0.750	1.000				
发音部位	描述	双唇	唇齿	舌尖前	舌尖中	舌尖后	舌面	舌根
	特征值	0.143	0.286	0.429	0.572	0.715	0.858	1.000
发音方法	描述	闭唇	鼻音	塞音	擦音	塞擦音	卷舌	边音
	特征值	0.143	0.286	0.429	0.572	0.715	0.858	1.000
舌位高低	描述	高	中	低				
	特征值	0.333	0.666	1.000				
舌位前后	描述	前	央	中后	后			
	特征值	0.250	0.500	0.750	1.000			

表 6　汉语音素表征表

汉字的语音表征我们参照 Li 和 MacWhinney(2002)提出的 PatPho，这种方法对英语的语素进行三个维度的特征描写，再将这三个维度进行等距离的数字表征。我们将汉字的音节按照其结构分为 6 个槽：声母部分包含一个槽，韵母部分包含四个槽，其中韵头一个槽，韵腹包含两个槽，韵尾包含一个槽，声调部分一个槽，将全部音节按照槽来排列音素。再将每个音素通过 5 个维度来表征，这 5 个维度分别是响亮度、发音部位、发音方法、舌位高低、

舌位前后等，每个维度又分别设有3到7个特征值，见表6。

如果某个槽没有音素，比如没有介音的音节，我们就将该槽的5个维度全部赋值为“0.000”，这样每个音节则由30个特征值来表示。

3.2.2 字形表征

3.2.2.1 汉字字形表征

我们从汉字的整字结构、汉字的部件数量和汉字部件特征3个大的部分来对汉字字形进行表征。下面分别叙述。

（1）汉字的整字结构。通常来说，汉字的结构主要有12种，加上独体字共13种，这13种结构是：左右结构、上下结构、右上包围结构、左上包围结构、左下包围结构、上三包围结构、下三包围结构、左三包围结构、全包围结构、框架结构、左中右结构、上中下结构、独体结构。我们对《现代汉语常用字表》的3500常用字的结构进行统计，发现左右结构占58.5%，上下结构占25%，独体结构占5.3%，其他结构所占的比例很少，各类包围结构合起来占9.2%。因此，我们将汉字的结构分成6大类进行表征，并对各个类型进行赋值，赋值的方法采用等分取值方法，就是将0—1分成6个等分，每个特征取0—1之间的一个特征值，6类结构以及它们的赋值情况分别是：独体结构(1.000)、左右、左中右结构(0.833)、上下结构(0.667)、上中下结构(0.500)、包围结构(0.333)、框架结构(0.167)。

（2）汉字的部件数。汉字的部件数是指汉字中包含的基础部件数量，我们首先统计了“小学汉字”中每个汉字的部件数，其中部件最多的汉字有7个部件。因此，特征值分别为单部件字(0.143)、两部件字(0.286)……。

（3）部件的视觉特征。部件的视觉特征通过部件结构、部件的笔画关系、部件的形状、部件的笔画等特征来描述。部件结构是指基础部件的结构类型，例如“木”是独体结构，“革”是上下结构，“非”是左右结构等，我们采用和整字结构相同的表征方法对部件结构进行了表征。部件的笔画关系可以分为5种。相离：两个笔画之间没有交汇点，例如“八”；相交：两个笔画之间有交叉点(例如“乂”)；相接：两个笔画之间有一个连接点(例如“丄”)；交接：部件中的笔画包含相交和相接关系，例如“土、才”等；交离：部件中的笔画包含相交和相离关系，例如“肉、书”等；另外，单笔部件(例如“一”“丶”等)，没有上述特征，我们把单笔部件单独作为一种属性。我们同样采用6等分赋值的方法对部件的笔画关系进行了赋值。部件的形状特征归纳为36组，每组又包含1到10个不等的特征，比如汉字部件中类似于“口”这样的方框结构，我们都归入一组，因此，“口、冂”和“田、日、曰、中、四”等字中的外框都在一组，并按照相似性和差别分别取0到1之间的等距离值。部件的笔画是从笔画形状的角度来表征的，如果一个部件少于4画，则按顺序排列，空的位置值为0.000，如果一个部件包含四个以上的笔形，则取该部件的首笔、第二笔、倒数第二笔和末笔的笔形。部件按照“横、竖、撇、点、折”的顺序采用等分取值的方法分别取1到0之间的平均值。

（4）部件的构字位置。这部分属性通过部件的构字位置和部件的排列方式来分析。汉字的部件构字时，常常具有固定的位置，我们按照一个部件构字的主要位置，对部件的构字位置特征进行描述，这些特征共有11种，分别是：独立成字的部件(如“日”)、只是笔画的部

件(如“一”)、部件常常在字的左边(如“亻”)、部件常常在字的右边(如“刂”)、部件常常在字的上边(如“⺍”)、部件常常在字的下边(如“廾”，“弄”字底)，部件常常是一个字的外框(如“冂”，“同”字框)、部件常常在汉字的腰部(如“⺘”，“变”字腰)、部件常常在汉字的内部(如“⺹”，“庚”字心)、部件常常在汉字的一角(如“夕”称为“餐”字角)。同样采取等分取值的方法进行赋值。

(5) 部件的排列方式。我们采用了分槽的表征方法对汉字部件进行排列，就是将汉字部件按照三个槽排列，第一个槽从部件1的位置开始，按顺序向后排列，第二个槽从部件4开始，由向后到向前向两个方向扩展，第三个槽从部件7开始，逐渐向前扩展。表7的例字说明了各个类型的部件排列情况。

例字	结构	部件1	部件2	部件3	部件4	部件5	部件6	部件7
为	独体				为			
副	左右	一	口	田				刂
斑	左中右	王			文			王
盒	上下	人	一	口				皿
曼	上中下	日			冖			又
蓓	上下	艹				亻	立	口

表7 汉字部件排列方式例字表(Xing, Shu and Li 2004:48)

3.2.2.2 形声字字形表征

我们将形声字的整字读音信息和声旁读音信息加入到形声字的字形表征中。如图2所示，整个形声字的形表征由字形表征加上声旁读音、整字读音组成，其中声旁和整字读音各30个特征值，而字形表征则又由部件数、整字结构、以及7个部件的依次拆分组成，每个部件又由部件结构、部件位置、笔画数、视觉特征、笔画关系以及笔形共44个维度取值组成，这样，整个形声字就有322个字形特征值，每个形声字共有382个特征值。

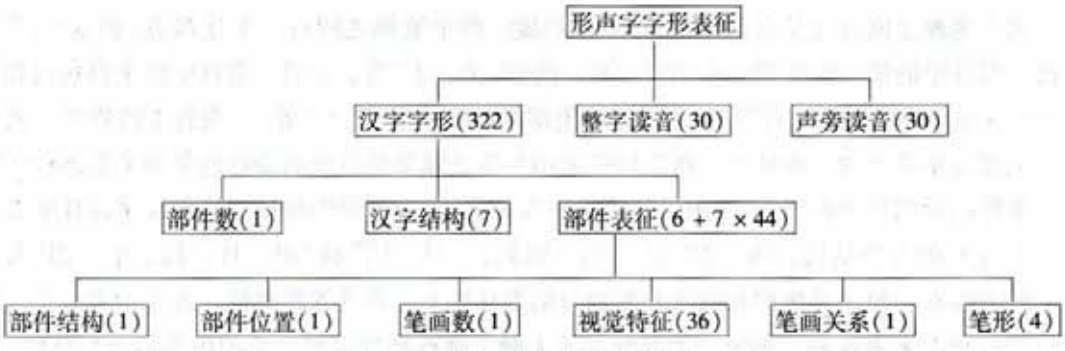

图2 形声字的字形表征结构图

4. 儿童汉字获得的自组织模型

4.1 模型的训练

训练样本的抽取：由于受到屏幕显示的限制，每个年级模型训练字限定在260字左右，按照下列原则抽取：(1)按照不同成员数家族比例抽取用于训练的家族，二年级、四年级、

六年级各个年级及以前出现的形声字和声旁的总数分别是1349、2537、3270个，抽取的训练形声字分别是该年级总数的18.91%、10.07%和7.81%；(2)抽取是以各个家族成员数量的大小为单位的，比如二年级两个成员的家族数148个，按照18.91%的比例，我们随机抽取了28个家族56个字，其他家族也是这样抽取的。(3)被抽取的家族包含该年级及以前学习的该家族的全部形声字；(4)家族的选取是完全随机的。我们以二年级的训练字为例。

家族成员	总字数	家族数	选取家族数	选取字数	例字
1	505	505	95	95	脑、部、办、晨
2	296	148	28	56	帮绑、风枫
3	255	85	16	48	共哄拱、峰蜂烽
4	88	22	4	16	忘忙芒茫、此些龇紫
5	95	19	3	15	胡湖蝴糊葫、白怕拍伯柏
6	72	12	2	12	干岸杆赶汗竿、方放房坊旁芳
7	21	3	1	7	挂蛙街鞋洼洼娃
8	8	1	1	8	青请情清晴蜻精睛
9	9	1	0	0	
合计	1349	796	150	257	

表8　二年级训练字的选取情况表

训练模型：每个字按其实际使用频率成比例训练，公式是“训练次数＝使用频率×训练周期”，因模型训练300次以上就基本学会该字，因而对于频率大于18次的就按18计算，这样每个字最少训练20次，最多训练360次。训练完成之后，字形网图、字音网图以及二者之间的联结正确率都大于70%。每个年级训练10个随机模型，相当于训练10个儿童，3个年级共训练30个模型。

4.2 模型的测试

4.2.1 测试材料

对训练后的模型从三个因素加入测试：3(二、四、六年级)×3(高频、低频、生字)×2(规则字、不规则字)。每个年级分别选择60个形声字作为测试项目，每种条件下10个测试项目。各年级的高低频率为相对高低频：二年级的高频字为在教材中出现次数大于5次的汉字，低频字为小于2次的汉字；四年级的高频字为大于8次，低频字为小于5次的汉字；六年级的高频字为大于18次，低频为小于8次的汉字。每个年级的生字是指在这个年级的模型中没有训练过的但是其家族成员已经得到过训练的汉字。测试项目的规则字指的是声旁读音与整字读音完全相同的形声字，不规则字指这个字的声旁读音与整字读音的声母、韵母以及音调都不相同的形声字。例如二年级的测试字分为6类：(1)高频规则字，如“帮、逗”等；(2)高频不规则字，如“岸、苦”等；(3)低频规则字，如“枫、绑”等；(4)低频不规则字，如“哄、蹄”等；(5)规则新字，如“犍、邱”等；(6)不规则新字，如“恍、驯”等。

4.2.2 测试方法

将测试形声字的字形表征与字音表征分别加入模型的两个表征文件中，测试的汉字对加入字对文件中，并对测试汉字加上标记，然后运行DISLEX的测试模块，模型运行完成之后

会在字形和字音网图上显示出所有汉字，点击字形网图上的测试汉字，模型会给出相对应的这个测试汉字的字形输出与字音输出。因在网图上的显示效果问题，对60个测试项目分6次测试，每次测试10个项目，三个模型共测试18次。

5. 模拟结果

声旁完全标音的形声字叫做规则字，声旁部分标音和不标音的形声字叫做不规则字，这种由于规则或不规则而造成的对汉字认读加工的影响叫做规则性效应(regularity effect)。我们对每个模型进行测试，只有当这个测试字的字形、字音输出都正确才计为读音正确，三个年级的平均命名正确率如表9：

规则性	频率	年级		
		二	四	六
规则	高	0.97	0.97	0.99
	低	0.33	0.73	0.89
	生	0.00	0.00	0.00
不规则	高	0.82	0.76	0.89
	低	0.23	0.56	0.66
	生	0.00	0.00	0.02

表9 模型在不同条件下的平均命名正确率

对测试结果进行三因素的方差分析，结果表明：年级的主效应显著，$F_1(2,27)=130.83$，$P=0.000$，$F_2(2,108)=10.99$，$P=0.000$，多重比较表明二、四、六年级的命名正确率依次上升($F<0.01$)；规则性的主效应显著$F_1(1,27)=94.82$，$P=0.000$，$F_2(1,108)=9.85$，$P=0.002$，表明规则字的命名正确率显著高于不规则字；频率的主效应显著$F_1(1,27)=397.30$，$P=0.000$，$F_2(1,108)=48.08$，$P=0.000$，高频字的正确率显著优于低频字。

年级与频率的交互作用显著，$F_1(2,27)=68.02$，$P=0.000$，$F_2(2,108)=8.73$，$P=0.000$，低频字从二年级到四年级的命名正确率迅速上升，而四年级到六年级的差距较小，表明由二年级到四年级是形声字获得的迅速发展的时期。

		规则	不规则	差值
高频	二年级	0.97	0.82	0.15
	四年级	0.97	0.76	0.21
	六年级	0.99	0.89	0.10
低频	二年级	0.33	0.23	0.10
	四年级	0.73	0.56	0.17
	六年级	0.89	0.66	0.23
生字	二年级	0.00	0.00	0.00
	四年级	0.00	0.00	0.00
	六年级	0.00	0.02	-0.02

表10 模型命名形声字在不同频率和年级条件下的规则性效应

年级与频率、规则性的三次交互作用显著，$F_2(2,27)=3.60$，$P=0.041$，$F_2(2,27)=$

.44，P=0.645，如表10所示，规则字与不规则字的命名正确率的差值表明了规则性效应程度，在高频字中，随着年级的上升，在四年级效应量最大，在低频字中，随着年级的上升，规则性效应量逐渐增大。而在生字中没有表现出规则性效应。

我们进一步对模型命名的错误模式做了分析，方法与对儿童命名形声字的错误分析相同。如果测试字为"猜"，模型字形输出为"请"，字音输出为"qing3"，则输出的错误类型为类推错误，对应于儿童使用与测试字相同家族的其他字来给测试字命名。如果模型字音输出为"qing1"，字形输出为"青"，则为声旁错误，对应于儿童用形声字的声旁来命名由该声旁构成的形声字。这样，对模型的命名错误模式进行分析（各种错误类型所占的比例见表11），模型在命名生字时，犯类推错误的比例随年级的上升逐渐增多，其他错误类型逐渐减少。

年级	声旁错误	类推错误	其他错误
二年级	0.06	0.45	0.50
四年级	0.02	0.68	0.31
六年级	0.00	0.72	0.28

表11　模型命名形声字的错误类型比例

6. 讨论

6.1 语言材料库的建设及其在模型中的应用

随着研究者对语言材料在语言习得研究中作用的认识的变化，语言材料库的建设、加工和统计研究日益受到人们的重视。近年来，越来越多的语言材料库被建立起来，例如英语儿童语言材料库CHILDES（MacWhinney 1991，1995）。这些数据一方面被大量应用在儿童语言发展的研究中，许多研究者认为（Derwing and Baker 1979；Nagy and Anderson 1984；舒华等 1998）语言材料库的分析为我们了解儿童语言和阅读获得提供了一条很好的途径。另一方面，也推动了从语言材料中提取知识的计算机模拟研究。很多学者已经对从语言材料中提取知识进行了理论探讨。Burgess 和 Lund（1997，1999）提出了HAL（hyperspace analogue to language）的理论，认为自然语言材料中词与词之间的关系提供了足够的语义信息。Landauer 和 Dumais（1997）也提出了类似的理论（latent semantic analysis），认为语义可以从词与篇章的关系中提取。这些理论已经在计算机模拟研究中得到应用，研究者通过模型学习大量自然语料，使模型获得语义、语法知识。例如词汇的发展模型——DevLex（Farkas and Li 2002），这个模型不限于固定的词汇学习，而是通过语料的增加而相应地增加新词，并可以不断增加网络中的单元数目及网图数目，这种逐步增加的过程可以更适当地反映儿童语言学习或成人外语学习的过程（李平 2002）。这个模型已经在中英双语的词汇表征中得到应用并取得很好的效果（Li and Farkas 2002）。目前，在英语模型研究中，大规模语料库主要应用于句法、语义特征的获得方面的计算机模型中，语料库途径的模拟使模型有可能突破人工建设句法、语义等知识表征的局限性。然而，通过语料库途径模拟形、音知识表征的模型还很少见。

本研究的目的是模拟儿童形声字的命名，我们建立了"小学语文课文库"和"汉字属性库"。前者为我们的模拟提供了形声字的分布和形声字的表音规律及其随着年级的升高而出

现的发展变化情况，后者提供了汉字的笔画、部件和结构类型、部件的构型特征等汉字字形特征。可以说，我们建设的"小学语文课文库"和"小学汉字库"为儿童语言习得的行为研究和计算机模拟提供了很多基本统计数据，为进一步的研究提供了很好的基础。本研究中，我们首先在建立模型表征上使用了来自数据库的统计信息，比如结构类型的频度、部件的频度等，为汉字字形表征的建立奠定了基础。在模型的训练上，我们也使用了基于语料库的统计信息，比如小学各年级不同家族成员数的家族的比例、各个形声字的使用频度的数据。从模型命名形声字的结果来看，本研究不仅进一步证实了语言材料和语言习得关系的假设，而且更进一步将语言材料中的一些非同现规律直接运用到计算机模型中，并且获得了成功，这将模型从语言材料中获得知识的方法在句法、语义知识获取(Li and Farkas 2002；Li，Farkas and MacWhinney 2004)的基础上更推进了一步。

6.2 汉字的字形表征

英语词汇中形的表征相对简单，比如计算英文字母点阵中黑白点的数量，并给每个字母取值，每个词形的表征来自字母的取值(Miikkulainen 1997)。汉字字形的表征，是汉字计算机模拟研究中必须解决的难点，也是关键。如何将汉字的构成成分和汉字的正字法规则通过数值的方法进行表征，这是汉字计算机模拟研究必须解决的问题。理论上说，汉字字形中包含的各种特征都会在儿童的心理特点中得到表征，这主要包括汉字的结构成分(包括部件、笔形等)和构造规则(结构类型、位置特点、使用频度等)，这方面的研究很少(陈鹰、彭聃龄 1994)。汉字字形表征的理想结果就是对汉字的构成成分特征和各种正字法规则进行表征。本研究对汉字字形特征进行了大量的分析研究，首先利用"汉字属性库"，对汉字的结构、部件、笔画等进行了标注，还建立了汉字的部件数据库，对汉字部件的结构、笔画、形状、位置等特征进行了标注，并进行了统计，根据统计结果对汉字字形进行表征，这种表征方法使得模型获得了字形特征，比如结构特征、相似性特征、部件数特征、笔画数特征等，这为汉字计算机模拟的进一步研究打下了基础。

形声字的表征方法是一个值得探讨的问题，但是这方面的研究还没有人涉及。我们从行为实验的研究结果(武宁宁、舒华 1999；Zhou and Marslen-Wilson 1999a，1999b；周晓林等 2000；杨珲等 2000)中得到启示：形声字的加工过程中，存在着亚词汇的加工过程。周晓林等(2000)认为无论是高频字还是低频字，亚词汇信息都得到激活，只是高频字的语音信息更容易激活，词汇和亚词汇之间的相互竞争使得亚词汇的语音激活效应很难表现出来。对低频字来说，整字的语音信息激活较慢，语言词汇的语音信息的竞争能力较弱，使亚词汇语音信息有机会表现出来。杨珲等(2000)的研究结果得出一些结论：整字与声旁的语音激活发生在汉字加工的早期阶段；语音的激活发生在高频字和低频字中；整字和声旁的语音存在着交互作用，影响这个交互作用的是整字和声旁的相对频率。

本研究对形声字的表征方法是将整字的读音和声旁的读音特征加入形声字的表征中，因为在熟悉声旁的情况下，儿童命名形声字的时候会激活声旁的语音信息，在测试模型的时候，我们去掉了测试字中的整字读音，这些方法在计算机模拟领域还没有人尝试过。

6.3 模型的知识获得和知识输出

计算机模拟的目的是要模拟人脑的信息加工过程。人脑之所以能够进行信息加工，和人的心理词典中储存的知识有关，计算机要能够模拟人脑的加工过程，就必须具备人脑的知识，因此，模型的知识获取就成为计算机模拟研究中的重要因素。就目前的研究来看，模型获得知识主要通过表征和训练。按照联结主义的分布表征理论，知识的存储特点是相似和差异的共存，两个学习的对象差异越大，共性越小，激活的共同特征越少，反之差异越小，共性越大，激活的共同特征越多，因此，表征在联结主义理论中占有重要地位。对表征方法做了大量研究，比如基于分析的方法(Ritter and Kohonen 1989；Miikkulainen 1997)和基于统计的方法(Li and Farkas 2002；Li，Farkas and MacWhinney 2004)。有了很好的表征以后，并不是说模型就能够获得知识。儿童的语言规则是逐步获得的，这是儿童接触到越来越多的语言材料以后，从语言材料中概括出来的。前面我们讨论过模型的学习规则，比如有指导的学习和无指导的学习，这只是学习的手段问题。其实，模型训练的关键不在于学习的手段，而是在于学习的结果。自然语言的统计规则是和儿童获得的语言规则高度相关的，因此，将语言材料的统计数据应用到模型的训练当中，才能真正模拟儿童语言获得的过程。

本研究不仅将语言材料的统计数据应用于汉字字形的表征中，而且将语言材料中的频度、分布规律、发展变化等属性直接应用于模型的训练，使得模型从表征和训练两个方面获得知识。从我们的研究结果来看，模型获得了和儿童非常一致的结果，这为汉语的计算机模拟找到了一条新的途径，为以后的汉语的计算机模拟研究打下了比较好的基础。

模型的学习过程就是模型通过表征和训练方法获得知识的过程，训练好的模型就获得了一定的知识，这些知识存储在模型的输入输出词典中。而测试模型实际上是模型的输出知识的过程。本研究通过生字(模型没有训练的字)判断的方法获得模型对生字的读音输出，实际上就是模型输出知识的过程。测试模型的另一个关键问题是测试项目的表征问题，因为模型就是从测试字的表征中获得测试字的特征，然后运用学习的知识，完成测试字的加工任务。本研究采用改变测试字的表征的方法对模型进行测试，这样就更好地模拟了模型运用学习的特征来对测试项目进行加工的过程。

7. 结论

本研究首先完成了汉字字音和字形的表征方法。从上面的研究结果我们得到以下启示：汉字字形表征的目的不仅仅是起到区别字形的作用，也不仅仅是字形的相似性问题，除了区别性和相似性以外，更重要的是如何通过表征的方法，使得模型获取和字形相关的知识，并能够在此基础上运用知识。语言材料的输入在儿童语言习得中起到非常重要的作用，儿童的语言规则的获得受到真实的语言材料的制约，儿童获得的语言规则也是对真实语言材料的反应，儿童语言习得过程近似于无指导的学习过程，因为教学只起促进作用，只要处在真实的语言环境中，儿童最终都能够获得语言规则。

本模拟成功地模拟了小学阶段二、四、六年级儿童的形声字学习过程，并且通过形声字命名的正确率和错误分析，得到了和儿童实验研究基本一致的结果。这些结果主要有：(1)模型较早就形成了规则性效应，模型命名规则形声字和不规则形声字时的正确率在二年级时

虽然有差异，但是不明显，到了四年级时比较明显，这说明四年级模型已经形成了比较好的规则性效应。(2)从年级的发展来看，规则字和不规则字在正确率上的差异随着年级的升高逐渐加大，这说明模型的规则性随着年级的升高而增加。(3)从模型使用的声旁线索来看，四年级使用的声旁线索比二年级和六年级高，说明四年级阶段出现了声旁过度规则化现象。(4)从规则性和频度的关系来看，高频条件下的规则字和不规则字的差别不明显，低频字差别明显；(5)对模型的错误模式分析表明，模型与儿童的表现一致，随着年级的升高，模型逐渐使用类推策略和读声旁的策略。

参考文献

Burgess, C. and K. Lund. 1997. Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes* 12,1 – 34.

———. 1999. The dynamics of meaning in memory. In E. Dietrich and A. Markman, eds., *Cognitive Dynamics: Conceptual and Representational Change in Humans and Machines*. Mahwah, NJ: Erlbaum. Pp. 17 – 56.

Derwing, B. and W. Baker. 1979. Resent research on the acquisition of English morphology. In P. Fletcher and M. Garman, eds., *Language Acquisition*. New York: Cambridge University Press. Pp. 209 – 23.

Farkas, I. and P. Li. 2002. DevLex: A self-organizing neural network model of the development of lexicon. In Wang L., J. C. Rajapakse, K. Fukushima, Soo-Young Lee and X. Yao, eds., *Proceedings of the 9th International Conference on Neural Information Processing*. Mahwah, NJ: Lawrence Erlbaum. Pp. 2546 – 51.

Harm, M. and M. S. Seidenberg. 1999. Reading acquisition, phonology, and dyslexia: Insights from a connectionist model. *Psychological Review* 106,491 – 528.

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43,59 – 69.

———. 1989. *Self-organization and Associative Memory*. Heidelberg: Springer-Verlag.

———. 1995. *Self-organizing Maps*. Heidelberg: Springer-Verlag.

Landauer, T. and S. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104,211 – 40.

Li, P. 2003. Language acquisition in a self-organizing neural network model. In P. Quinlan, ed., *Connectionist Models of Development: Developmental Processes in Real and Artificial Neural Networks*. Hove / New York: Psychology Press. Pp. 115 – 49.

Li, P. and B. MacWhinney. 2002. PatPho: A phonological pattern generator for neural networks. *Behavior Research Methods, Instruments, and Computers* 34,408 – 15.

Li, P., C. Burgess and K. Lund. 2000. The acquisition of word meaning through global lexical co-occurrences. In E. V. Clark, ed., *Proceedings of the Thirtieth Stanford Child Language Research Forum*. Stanford, CA: Center for the Study of Language and Information. Pp. 167 – 78.

Li, P. and I. Farkas. 2002. A self-organizing connectionist model of bilingual processing. In R. Heredia and J. Altarriba, eds., *Bilingual Sentence Processing*. North-Holland: Elsevier Science Publisher. Pp. 59 – 85.

Li, P., I. Farkas and B. MacWhinney. 2004. Early lexical development in a self-organizing neural networks. *Neural Networks* 17,1345 – 62.

Li, P., X. Zhao and B. MacWhinney. 2007. Dynamic self-organization and early lexical development in children. *Cognitive Science* 31,1 – 32.

MacWhinney, B. 1991. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum.

———. 1995. *The CHILDES Project: Tools for Analyzing Talk*. 2nd edition. Mahwah, NJ: Lawrence Erlbaum.

Miikkulainen, R. 1993. *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory*. Cambridge, MA: MIT Press.

———. 1997. Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and Language* 59,2:334 – 66.

Nagy, W. E. and R. C. Anderson. 1984. How many words are there in printed school English? *Reading Research Quarterly* 19,304 – 30.

Plaut, D. C., J. L. McClelland, M. S. Seidenberg and K. Patterson. 1996. Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* 103,1:56 – 115.

Peng, D. L., H. Yang and Y. Chen. 1994. Consistency and phonetic independency effects in naming task of Chinese phonograms. In Q. C. Jing, H. C. Zhang and D. L. Peng, eds., *Information Processing of Chinese Language*. Beijing: Beijing Normal University Press.

Ritter, H. and T. Kohonen. 1989. Self-organizing semantic maps. *Biological Cybernetics* 61,4:241 – 54.

Wu, N., Zhou X. and Shu H. 1999. Sublexical processing in reading Chinese: A developmental study. *Language and Cognitive Processes* 14,503 – 24.

Xing, H., H. Shu and P. Li. 2004. The acquisition of Chinese characters: Corpus analyses and connectionist simulations. *Journal of Cognitive Science* 5:1 – 49.

Zhou X. and W. Marslen-Wilson. 1999a. Sublexical processing in reading Chinese. In J. Wang, A. Inhoff, H. C. Chen, eds., *Reading Chinese Script: A Cognitive Analysis*. Mahwah, NJ: Lawrence Erlbaum. Pp. 37 – 63.

———. 1999b. The nature of sublexical processing in reading Chinese characters. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 25,819 – 37.

陈 鹰、彭聃龄，1994，汉字识别和命名的联结主义模型，见 Hsing-Wu Chang, Jong-Tsun Huang, Chih-Wei Hue and Ovid J. L. Tzeng 编，*Selected Writing from the Sixth International Symposium on Cognitive of the Chinese Language*. 211 – 39 页。

李 平，2002，语言习得的联结主义模式。《当代语言学》第 3 期，164 – 75 页。

舒 华、武宁宁、郑先禹、周晓林，1998，小学汉字形声字表音特点及其分布的研究。《语言文字应用》第 2 期，63 – 8 页。

武宁宁、舒华，1999，汉字亚词汇加工研究。《心理科学》第 6 期，537 – 41 页。

杨 珲、彭聃龄、Charles A. Perfetti，谭力海，2000，汉字阅读中语音的通达与表征(I)——字水平的语音与亚字水平的语音及其交互作用。《心理学报》第 2 期，144 – 51 页。

周晓林、鲁学明、舒华，2000，亚词汇水平加工的本质：形旁的语音激活。《心理学报》第 1 期，20 – 4 页。

张亚旭、周晓林、舒华、邢红兵，2003，汉字识别中声旁与整字语音激活的相对优势。《北京大学学报(自然科学版)》第 39 卷，第 1 期，126 – 33 页。

第一作者简介

邢红兵，男，博士，北京语言大学对外汉语研究中心教授。研究兴趣：语言习得研究、汉语认知研究、语料库语言学、计算语言学。代表作：《基于统计的汉语字词研究》，“汉语词语重叠结构统计分析”。电子邮件：xinghb@blcu.edu.cn

XING Hongbing, male, Ph. D., is a professor at the Centre for Studies of Chinese as a Second Language, Beijing Language and Culture University. His research interest includes language acquisition, cognitive research of Chinese language, corpus linguistics and computational linguistics. His major publications are: *Research on Chinese Words and Characters Based on Statistics*, “Statistic analysis on reduplication of modern Chinese words”. E-mail: xinghb@blcu.edu.cn

作者通讯地址：邢红兵 100083 北京语言大学对外汉语研究中心
舒 华 100875 北京师范大学认知神经科学与学习研究所
李 平 University of Richmond, Richmond, Virginia 23173, U.S.A.