小学儿童汉字学习的计算机模拟

王小娟[1]，舒　华[1]，杨剑峰[2]

（1. 北京师范大学认知神经科学与学习国家重点实验室，北京　100875；
2. 中国科学院心理研究所行为科学重点实验室，北京　100101）

摘要：采用联结主义模型，使用小学儿童的真实文本语料分别训练二、四、六年级模型，研究儿童的阅读发展。模型能正确命名几乎全部训练汉字，并成功模拟出小学儿童规则性意识先发展，一致性意识后发展的结果。模拟结果与语料库分析以及儿童的行为实验结果一致，表明在真实的语料输入下，联结主义模型能成功模拟出儿童的汉字阅读发展过程，揭示阅读加工与发展具有相同的内部机制，而在阅读发展的不同阶段，输入语料的属性差异决定了最终阅读加工的外在行为表现。

关键词：心理语言学；阅读发展；联结主义；规则性；一致性
中图分类号：B842.5　　**文献标识码**：A　　**文章编号**：1674-2850(2011)03-0469-6

A computer simulation on Chinese characters learning of primary school children

WANG Xiaojuan[1], SHU Hua[1], YANG Jianfeng[2]

(1. *State Key laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing* 100875, *China*;
2. *Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing* 100101, *China*)

Abstract: Connectionist models were trained by real corpus from primary textbook to capture 2nd, 4th and 6th grade's reading performance to study the reading development of children. After training, the regularity effect was found in all three grades models, but the consistency effect just showed difference in 4th grade and reached significant in 6th grade. The result was consistent both with corpus analysis and children's development. It revealed a general mechanism both for skilled reading and its development, in which the difference of reading in different stage came from the statistic learning on the properties of input corpus. It also suggested that skilled reading was the stable stage of reading development.

Key words: psycholinguistics; reading development; connectionist; regularity; consistency

0　引言

长期以来，致力于阅读加工和阅读获得的研究各自关注不同的问题，没有相互结合。最新的联结主义把二者结合起来，认为阅读加工是阅读发展的结果，阅读获得是对大量输入语料统计学习并从中抽取形-音对应规则的过程，而获得过程会受各种语言与非语言线索的影响。不同正字法深度[1]的文字系统，因为其形-音对应粒度单位[2~3]不同，学习者采用的学习策略以及教学中的教学方法就相应不同，从而最终形成不同的阅读加工模式。

基于联结主义的计算机模型在英语词汇阅读中得到了广泛应用，并模拟和解释了大量成人的阅读

基金项目：国家自然科学基金（30870758）；教育部长江学者与创新团队发展计划（IRT0710）；北京市自然科学基金（7092051）；中国科学院心理研究所青年人才启动基金（Y0CX122S01）
作者简介：王小娟（1974—），女，博士研究生，主要研究方向：阅读加工的认知神经机制
通信联系人：杨剑峰，助理研究员，主要研究方向：视觉词汇识别、情绪韵律加工的认知神经机制. E-mail：yangjf@psych.ac.cn

现象[4~6]，对发展性阅读障碍[7]以及获得性阅读障碍[8]的形成原因都作了相应的理论解释。在充分考虑教学因素时[9~10]，模型能在一定程度上模拟出阅读发展。但这些模型具有很大的人为性，训练的是成人语料，不能真实反映出儿童阅读发展随语料输入不同的变化结果。

联结主义模型已经能推广到汉字阅读[11]，并模拟出成人汉字阅读的规则性效应，即声旁与整字读音相同的形声字（规则字）阅读起来要比不相同（不规则）的形声字快而准确[12~14]。一致性效应，即被试在同声旁且读音都一致的形声字上的命名要比不一致的快而准确[15~16]，以及二者与频率的交互作用[17~18]。但模型并没有反映出儿童的阅读发展：模型[11,18]的规则性和一致性是同步发展的，而儿童的一致性意识要比规则性意识发展晚[19]。

汉字阅读模型模拟阅读发展的不足可能在于训练语料。模型在不同阶段的训练语料都相同，不同阶段的阅读差异仅是熟练度的不同。而语料库统计发现[20~22]，低年级儿童学习的形声字主要是规则字，家族数较少且大部分都是一致家族。高年级儿童学习的生字中出现大量不规则字，形声字家族变得不再一致，这时声旁一致性意识才开始发展[20,23]。事实上，基于不同测试语料的行为实验所得到的一致性意识发展早晚也有差异[19,24]。

因此，使用小学课本作为训练模型的语料，应该能成功模拟出儿童的阅读发展，从而在汉字特异的书写系统下，考察阅读发展与阅读加工 2 个领域的研究能否统一在联结主义理论框架下。

1 研究方法

1.1 模型的建构

1.1.1 模型结构

汉字阅读的形-音对应模型采用与英语词汇阅读模型[7]完全相同的结构，输入层为 270 个单元的字形层，输出层为 92 个单元的语音层，输入层与输出层经 200 个隐单元相联结，层与层之间实行单向全联结，组成一个由字形到语音的前馈（Feedforward）网络；另外，语音层与自身单元之间实行全联结，同时，50 个单元的 cleanup 层又与语音层之间实行双向全联结，组成一个语音的吸引子（attractor）网络[25]。整个模型是由形-音对应的前馈网络与语音吸引子网络 2 个子网络组成的一个复杂网络，如图 1 所示。

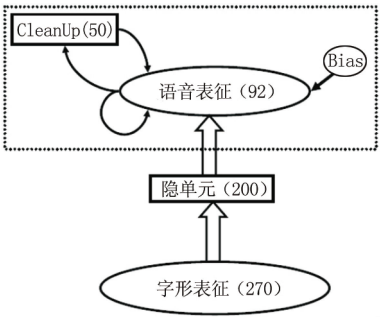

图 1 汉字阅读的形-音对应模型
Fig. 1 Orthography-phonology mapping model of Chinese characters reading

1.1.2 字形表征

XING 等[20,26]根据汉字的拆分结构信息，对每个汉字进行了详细表征，因而这里主要对其表征方案加以修正，以适合联结主义模型，主要进行 3 方面的修正：

首先，去除字形表征中整字和声旁的语音表征编码。因为在字形表征中加入语音表征，就相当于给模型预先设定了字形与语音之间的对应关系，而这种形—音对应关系需要模型在学习过程中通过统计学习而获得。

其次，把连续值表征改成二进制表征，去除每个部件冗余的 36 种视觉特征。为避免出现重码现象，增加 6 个随机单元来区分 53 个出现重码的部件（如，土与士，人与入）。

最后，对声旁独立成字时的字形表征重新编码。根据 XING 等[20,26]的表征方案，声旁独立成字时是按照字的方案表征的，这样一来，成字声旁（如“乔”）的 3 个部件就与它在形声字（“侨、娇、荞、桥”）中作为声旁时的表征位置不同。根据声旁家族中所有形声字的声旁典型位置来安排声旁成字时的表征，这样就保证了声旁与整字在字形上具有相似性（如表 1 所示）。事实上，这种表征方案保证了模型成功模拟出汉字命名的规则性效应。

表 1 汉字形表征的部件序列

Tab. 1 Representation slots for sequences of characters' radicals

		部件序列						
		1st	2nd	3rd	4th	5th	6th	7th
乔	声旁（修正前）	丿	大					冂
乔	声旁（修正后）					丿	大	冂
侨	左右结构	亻				丿	大	冂
矫	左右结构	𠂉	大			丿	大	冂
荞	上下结构	艹				丿	大	冂

1.1.3 语音表征

汉语音节由 3 部分组成：声母、韵母和音调，韵母又由介音、韵腹和韵尾组成。因而对于每一个音节，都用 5 个槽位来表征声母、介音、韵腹、韵尾和音调 5 部分信息。除音调由 4 个单元表征 5 种音调之外，其他部分均由 22 个单元来表征各音素的发音位置、发音方法等发音特征，这样，每个音节有 92 个单元表征。

在具体表征时，主要解决 2 个主要问题：首先根据汉语拼音方案，把实际的拼写转换成标准的拼音形式，如/ga1/的声母为/g/，韵腹为/a/，声调为 1，介音与韵尾则用“—”表示空，对于用/w/和/y/ 书写的零声母（如/wu1/）音节，则把声母还原为空，把简写的韵母（/iu/）还原成标准形式（/iou/）。其次是根据国际音标（International Phonetic Alphabet，IPA）方案，把书写形式相同但发音不同的同位异音（Allophone），如/ga1/与/gang1/的/a/，用不同的发音特征来区别表征。

1.2 模型的训练

模型的训练方法与成人阅读模型相同，只有训练材料来自小学课本的真实文字材料。训练 3 个分离的模型。模型 1 使用二年级所有 4 册课本的汉字（1 364 个）作为训练材料，根据其频率分布，同样使用平方根压缩的方法转化为模型训练时的抽取概率（抽取概率＝汉字频率/训练样本中最高频率，最高频率取值按高于 20 的取 20）；模型 2 使用四年级儿童学习过的所有 8 册课本的全部汉字（2 564 个）作为训练材料，最高频率取值为 40；模型 3 使用六年级儿童学习过的所有 12 册课本的全部汉字（3 306个）作为训练材料，最高频率取值为 70. 这里的大于最高频率取值的汉字占各年级所有训练汉字的 10%左右。

根据 3 个年级的汉字抽取概率的总和比例，总共对六年级模型训练 3 百万次，对四年级模型训练 2.3 百万次，对二年级模型训练 1 百万次。训练完成后考查模型在三类字上命名正确率及误差总平和（sum of squared errors，SSE）输出。

1.3 模型的测试

从 3 个年级的训练材料中分别挑选规则一致、规则不一致和不规则不一致汉字各 20 个。为便于年级之间进行比较，在每种条件下的 20 个形声字中，有 12 个是 3 个年级完全相同的，这些汉字从二年级到六年级的使用频率、形声字家族数都是增加的。有 4 个是二年级与四年级相同的汉字，4 个四年级与六年级相同的汉字，另外 4 个是只在二年级与六年级分别使用的测试汉字。

儿童从二年级到六年级的发展过程中，汉字出现的频率与声旁家族大小都会增长，从而促进其规则性、一致性意识的发展，因而测试材料保证在 3 个年级的频率与家族成员数都是呈递增的趋势（如表 2 所示）。

表 2 各年级测试材料的频率和家族成员数举例

Tab. 2 Examples for increased number of materials' frequency and family size

	频率/次			家族成员数/个		
	2nd	4th	6th	2nd	4th	6th
规则一致（塘）	2	5	18	1	2	4
规则不一致（评）	4	9	16	4	6	8
不规则不一致（治）	4	10	17	3	5	7

已有研究发现：规则性、一致性都只在低频条件下表现[12~16]，因而测试汉字全都选取相对各年级低频的汉字。频率确定以训练时的抽取概率小于 0.5 为标准：二年级模型的低频字低于 5 次，平均 2.2 次；四年级模型中的低频字小于 10 次，平均为 4.9 次；六年级模型的低频字小于 20 次，平均为 11.1 次。

模型训练结果后，从模型语音输出层的实际语音输出以及输出误差两方面进行评价。在测试模型的语音输出时，计算实际输出的语音层各单元与所有汉字音节表征之间的欧氏距离，距离最小的音节即为模型的输出语音；输出误差则计算语音输出层各单元与目标语音表征各单元的 SSE.

2 结果

各年级模型训练到预期次数之后，3 个年级的模型都能命名全部 60 个测试材料，因而只考查模型在测试材料上的 SSE 输出。

对各年级模型的 SSE 输出进行方差分析（analysis of variance，ANOVA），结果发现：二年级模型在 3 种测试字的主效应显著，$F(2, 59)=4.0$，$P<0.05$；多重比较结果表明：规则不一致汉字输出误差要显著小于不规则不一致汉字（$P<0.05$），表现出显著的规则性效应；但规则不一致汉字与规则一致汉字的 SSE 差异不显著（$P=0.79$），没有一致性效应。

四年级模型的 SSE 输出同样表现出 3 种测试汉字显著的主效应，$F(2, 59)=4.25$，$P<0.05$；多重比较结果表明：规则不一致汉字输出误差要显著小于不规则不一致汉字（$P<0.05$），表现出显著的规则性效应；但四年级模型的规则一致汉字与规则不一致汉字的 SSE 输出表现出差异，但还没有达到显著水平（$P=0.48$）。

六年级模型与前 2 个年级的模型相同，3 种测试字 SSE 输出的主效应显著，$F(2, 59)=4.64$，$P<0.05$；多重比较结果表现出显著的规则性效应（$P<0.05$），同时，与前 2 个模型相比，规则一致汉字的 SSE 输出要比规则不一致汉字的更小，达到了边缘显著（$P=0.07$）。

从 3 个年级的模型在规则性和一致性上的结果（如图 2 所示）可以看出：3 个年级都表现出规则性效应，一致性效应只有在六年级时才表现出来，这与儿童发展的结果[19]相一致：规则性意识先发展，一致性意识的发展正是因为儿童学习了大量的同声旁的形声字之后才发展起来的。

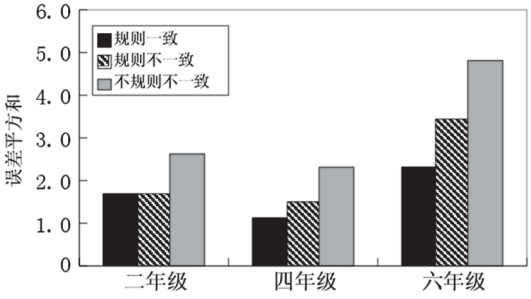

图 2 3 个年级模型在不同类型测试字的 SSE 输出
Fig. 2 SSE of different types of character in three grades' models

3 讨论

汉字阅读的联结主义模型成功模拟出儿童的阅读发展。使用小学课本中的汉字作为训练材料，二、四、六年级 3 个模型都表现出规则性效应，一致性效应在四年级开始表现出差异，但只有六年级模型才达到显著，表现出规则性意识先发展、一致性意识后发展，模拟结果与语料库的分析[20~22]以及儿童的阅读发展过程[19,27]基本一致，表明汉字阅读的联结主义模型能成功模拟出儿童的阅读发展。

与英文的阅读发展模型[9~10]不同，汉字的阅读发展模型没有人为地使用各种干预手段，只是使用儿童真实的语料输入，成功模拟出儿童的阅读发展过程。为模拟出跨语言的儿童阅读发展，研究者不得不增加一些额外的语音训练任务或者同时采用新的模型结构来解决。这里的模型采用与成人模型完全相同的模型结构、学习算法、训练和测试方法，得到了与自组织模型[20]基本一致的结果，从而表明阅读加工和阅读发展可能具有相同的内部机制，它们能被统一在联结主义的模型框架下。

联结主义模型在内部机制上就是一种对输入语料统计属性的抽取，在不同的语料输入下表现出不

同的阅读模式。阅读加工是阅读发展的结果：一方面，不同语言的语料属性不同，统一的机制既能体现出语言的共性，还能根据语料属性的不同体现出语言的特异性。另一方面，阅读发展不同时期的输入语料不同，从而导致阅读加工在不同阶段表现出不同的行为模式，形成阅读发展的阶段性。汉字形声字的绝大多数声旁都能独立成字，并且最早学习，对声旁的学习就直接促进了对规则形声字的学习，所以低年级儿童就能意识到声旁与整字读音的关系，体现出声旁规则性效应。随着年级的增加，学习的汉字绝对数量增加的同时，原来一致的家族变得不再一致[20~21]，儿童开始意识到不能再简单地根据声旁来推测整字的读音，开始受到家族内其他字读音的影响[28]，一致性意识才开始发展起来。但它们都是在不同的阶段对不同输入语料统计学习的结果。

对于汉字阅读发展的计算机模拟研究，不仅能从具体可运算的角度为阅读加工及其发展理论提供研究证据。同时，这种研究能为汉字教学提供实践的指导意义，例如一次汉字教材和教学方法的改革需要非常长的时间周期，反复的改革尝试也需要以学生的实际发展为代价，而计算机模拟方法能提供一种经济有效的改革尝试和评估系统，为实际的汉字教材乃至教育改革方案提供先期预测作用。

4 结论

1）使用真实的儿童学习材料，联结主义模型能成功模拟出汉字阅读的规则性、一致性意识的发展。

2）阅读发展与阅读加工可能具有相同的内部机制。

3）阅读发展受到输入语料的影响，不同的语料输入决定了阅读的不同加工模式。

[参考文献] (References)

[1] FROST R, KATZ L, BENTIN S. Strategies for visual word recognition and orthographical depth: a multilingual comparison[J]. Journal of Experimental Psychology: Human Perception and Performance, 1987, 13(1): 104-115.

[2] ZIEGLER J C, GOSWAMI U. Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory[J]. Psychological Bulletin, 2005, 131(1): 3-29.

[3] ZIEGLER J C, GOSWAMI U. Becoming literate in different languages: similar problems, different solutions[J]. Developmental Science, 2006, 9(5): 429-436.

[4] PLAUT D C, MCCLELLAND J L, SEIDENBERG M S, et al. Understanding normal and impaired word reading: computational principles in quasi-regular domains[J]. Psychological Review, 1996, 103(1): 56-115.

[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] ZEVIN J D, SEIDENBERG M S. Simulating consistency effects and individual differences in nonword naming: a comparison of current models[J]. Journal of Memory and Language, 2006, 54(2): 145-160.

[7] HARM M W, SEIDENBERG M S. Phonology, reading acquisition, and dyslexia: insights from connectionist models[J]. Psychological Review, 1999, 106(3): 491-528.

[8] WOOLLAMS A M, LAMBON RALPH M A, PLAUT D C, et al. SD-Squared: on the association between semantic dementia and surface dyslexia[J]. Psychological Review, 2007, 114(2): 316-339.

[9] HUTZLER F, ZIEGLER J C, PERRY C, et al. Do current connectionist learning models account for reading development in different languages?[J]. Cognition, 2004, 91(3): 273-296.

[10] POWELL D, PLAUT D C, FUNNELL E. Does the PMSP connectionist model of single word reading learn to read in the same way as a child?[J]. Journal of Research in Reading, 2006, 29(2): 229-250.

[11] 杨剑峰，舒华. 汉字阅读的联结主义模型[J]. 心理学报，2008，40 (5)：516-522.
YANG J F, SHU H. A connectionist model of Chinese characters reading[J]. Acta Psychologica Sinica, 2008, 40(5): 516-522. (in Chinese)

[12] SEIDENBERG M S. Constraining models of word recognition[J]. Cognition, 1985, 20(2): 169-190.

[13] 舒华，张厚粲. 成年熟练读者的汉字读音过程[J]. 心理学报，1987，19 (3)：282-290.
SHU H，ZHANG H C. The processing of pronouncing Chinese characters by proficient mature readers[J]. Acta Psychologica Sinica，1987，19(3)：282-290. (in Chinese)

[14] HUE C. Recognition processes in character naming[A]. CHEN HUSANCHIH，J. L. T O. Language Processing in Chinese[C]. North-Holland：Elsever Science Publisher，1992. 93-107.

[15] PENG D，YANG H，CHEN Y. Consistency and phonetic-independcency effects in naming tasks of Chinese phonograms[A]. JING Q，ZHANG H，PENG D. Information Processing of the Chinese Language[C]. Beijing：Beijing Normal University Press，1994. 26-41.

[16] PENG D，YANG H. The phonological processing of Chinese phonograms[J]. Asia and Pacific Journal of Speech，Language and Hearing，1997，2(2)：177-195.

[17] CHEN Y，PENG D. A connectionist model of recognition and naming of Chinese characters[A]. CHANG H W，HUANG J T，HUE C W，et al. Advances in the Study of Chinese Language Processing[C]. Taipei：National Taiwan University Press，1994. 211-240.

[18] YANG J，ZEVIN J D，SHU H，et al. A 'Triangle' model of Chinese reading[A]. Proceedings of the Twenty eighth Annual Conference of the Cognitive Science Society[C]. Mahwah：Lawrence Erlbaum，2006. 912-917.

[19] 舒华，周晓林，武宁宁. 儿童汉字读音声旁一致性意识的发展[J]. 心理学报，2000，32 (2)：164-169.
SHU H，ZHOU X L，WU N N. Utilizing phonological cues in Chinese characters：a developmental study[J]. Acta Psychologica Sinica，2000，32(2)：164-169. (in Chinese)

[20] XING H B，SHU H，LI P. The Acquisition of Chinese characters：corpus analyses and connectionist simulations[J]. Journal of Cognitive Science，2004，5(1)：1-49.

[21] SHU H，CHEN X，ANDERSON R C，et al. Properties of school Chinese：implications for learning to read[J]. Child Development，2003，74(1)：27-47.

[22] 舒华，武宁宁，郑先隽，等. 小学汉字形声字表音特点及其分布的研究[J]. 语言文字应用，1998，26 (2)：63-68.
SHU H，WU N N，ZHENG X J，et al. A study on the distribution and properties of phonological cues in school chinese phonograms[J]. Applied Linguistics，1998，26(2)：63-68. (in Chinese)

[23] CHEN X，SHU H，WU N，et al. Stages in learning to pronounce Chinese characters[J]. Psychology in the Schools，2003，40(1)：115-124.

[24] YANG H，PENG D. The learning and naming of Chinese characters of elementary school children[A]. CHEN H C. Cognitive Processing of Chinese and Related Asian Languages[C]. H K.：The Chinese University Press，1997. 323-346.

[25] HINTON G E，SHALLICE T. Lesioning an attractor network：investigations of acquired dyslexia[J]. Psychological Review，1991，98(1)：74-95.

[26] XING H，SHU H，LI P. A self-organizing connectionist model of character acquisition in chinese[A]. GRAY W D，SCHUNN C D. Proceedings of the Twenty Fourth Annual Conference of the Cognitive Science Society[C]. Mahwah：Lawrence Erlbaum.，2002. 950-955.

[27] 舒华，曾红梅. 儿童对汉字结构中语音线索的意识及其发展[J]. 心理学报，1996，28 (6)：160-165.
SHU H，ZENG H M. Awareness of phonological cues in pronunciation of chinese characters and its development[J]. Acta Psychologica Sinica，1996，28(6)：160-165. (in Chinese)

[28] SHU H，WU N. Growth of orthography-phonology knowledge in the Chinese writing system[A]. LI P，BATES E，TAN L H，et al. Handbook of East Asian Psycholinguistics[C]. Cambridge：Cambridge University Press，2006. 103-113.