基于系统思想和网络分析方法的繁简体字比较

沈哲思 [1], 闫小勇 [1,2], 狄增如 [1,3], 吴金闪 [1,3,†]
1. 北京师范大学系统科学系, 北京,100875, 中国
2. 石家庄铁道大学复杂系统研究中心, 石家庄,050043, 中国
3. 北京师范大学复杂性研究中心, 北京, 100875, 中国

January 25, 2013

摘要

本文从系统科学的思想出发，用复杂网络的方法分别考察了简体与繁体汉字之间的构成关系，并在这两个汉字构成关系网络上讨论了汉字学习效率，并以此为基础对比简体字网络与繁体字网络。研究发现：在不考虑汉字理据性的条件下，简体字的学习效率略高于繁体字；在考虑来自于专家的汉字理论理据性的条件下，简体字的学习效率略高于繁体字，但是学习效率提高的幅度很小；在考虑来自于大学生问卷调查所得的汉字实际理据性的条件下，简体字与繁体字的学习效率基本不可区分。

关键词：简繁之争，复杂网络，系统科学，汉字学习

Chinese characters, in either their simplified or the traditional forms, are connected via hierarchical structural relations: some simple characters are components of other more complicated characters. Based on this connection and guided by the basic idea of systems science — thinking of the systems as whole instead of each unit of the system, we perform a network analysis of the networks of simplified and traditional Chinese characters. Especially, the problem of learning efficiency is discussed on each of the two networks. Our investigation shows that if taking no consideration of rationality of Chinese characters, the simplified system has better learning efficiency than the traditional one; if taking theoretical scores of rationality from experts on Chinese characters, the simplified one is still better than the traditional one while the difference is much smaller than the previous case; if taking empirical scores of rationality from college students, learning efficiency of the simplified and traditional characters are almost indistinguishable.

Keywords: Simplified v.s. traditional characters, complex networks, systems science, learning Chinese characters

目录

1 引言

汉字记录着中华五千年的悠久历史和文化，在其发展过程中先后经历了甲骨文、金文、隶书、大篆、小篆、繁体字等不同的文字。在新中国成立后，颁布推行了简体字 [1]，并实施到现在，取得了非常不错的成果，帮助解决了文盲等问题 [2]。但关于繁体字和简体字的争论就一直没有停止过，在 09 年两会期间，政协委员梅葆玖 [3]、潘庆林 [4] 提出了“利用 10 年时间废除简体字，恢复繁体字”的提案更引起了社会各界的广泛争论 [5]。支持恢复繁体字的观点主要基于中华文化的传承和华语地区的交流需要，认为繁体字承载着中华的文化，更利于文化的传承和发扬 [6]，同时一字两体的存在阻碍了华语地区的交流；在非华语海外地区的中文教学与研究中，繁体字占了很高的比例，而近些年孔子学校的发展使得简体字与拼音教学也占了一定的比例，一字两体也会在海外汉语教学和研究中造成混乱；有的学者 [7] 认为简体字的简化并没有遵从汉字发展的一般规律，是一种“摧损六书，自乱体制”的做法。另一派的学者认为简体字通过对部分繁体字的简化，减少了笔画数 [8]，对于消除文盲起到了积极的作用，而且汉字简化已有很长一段时间，恢复繁体字反而会增加认字难度。

除了以上这些观点上的争辩，眞正对简体字与繁体字做系统地对比，关注目前的汉字简化系统整体的合理性的工作并不多。大多数社会上的一些争论则主要是集中在个别字之中，比如“爱（愛）无心、亲（親）不见”等，或者主要讨论某几个字的简化是否合理 [9]。据我们所知，有少数几位学者从理据性角度对部分简体字与繁体字做了对比 [10, 11, 12]。我们认为，暂时抛开民众基础、历史原因等因素，纯粹从科学的角度来看，对比简繁体字的主要着眼点应该是对比两者在学习难度上的差别。因此，在这里我们提出一种系统地、整体地对简繁体字的学习效率进行比较的方法。这种方法的基本思想是把所有汉字（简体字或者繁体字各自独立地），以及汉字之间结构上的联系作为一个相互联系的网络系统，然后寻找这个网络上的最优学习方式，最后对比简繁体字在这两个网络上的最优学习方式的学习效率。有关构造网络、寻找最优学习方式的细节我们会在下面做详细的介绍。这里先介绍我们这样做的出发点。我们认为汉字与汉字是相互联系的——一个字可以从意义或者读音上成为另一个字的一部分；如果我们能够系统地找出这样的依赖关系并且进一步利用这样的关系来帮助学习，那么我们就有可能高效地学习汉字；因此，比较简繁体字就是比较两

者多大程度上把汉字有机地关联起来了，或者说汉字简化的过程中保留了汉字之间的多少合理的结构上的联系。因此，在这个工作之中，我们主要考察与对比简繁体字组成的构字网络，以及在这两个网络上汉字的学习效率。在考虑学习效率的时候，我们还综合考虑了汉字理据性的影响。下面，我们首先对这个研究工作的数据来源、研究方法和已经完成的相关工作做一个介绍；然后，报告最主要的有关汉字学习效率的对比和结论；接着，我们会讨论简繁体字两个网络的基本结构的对比，基本结构的对比可以看作对两个网络学习效率上的对比关系的一个补充说明；最后我们给出这个研究的基本结论以及对结论做一些讨论。

2 数据来源、研究方法与已经完成的相关工作

2.1 数据来源

在汉字形成的历史中，有上万个汉字，但我们现在实际使用的汉字仅占了其中很小的一部分，而对于普通人，掌握 3000 ～ 4000 个常用汉字便可以满足其日常的生活。为此，在本文中，我们选择的汉字集合是《现代汉语常用字表》[13]，包含 3500 简体字，然后我们通过《简化字总表》找到这些字相对应的繁体字。在实际应用中这 3500 个汉字的覆盖率超过 99%。在汉字理据性方面，全面的讨论也不多。根据我们目前的调研，关于汉字理据性的研究主要集中在汉字的构字理据性，即比较现有汉字字形的表意度和表音度，而不追溯其原始的造字理据性，而且得到的是一些半定性半定量的比较结果。在这里，我们采用文献 [10, 11, 12] 中关于繁简字理据性分析的结论，用于汉字学习成本的讨论。同时，我们还组织了问卷调查来获取 48 名来自于全国各地的前来参加复杂性研究暑期学校的大学生对理据性的实际认知，而不仅仅是通过字音、字义等分析方法得到的理论理据性。

在汉字构成关系方面，我们把汉字与汉字之间的直接联系——也就是一个字成为另一个字的直接组成部分（如果汉字 A——例如召，先构成 B——昭，B 再构成 C——照，我们不认为 A 与 C 有直接关系）作为汉字之间的构成关系。在拆分汉字获取这样的联系的数据方面，我们主要参考《说文解字》[14]、《文字学概要》[15]，《常用字解》[16]，按照以下原则完成：

1. 象形字不拆分。在中国汉字中，象形文字占有一定的比重，它能让人比较好的联想到字和事物的关系，有利于整体记忆，不做下一步拆分。例如日、月、人、口等。我们认为笔画以及笔顺对于识字意义有限。

2. 指示、会意和形声字逐级拆分。这类汉字一般由若干汉字或部首构成，有比较明显的结构特征。例如形声字由形旁和声旁组成，而每个部分均有其特有的含义，方便学习记忆。这种拆分的理据性越强，则越有利于学习。

3. 偏旁处理。在汉字中有许多偏旁有其自身对应的汉字，比如“氵”与“水”，“扌”与“手”等，针对这样的偏旁，本文均将其替换为相应的原始汉字，以便用于简体和繁体网络的比较。当然也存在一些没有对应汉字的部首，比如“宀”、“冖”等，在处理这些偏旁时，本文保留这些偏旁，不作替换。

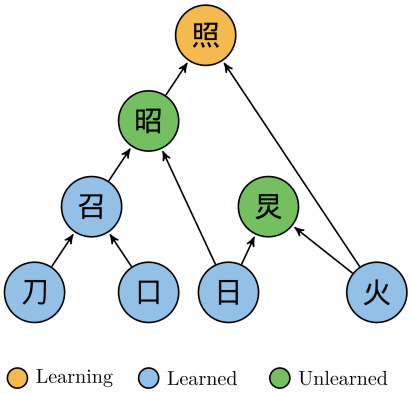

图 1: “照”字组成示意图。汉字左边的数字表示层级，右边表示其学习成本。把这几张结构示意图合起来，层级显示在左边，学习成本显示在右边，用 learning cost 那张看起来更好一点。加上已认识、未认识等图例

4. 按照上述原则进行汉字拆分后，会有许多半边字（其本身并不在原始数据中），比如“□”、“㠯”等。本文根据半边字是否构成了多于一个常用字来确定其是否需要单独列出，归入原始数据中，以待进一步拆分。比如“㠯”构成了“端”和“瑞”等常用字，于是将其单独列出；而“□”只构成了“漏”，且是生僻字，于是将其拆分为“尸”和“雨”，而“□”不再单独列出。

汉字拆分可以遵循不同的原则，例如字源、用字、构字等等几个层次 [17, 18, 19, 20, 21]。我们称目前这个工作中的拆分方法为表面结构拆分方法，基本上按照构字关系来拆分，考虑了一定的字源因素。我们不能保证我们拆分得到的结果每一个字都是合理的，更加不能保证我们的拆分完全把握了每一个字的字形与字义或者字形与读音的关系。我们只能够说在我们能力所及的范围之内，大部分的汉字的拆分是合理的，而且我们的拆分是在考虑了目前所能够收集到的关于汉字字源汉字结构的权威可靠的资料的基础上完成的。我们按照以上资料与原则所得到的简体字与繁体字拆分关系可以从我们的研究网站 [?] 下载。

2.2 研究方法

我们将汉字及其构字关系作为一个整体，采用网络的方法对其进行描述。所谓网络，最基本的元素就是顶点与边。在这里，我们把所有的汉字当作网络上顶点的集合。如果两个汉字之间存在这上面整理出来的表面结构关系，则在这代表两个汉字的顶点之间连接一条有向边，其方向我们定义为从构成用字指向所构成的字，也就是从简单汉字指向其所构成的复杂汉字。

复杂网络 [] 作为一种研究单元与单元之间复杂关系的方法被广泛应用于各类社会科学的研究之中，比如人际接触网与传染病研究 [22]、电力网与电力控制网耦合研究 [23] 等。我们提出用网络的方法来呈现的汉字之间构成关系这样一个设想，主要是为了讨论能否综合利用构字关系来寻找高效的汉字学习策略。我们认为对于一个理据性比较强的汉字，学习了它的组成部分就相当于

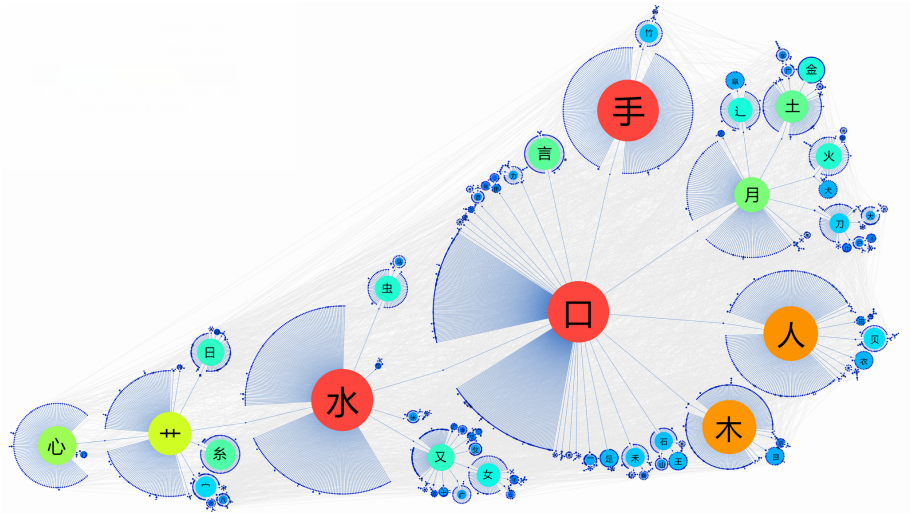

图 2: 简体字网络，来自于 [?]

图 3: 繁体字网络，让闫小勇帮你做一下这张图。

学习了这个字。例如前面所举的例子（见图 1），假设我们已经学习和理解了“日”和“召”两个汉字（其实“召”本身还可以继续拆分），那么我们很容易就理解和记住“昭”这个汉字。进一步，如果我们理解“灬”就是火的含义，我们很容易就可以学会“照”这个汉字。如果我们能够把所有汉字之间的构字关系都整理并类似地呈现出来，那么我们就能够大大地降低学习汉字的难度。尤其对于国外的或者少数民族的汉语学习者，甚至小学高段的学生，这样的系统都会很大程度上提高他们的学习效率。当然，我们也要注意到，并不是每一个汉字的字义或者读音与它的字形上的构成单位都存在很好的联系。因此，这个系统不可能是完备的系统。但是，至少通过网络的描述方法，我们能够把所有汉字之间的构成关系以一种整体的方式呈现出来，供下一步分析计算使用。这样的一种呈现方式，是任何在单个汉字个体的层次的讨论都无法做到的。为了给出一个直观的对比，下面我们分别制作了简繁体字的网络效果图，见图（2）和图（3）。粗看起来，这两个网络确实有很大的相似性。下面我们将进行更详细的分析。

2.2.1 分布式顶点权与汉字学习顺序的计算

那么有了这个网络之后，在这个网络上，该如何通过分析计算来利用这些关系，寻找更高效的汉字学习方法呢？这个是我们最近完成的工作主要提出和解决的问题 [24]。因为这个工作是我们这个简繁汉字对比工作的基础，在这里，我们对它做一个简短的概述。我们主要讨论汉字学习顺序的问题。具体到每一个汉字如何学习，我们认为，我们的汉字构成关系网络也是有借鉴意义的，但是对此我们暂时不做讨论。对于学习顺序，首先，一个简单直觉就是，汉字学

习要尽量遵循从简单基本到复杂的汉字的原则。也就是说，在我们的汉字构成关系网络上，优先从处于网络底层的、基本的汉字开始学习。其次，我们也希望优先学习使用频率高用频率高的汉字。这样，我们就可以避免花很多时间学习某一些很基本的但是基本上用不着的汉字。还有应该优先学习构字能力强的，也就是网络中顶点度大的汉字。但是，这三个优先原则给出的顺序并不一致。为此我们提出了一个分布式顶点权的概念，把它作为网络顶点重要性的度量。然后我们发现，按照这个重要性顺序来学习汉字，确实有比较高的效率。分布式顶点权的基本思想是把前面三个原则结合起来：网络上每一个顶点的初始权重设为所对应汉字的使用频率，从网络的最高层开始往下传播权重，直到最底层，传播以后每一个顶点得到的调整权重就是这个汉字的分布式顶点权。传播方法是把每一个当前汉字的总权重乘以 0.5 然后分配到这个汉字的所有直接组成部分（也就是父节点）上去。用公式来表达就是：

$$\widetilde{W}_i = W_i + 0.5 \times \sum_{<i,j>} \widetilde{W}_j \tag{1}$$

其中 W_i 表示节点 i 的原始权重，$\widetilde{W}_i$ 表示 i 节点的经传播后的调整权重，$\widetilde{W}_j$ 表示 j 节点的调整权重，$\sum_{<i,j>}$ 表示对节点 i 的所有子节点求和。关于 0.5 的取值见文献 [?] 的补充说明。

2.2.2 学习成本与学习效率的计算

按照上述方法得到理论上的汉字学习顺序后，我们对比简繁体字两个汉字构成关系网络上的高效学习顺序的学习效率。关于效率的衡量，我们的基本思想是：组成部分越多的汉字学习成本越高，组成部分中没有学过的汉字的个数越多成本越高，理据性越低说明该汉字组成部分表音表意越差，学习成本越高。在未考虑汉字理据性的情况下，具体的计算方式如公式 2：从一个被学习的汉字 A 开始，我们统计这个字的组成部分的个数，记为 N_A^{com}，然后统计每一个组成部分的成本（$C_j,\quad j = 1,2,\ldots,N_A^{com}$）。其中 C_j 的计算方法是如果汉字 j 已经学过则 $C_j = 0$，如果没有学过但是已经是最底层节点则 $C_j = 1$，否则说明汉字由更基本的汉字构成。在这种情况下，把汉字 j 当作被学习汉字按照汉字 A 的方法递归地计算其成本。

$$C_A = N_A^{com} + \sum_{j=1}^{N_A^{com}} C_j \tag{2}$$

在考虑汉字理据性的情况下可将公式 2修改为：

$$C_A = N_A^{com}(1 - R_A) + \sum_{j=1}^{N_A^{com}} C_j \tag{3}$$

其中 R_A 表示汉字 A 的理据性，可取 1（有理据）、0.5（半理据）和 0（无理据）值。例如当我们要学习图 1中的“照”时，我们需要学习“昭”和“火”，即 $N_A^{com} = 2$；而其中“火”已经学了，其成本为 0，而“昭”没有学，需要计算其学习成本，即“召”和“日”的组合成本及各自的学习成本，在这个例子中，“召”和“日”均已学，所以“昭”的学习成本为 2，即“照”的学习成本

为 4。当我们考虑理据性时，由于“照”是有理据的汉字，即 $R_A = 1$，根据公式 3得到其学习成本为 2。

汉字学习效果的衡量主要有以下两个方面：习得汉字数量的多少和习得汉字的累积使用频率的多少。每一个字的使用频率并不完全相同，因此这两种效果衡量方法的计算结果也有所不同。汉字学习效率就是在特定成本下学习效果的多少。在这里，我们比较系习得累积使用频率与成本的关系，以及习得字数与成本的关。

3 主要研究结果

根据 2.1中的拆分原则，本文分别将简体字和繁体字进行了拆分，分别得到 3688 个和 3818 个部件（汉字和组件），用于构建简繁体汉字网络。将这些部件作为网络的节点，根据汉字与汉字和组件之间的直接构成关系，连边建立有向网络。为了说明本方法的有效性，我们引入二简字作为对比，按照相同的原则进行拆分并构建了网络。二简字是中国文字改革委员会于 1977 年 12 月 20 日提出的《第二次汉字简化方案（草案）》中的简化汉字, 分为两表：表一收录了 248 个简化字，表二收录了 605 个简化字。一般认为该简化方案并不合理，造成了社会上的使用混乱，未能被广大人民接受。在 1986 年 6 月 24 日，国务院发出《国务院批转国家语言文字工作委员会〈关于废止《第二次汉字简化方案（草案）》和纠正社会用字混乱现象的请示〉的通知》，宣布废除“二简字”，并指出了“今后对汉字的改革要持谨慎态度，使汉字形体在一个时期内保持相对稳定，以利社会应用”。

3.1 简繁体字学习效率的对比

我们按照分布式顶点权方法给出的学习顺序和成本计算方式，对简体字、繁体字的学习效率做了计算，结果如图 4所示。从图中我们可以看到，在不考虑理据性的条件下，相比于简体字，繁体字的学习曲线处于下方，说明学习相同累积使用频率的汉字，繁体字花费的成本更高，而花费相同的学习成本，我们能学会更多的简体字、更高的累积使用频率。这样的结果是与繁体字简化过程中的组件合并、组建数减少和层级降低是相一致的，在某种意义上，简体字较繁体字更容易学习。在考虑理论理据性的情况下，我们可以看到，相比于简体字，繁体字的学习曲线依旧处于下方，但同时对比图中的无理据性曲线，可以注意到，简繁体字的学习曲线都有所提升，且繁体字的学习曲线的提升更加明显。理论理据性仅关注了汉字本身的表音表意程度，而没有从人的眞实认知角度进行考量，为此我们将实际理据性纳入考量，结果如图 4中的蓝黑色点线所示，可以看到在考虑实际理据性的简繁体字学习曲线和相互关系与理论理据性曲线是一致的，但相比于理论理据性曲线，两条曲线间的差距更加小了，也就是说在考虑实际理据性的情况下，学习繁体字的成本的降低更加明显。

二简字作为对简体字的进一步简化，我们也从学习效率的角度对二简字和简体字进行了对比，在图 5中我们可以看到二简字的学习效率较简体字只有轻微提高，说明二简字实际的简化效果并不明显。

为定量比较这样的提升并使不同汉字间的学习效率具有可比性，我们定义了汉字平均学习速度 V_N、 V_F，见公式 (4)(5)。我们将在实际理据性下学习二简字的成本定义为 C_{min}（本文取 $C_{min} = 6850$），在该成本下学习到的二简字

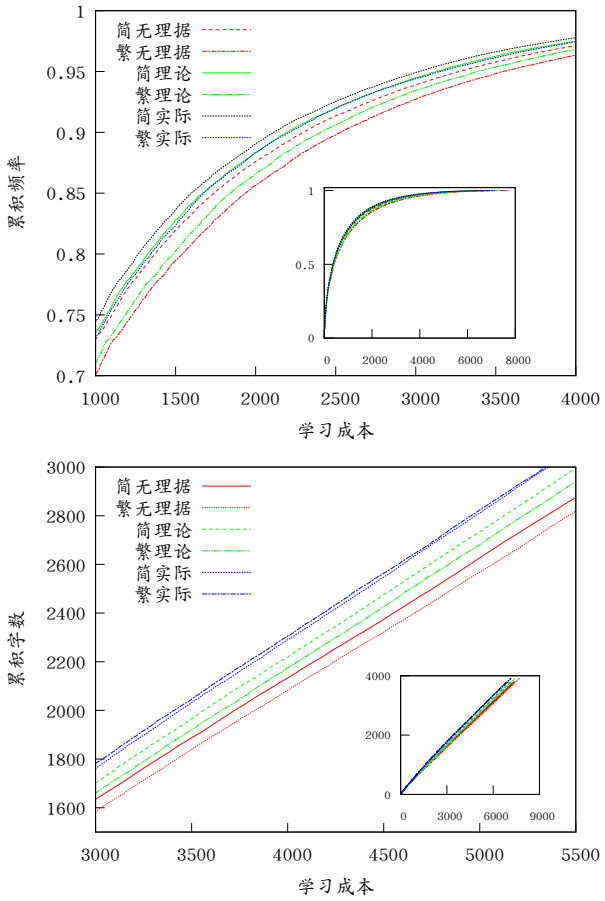

图 4: 繁简体字学习效率对比图

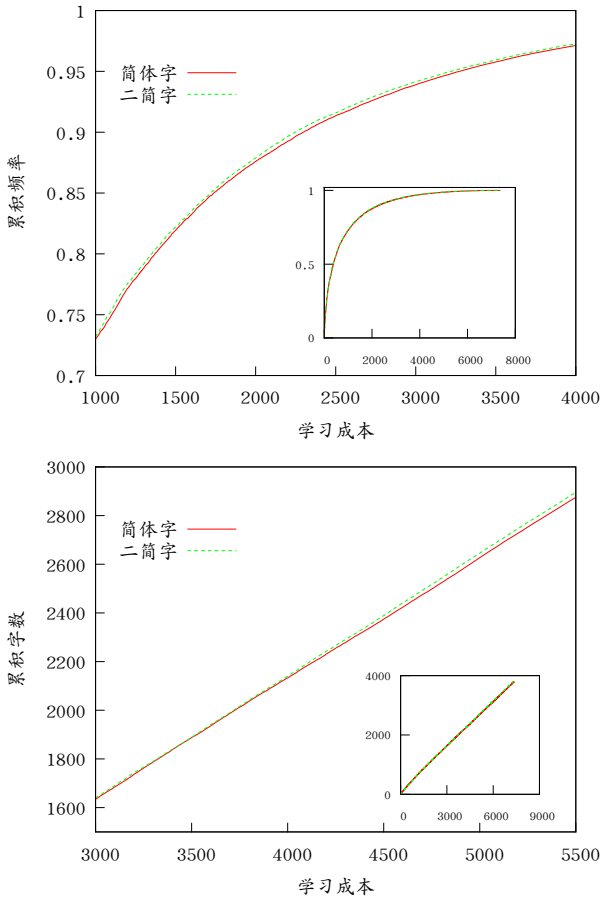

图 5: 简体字与二简字学习效率对比图

数量定义为 N_{min}，其相应的累积使用频率为 F_{min}；然后计算汉字学习曲线与 C_{min} 所包围的面积，根据学习目标的不同分别定义为 S_N（以汉字数量为学习目标）和 S_F（以汉字累积使用频率为学习目标）；这样汉字平均学习速度即为 S_N（S_F）与 $C_{min}N_{min}$（$C_{min}F_{min}$）的比值，比值越大说明学习效率越高。结果如表 1所示。在考虑理据性后，繁体字和简体字的平均学习速度都有提高，而繁体字提高的更加明显（其相对平均学习速度较未考虑理据性情况下有较大提高）；特别是在考虑实际理据性情况下，繁体字的平均学习速度提高更加明显。

$$V_N = \frac{S_N}{C_{min}N_{min}} \tag{4}$$

$$V_F = \frac{S_F}{C_{min}F_{min}} \tag{5}$$

表 1: 汉字学习速度对比表

	无理据性		理论理据性		实际理据性	
	V_F	V_N	V_F	V_N	V_F	V_N
二简字	0.881	0.486				
简体字	0.879	0.483	0.883	0.502	0.887	0.517
繁体字	0.867	0.472	0.873	0.492	0.882	0.520

通过上面的对比，我们发现不管考虑汉字理据性与否，简体字的学习效率均要比繁体字高（除了右下角最后一个数据，但是 0.520 与 0.517 之间差别非常小，而且同时 $0.882 < 0.887$），这说明简体字的简化是有利于汉字的学习和掌握的，而二简字的简化效果在汉字学习效率上体现的并不明显。当我们考虑了汉字理据性（即汉字本身的表音表意对学习的影响）后，繁体字学习效率的提升更加明显，特别是在考虑实际理据性后，提升的效果更加明显，说明繁体字在理据性上有一定优势，在一定程度上更加符合人的认知，但由于繁体字本身的构字复杂度使其并没有充分体现这种优势。目前，我们正在研究繁体字的造字理据性，如果繁体字的造字理据性能够大大高于简体字的造字理据性，那么一个能够很好地利用这一理据性的教学方法就有可能使得繁体字的学习效率更高。这一工作正在进行中。

在另一个工作中 [?]，我们用同样的方法研究对比了实际教材识字顺序的学习效率。在此，我们制作了一张表显示排在前面的几部教材的学习效率。我们看到第一实际教材的效率平均值大约在 $V_N \approx 0.3, V_F \approx 0.6$，还远远低于我们的理想效率，这说明实际教材的学习效率还有很大的提高的空间。另外我们看到实际教材之间的差别大约是 0.04 的量级。后者说明我们计算所得到的简繁体字最优效率的差别在 0.01 左右，小于实际教材的效率之间的区别的通常大小。

3.2 简繁体字网络基本结构的对比

为了对上面的学习效率的分析结果做一个补充说明，也为了分析对比简繁体字网络基本结构的异同，本文采用网络的分析方法对上述网络的基本结构做了分析。首先我们所构造的汉字网络是一个有向无环图，是一种完全的层次网络，可以采用拓扑排序的方法标记出网络中各个节点所处的层级，具体规则如下：

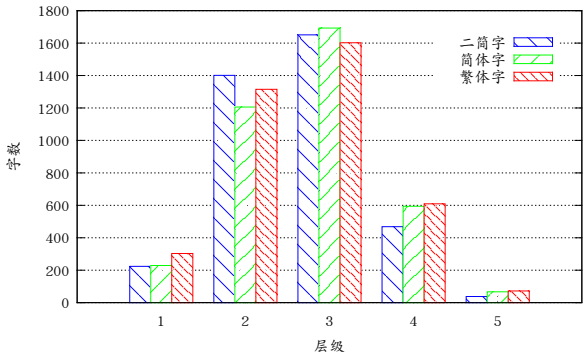

图 6: 层级结构分布图

1. 对于某一节点，若其不存在指向该节点的边，则该节点为第 1 层级；
2. 对于某一节点，若所有指向它的边连接的节点的最大层级为 N，则该节点的层级为 N+1。

例如图 1中，“石”、“氵”、“艹”、“甫”和“寸”为 1 层级，“尃”为 2 层级；“溥”由 2 层级的“尃”和 1 层级的“氵”组成，为 3 层级；“薄”由 3 层级的“溥”和 1 层级的“艹”组成，为 4 层级；“礴”由 4 层级的“薄”和 1 层级的“石”组成，为 5 层级。节点的层次在一定程度上反应了该汉字学习的难易程度。做相应修改。分别将简体字、繁体字和二简字网络的节点的层级进行频数统计，得到图 6：层次分布图。

从图 6中可以看到低层级汉字（基础组件及简单汉字）仅占全部汉字的 10% 左右；超过 80% 的汉字位于第 2、第 3 和 4 层，这说明汉字更倾向于采用逐级组合的方式构字。但超过 4 层的汉字所占比例不到 5%，这是因为层次数太多会使得汉字结构过于复杂（例如图??中的“礴”字，它是 5 层级汉字）。而从简体汉字网络和繁体汉字网络对比上可以看到，两个网络的层次结构在整体上是基本一致的，这说明在简化繁体字的过程中并没有改变汉字的结构（逐级组合）构字这一最本质的东西；而在一些细节上，繁体字网络在低层级汉字（基本组件和简单汉字）和高层级汉字上相对多一些，这说明在繁体字简化过程中，简化合并了一些简单的字或者组件，并且将少量的高层级的字简化到低层级中去了。同时我们可以注意到二简字在二层级的字明显多于简体字，而在三、四、五层级的字较少，这说明二简字将通过对中高层级汉字的简并、部件的替换大大降低了层级结构。汉字网络的这种层次结构对于汉字学习是有意义的。最底层的汉字需要独立地记忆，最高层的汉字依赖于多个其它的汉字。一个容易学习的系统，两者都不能过多。当然由于字义的需要，某些字必须依赖于多个汉字，这一点是避免不了的。这一点，我们看到简体字、繁体字都比较符合。然而二简字中层级为二的字尤其多（简体字、繁体字都是层级为三的汉字居多），高层级的字明显减少，这一点应该是强行简化的结果。

汉字的基础组成部件数则从另一方面对繁简字的学习难度进行了刻画。在网络中，某节点所连接的 1 层级节点数表示了该节点汉字由多少基础部件（汉字）组成，若该节点数越大，说明该节点汉字需要的基础部件（汉字）越多，组成更加复杂，反之该节点汉字比较简单。通过对繁简字网络各节点基础组件

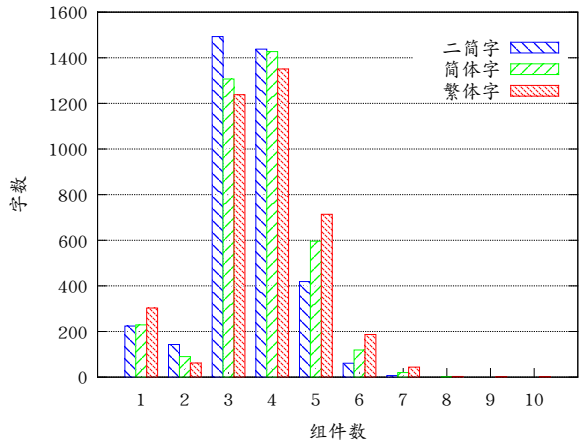

图 7: 汉字基础组件数分布图

数的统计，得到下图： 从图 7中可以看到，汉字的组成难易程度跨度还是非常大的，简单的汉字仅由一、二个基础部件（汉字）组成，而复杂的汉字则需要九、十个基础部件组成。繁简字组件数分布还是较一致的，同时我们也能注意到在低组件数（组件数为 1）和高组件数（组件数大于 4）的分布上，繁体字均多于简体字，这与层级结构分布上的差异是一致的，主要是由于简化过程中的基础部件的合并和高层级汉字的简化造成的。当我们综合比较图 6和图 7时，可以看到虽然繁体字的高组件数汉字明显大于简体字，但在层级结构上并没有反映出相一致的特性（即高层级繁体字汉字大于简体字），反而在 3 层级分布上，简体字还要多一些，这说明部分繁体字虽然基础部件数较多，但层级不高，各基础组件间只是简单的堆积，并没有构成简单结构的成字；而简体字利用较少的基础部件数，组成了中低层级成字并在此基础上进一步组合形成中高层级汉字。在综合比较简体字和二简字时，我们可以注意到，在基础字数量基本相同的情况下，二简字中低层级的字和中低组件数的字都更多，这说明二简字在基于简体字的简化过程中，是通过用低层级的简单汉字替换中高层级的复杂汉字，特别是利用一些相同的基础字代替表达不同意义的复杂字，以达到降低层级结构的目的。但这样的简化并没有考虑汉字的层级意义以及汉字的理据性。二简字的层级结构和组件数分布与简体字和繁体字有非常大的区别（高层级和高组件数汉字大大减少，中低层级和中低组件数的汉字显著增加），通过减少汉字的组件数（大量合并复杂汉字，和形近汉字）大大降低了汉字的层级结构，在很大程度上减少了汉字的区分度，不利于汉字的学习和识别。汉字的简化优化不仅仅是个别汉字的简化优化，更应该在此基础上达到整个汉字体系的优化引用王宁：汉字的优化与简化。

4 研究结论与讨论

本文通过对繁简字的结构拆分，运用复杂网络分析的方法，从一个系统的、宏观的角度对繁简字的差异进行了分析。我们看到简体字和繁体字在层级结

构分布和组件数分布上是基本一致的，而与二简字有很大的差异，同时在细节上简体字较繁体字减少了汉字构成的组件数量、并降低和优化了汉字的层级结构，同时也一定程度上损害了汉字的理据性。我们通过汉字学习的效率度量得到了在不考虑理据性的情况下简体字较繁体字有一定的学习优势，而在考虑了理据性的情况下，简体字的学习效率仍然高于繁体字，但是提高的幅度非常有限，甚至在考虑实际理据性的条件下基本不可区分。也就是说，第一简体字带来的对汉字理据性的损害程度针对学习而言还在可接受的范围内，第二那么简体字带来的学习效率提升的效果有限，如果我们的汉字教学系统能够有机地利用汉字之间这种结构关系。但是，另一方面，如果我们的汉字教学系统不能很好的利用这样的结构关系，那么简体字在提高识字率方面确实发挥了比较大的作用。那么，下一步的问题就是，当年我们没有这样的好的汉字教学系统，现在我们有了吗？在 [?] 中，我们利用同样的学习效率的计算方法比较了 20 套主流汉语教材。我们发现，其效率都比我们这里得到的学习效率指数小很多。这表明，如果我们的汉字教学系统能够利用汉字之间的结构上的联系 [?]，那么在目前的基础上，不论繁体字简体字，都存在很大的学习效率提升的空间。

5 致谢

参考文献

[1] 简化字总表.

[2] 邢锐. 浅谈汉字繁简问题. 安徽文学 *(*下半月*)*, (6):326–328, 2009.

[3] 《视频：京剧大师梅葆玖建议恢复繁体字》.

[4] 《全国政协委员潘庆林建议恢复使用繁体字》.

[5] 谢金良. 关于繁体字与简体字的若干思考. 闽江学院学报, (4):45–49, 2009.

[6] 吴燕敏黄闵. 繁体字优势及其文化意义. 当代小说：下半月, 10(10):73, 2010.

[7] 张善文. 古典文献研究与繁简字的思考. 闽江学院学报, (3):66–70, 2009.

[8] 郭曙纶. 简化字与繁体字笔画数的动态统计与比较. 北华大学学报 *(*社会科学版*)*, (2):50–56, 2009.

[9] 徐冬梅. 简化字“发”及其两个繁体字“發”“□”问题研究. 安徽文学 *(*下半月*)*, (8):295–296, 2009.

[10] 宣丽娟. 繁简字理据性比较分析. 语文学刊：高等教育版, 1(1):71–74, 2004.

[11] 邱理萌. 繁简字理据度分析. 汉字书同文研究, 4.

[12] 吴芳芳.《简化字总表》中简化字与其对应繁体字理据性比较与分□ . PhD thesis, 河北大学, 2009.

[13] 现代汉语常用字表.

[14] 许慎. 说文解字. 中华书局, 2004.

[15] 裘锡圭. 文字学概要. 商务印书馆, 1988.

[16] 白川静. 常用字解. 九州出版社, 2010.

[17] 王宁. 汉字构形理据与现代汉字部件拆分. 语文建设, (3):4–9, 1997.

[18] 潘德孚. 论汉字拆分的系统性. 汉字文化, (04):26–30, 2002.

[19] 叶平. 论汉字结构规律及汉字拆分方法. In 中国中文信息学会汉字编码专业委员会第九届年会暨学术 讨会, page 9, 中国江苏苏州, 2011.

[20] 张玉金. 论汉字的部件拆分和字符拆分. 辽宁师范大学学报, (04):67–69, 2000.

[21] 杜朝科. 2500 常用汉字的尝试性拆分. 科教文汇 *(*上旬刊*)*, (07):242, 2008.

[22] R. Pastor-Satorras and A. Vespignani. Epidemic dynamics and endemic states in complex networks. *Physical Review E*, 63(6):066117, June 2001.

[23] Roni Parshani, Sergey V. Buldyrev, and Shlomo Havlin. Interdependent networks: reducing the coupling strength leads to a change from a first to second order percolation transition. *Physical review letters*, 105(4):048701–048701, July 2010.

[24] 闫小勇吴金闪. Strategy for effcient learning of chinese characters offered by network analysis.