汉语中介语词同现网络研究

钱隆[1]，赵慧周[2]，丁芊[3]，王治敏[4]*
1.北京语言大学 国际中文教育研究院/北京
2. 北京语言大学 信息科学院/北京
3. 安徽信息工程学院 通识教育与外国语学院/安徽芜湖
4.广东外语外贸大学 中国语言文化学院/广东广州
qianlong_1221@163.com zhaohuizhou@blcu.edu.cn
dingqian0801@163.com wangzm000@qq.com

摘要

近年来，运用复杂网络方法进行语言学研究已成为数字人文研究的一条新路径。本文基于214篇日本汉语学习者的书面作文，构建了6个不同能力水平的汉语中介语词同现网络，并探讨了这些网络的结构特性及其动态演变过程。研究结果显示，所有的汉语中介语词同现网络均呈现出小世界属性、无标度属性、异配性和层级结构等复杂网络的特性。这些特性揭示了汉语学习者在词汇使用方面的特定模式：低水平学习者更倾向于将低频词汇与高频词汇进行连接，这可能与学习者减轻认知负荷的习得模式有关；学习者语言水平的提升，中介语网络参数会逐渐向母语者靠拢，但是无法达到母语者的水平；此外，本研究还观察到，语言错误会对中介语网络结构产生影响，引起网络结构的变异。

关键词： 复杂网络；汉语中介语；小世界属性；无标度属性；异配性

A Study on Chinese Interlanguage Co-occurrence Networks

QIAN Long[1] ZHAO Huizhou[2] DING Qian[3] WANG Zhimin[4]
1. Research Institute of International Chinese Language Education,
Beijing Language and Culture University / Beijing
2. College of Information Science,
Beijing Language and Culture University / Beijing
3. Faculty of General Education and Foreign Languages,
Anhui Institute of Information Technology / Anhui Wuhu
4.Faculty of Chinese Language and Culture,
Guangdong University of Foreign Studies / Guangdong Guangzhou

Abstract

Within the framework of complex network theory, this paper constructs 6 Chinese interlanguage co-occurrence networks based on 214 written compositions by Japanese L2 learners of Chinese, categorized by different proficiency levels. It explores the structural characteristics and dynamic evolution of these networks. The results reveal that all the Chinese interlanguage co-occurrence networks exhibit complex network properties such as small-worldness, scale-freeness, disassortativity, and hierarchical structure. These characteristics highlight specific patterns in vocabulary usage among Chinese learners: learners at lower levels tend to connect new words with high-frequency words, possibly related to the anchoring effect and the principle of least effort; as learners' proficiency

*通讯作者 Corresponding Author

基金：本文系2018 年国家社会科学基金重大项目"基于'互联网+'的国际汉语教学资源与智慧教育平台研究"（18ZDA295）的研究成果。

increases, the parameters of the interlanguage networks gradually converge towards those of native speakers, yet they do not reach the native speakers' level; moreover, this study observes that language errors impact the interlanguage network structure, causing variations.

Keywords: Complex Networks , Co-occurrence , Chinese Interlanguage , Small-World Property , Scale-Free Property , Disassortativity

1 引言

在自然界和人类社会中，网络现象无处不在。网络是可以用图来表示的系统，系统中的元素是网络的节点，而元素之间的关系则是网络的边(Barabási, 2002; Watts, 2003)。绝大多数现实世界的网络，例如：社交网络(Davidsen et al., 2002)、学术引文网络(Greenberg, 2009)、生物基因网络(Levine and Davidson, 2005)、流行病传播网络(Danon et al., 2011)和电力网络(Pagani and Aiello, 2013)等都属于复杂网络范畴，它们既不是完全规则的，也不是完全随机的(Watts and Strogatz, 1998)。自Watts and Strogatz(1998)和Barabási and Albert(1999)在复杂网络研究中分别引入了"小世界网络"和"无标度网络"的概念以来，复杂网络理论便迎来了快速发展的阶段。此后，广泛的研究表明，小世界属性和无标度分布在自然界和人类社会中均普遍存在。这些研究成果不仅进一步丰富了复杂网络的理论，也为跨学科研究提供了新的视角与方法。

语言普遍被视为一种复杂适应系统(Briscoe, 1998; Ellis and Larsen-Freeman, 2006; Ellis and Larsen-Freeman, 2009; Cameron and Larsen-Freeman, 2007)。这使得复杂网络分析方法应用于语言学研究成为一种可能。复杂网络方法能够捕获网络的整体特征，揭示相变，并为系统分析提供有效的定量测量(Cong and Liu, 2014)。在语音、词汇、句法、语义等不同层面上，自然语言都展现出复杂网络的特性(刘海涛, 2022)。研究发现，汉字网络(Li and Zhou, 2007)、句法依存网络(Cancho et al., 2004)、语义网络(Borge-Holthoefer and Arenas, 2010)等均具有复杂网络特性。随着复杂网络理论的不断深化，研究者开始关注于中介语网络的结构特性以及这些特性在语言习得和应用中的影响。

中介语是指第二语言学习者在其母语（第一语言）和目的语（第二语言）之间形成的一种心理语言系统，其特点包括动态性、系统性、过渡性和创造性(Selinker, 1972)。近年来，随着复杂网络方法的应用与发展，越来越多的研究聚焦于中介语网络的结构特性及其发展过程。例如：Jiang et al(2019)在研究英语中介语句法依存网络时指出，二语学习者的句法发展并不会像一语学习者那样出现"句法涌现"现象。Hao et al(2021)揭示了复杂网络参数能有效区分二语学习者的语言熟练度。韩笑等(2021)发现中介语句法复杂度网络的演变主要体现在网络强度及中心节点的变化上。王士丛(2021)则观察到汉语中介语网络的国别化差异，不同国别学习者词汇网络内部存在致密性差异。虽然这些研究丰富了我们对中介语网络结构与动态性的认识，但依然存在一些不足。首先，大部分研究集中于探讨中介语网络的总体特性及其演变，却往往未能深究这些特征以及变化背后的具体成因。其次，当前的中介语网络研究主要集中在句法依存网络的构建上，而较少使用词同现网络方法，尽管后者在直观反映词汇间的搭配关系和结构特征方面具有明显优势(Liu and Cong, 2013)。最后，二语学习者的语言输出会产生较多的语言错误，这些错误是中介语的重要组成部分，然而现有的研究对此并未给予足够的关注和讨论。

鉴于此，本文通过构建与分析汉语中介语词同现网络，旨在揭示汉语作为第二语言学习者在语言习得过程中的模式与特征，深入分析这些模式和特征背后的成因，并考察语言错误对中介语网络的具体影响，进而为汉语教学提供理论依据与实践指导。

2 研究设计

2.1 语料

本研究使用的语料分为中介语语料和母语者语料两部分。中介语语料采集自全球汉语中介语语料库（简称"全球库"）[-1]。全球库是一个大型的汉语中介语语料库，迄今为止已累积收集了约2367万字的中介语资料(张宝林、崔希亮, 2022)。本研究选取了全球库中1530篇日本汉

[-1]全球汉语中介语语料库网址：http://qqk.blcu.edu.cn/

语学习者的书面作文，总字数约为41.7万字。母语者语料则采集自作文网[0]，这是一个开放的中国中小学生作文投稿平台，涵盖了广泛的年级和题材。本研究从该平台收集了60篇中学生的期末考试作文，总计3.9万字。语料采集完毕后，依据《欧洲语言共同框架》（CEFR）中的整体书面产出（Overall Written Production）和写作评定网格（Written Assessment Grid）对1530篇作文进行分级。《欧洲语言共同框架》将语言使用者的语言水平分为六个等级，分别是：A1、A2级表示基础使用者（Basic user），B1、B2级表示独立使用者（Independent user），C1、C2级表示熟练使用者（Proficient user）(of Europe, 2018)。分级工作由三名评分员和一名专家共同完成，三名评分员均拥有超过三年的评分经验，而专家则多次担任国家级汉语水平考试的阅卷组长。具体步骤如下：首先，每位评分员独立地对每篇作文按照A1至C2的等级进行分级。若三位评分员给出的分级完全一致，或者其中两位的评分一致且第三位的评分与此一致评分相距一个等级，则认定该分级结果有效，据此确定作文的最终等级。如果三位评分员的评分全都不一致，或两位一致但第三位的评分与此一致评分相距两个等级或以上，则认定该分级结果无效，需提交给专家进行重新分级。最终，有效分级的作文数量为1231篇，无效分级的作文数量为299篇，有效分级的比率达到了80.46%。为验证分级结果的一致性，采用SPSS 27.0软件进行了肯德尔和谐系数（Kendall's coefficient of concordance）检验，具体检验结果展示于表1中。

评分员数	肯德尔W^a	卡方	自由度	渐近显著性
3	0.747	3425.435	1529	0.000

Table 1: 分级数据肯德尔和谐系数检验结果

由表1可知，肯德尔和谐系数检验呈现出显著性(p=0.000<0.05)，三位评分员的分级结果具有一致性。同时，肯德尔和谐系数为0.747，介于0.6-0.8之间，说明三位评分员的分级一致性较强，分级结果可以采纳。结合有效的分级结果和专家的分级结果，本研究对1530篇语料全部进行了分级。

中介语语料分级结束后，使用Jieba软件对母语语料和中介语语料进行了分词处理，并人工修改了部分错误。母语语料与中介语语料详细信息见表2。

子库	A1	A2	B1	B2	C1	C2	母语
篇数	94	410	582	290	115	39	60
字数	11828	79943	158241	92665	44626	23218	39415
词数	10526	67585	127461	72837	34672	17597	25227

Table 2: 母语语料与中介语语料详细信息

考虑到六个等级的中介语语料与母语语料总量不等，为避免语料规模差异对研究结果产生影响，本研究对各子库内的语料进行了平衡处理。具体措施包括：首先，从每个等级的作文中剔除词数的极端值，以保证各等级学习者的写作水平具有一致性；随后，通过简单随机抽样的方式调整每个子库的规模，确保其词数控制在5000±50词的范围内，以实现语料库之间的容量均衡。最终我们抽取了214篇中介语作文以及12篇母语作文。语料抽样信息见下表。

子库	A1	A2	B1	B2	C1	C2	母语
篇数	69	52	31	26	21	15	12
字数	17627	17827	17785	17881	18140	18111	17893
词数	5036	5014	5018	4957	5020	5021	4981

Table 3: 母语语料与中介语语料抽样信息

[0]作文网网址：https://www.zuowen.com/

2.2 网络构建

无论网络如何复杂，其构成的基本要素都很简单。网络构成的基本要素就是节点（vertices）和边（edges）(Newman, 2002)。在语言网络中，节点通常由各类语言单位充当，如：词、音素、语素、汉字、汉字部件等；而边则表示节点与节点之间的关系(Liu, 2008; Cong and Liu, 2014)。

在构造以词为节点的网络时，采用词同现的方式来构建语言网络是一种常用方法。词同现揭示了语言的基本属性之一——线条性，即词汇在文本中按照先后顺序线性排列。词同现网络以词（词型，word types）为节点，两个同时出现（相邻）的词通过边相连(Liang et al., 2009)。词同现关系存在有向和无向之分。尽管有向网络能够展示词汇序列的方向性，但是无向网络则更适合于分析词汇同现的结构及其动态变化(Garg and Kumar, 2018)。因此，本研究采用无向词同现网络的构拟方法来构建语言网络，即从语料库中提取所有不同的二元词同现对，见图1（a），并将其转换成词同现网络，见图1（b）。

(a) 词同现分析　　(b) 词同现网络

Figure 1: 例句的词同现分析和词同现网络

在构建网络之初，本研究首先对语料进行了处理，去除了所有的非完句标点符号（“，”“：”“；”等）和外文符号。接着，参考刘海涛(2022)的做法，从各子库中提取二元词同现对，直到遇到完句符号（“。”“！”“？”），此时跳转至下一句继续提取，确保二元词同现对不跨句提取。随后，将提取的二元词同现对依次放入Excel表格中。最后，使用Createpajek软件，将Excel中的二元词同现对转化为词同现网络。据此，本研究成功构建了六个不同等级（A1、A2、B1、B2、C1、C2）的中介语网络以及一个母语网络。

为了探究中介语网络是否具有小世界属性，本研究进一步构建了六个相应的随机网络进行比较分析，这些随机网络与对应的中介语网络具有相同的节点数、边数和平均度。此外，为了观察语言错误对中介语网络结构的影响，本研究还对六个中介语网络中的语言错误进行了人工修正，创建了六个无语言错误的中介语词同现网络。构建完成所有网络后，使用Pajek软件对这些网络的特征进行了数据提取。

2.3 网络的统计特征

2.3.1 网络密度

网络密度D指的是网络中实际存在的边（即连接）数与网络中可能存在的最大边数之间的比例。该指标可以描述网络中节点间连接的紧密程度(Wasserman and Faust, 1994; Scott, 2000)，公式如下：

$$D = \frac{2M}{N(N-1)}$$

其中，M是网络中实际边的数量，N是网络的节点数。网络密度的值范围从0到1。0表示一个完全无连接的网络（即没有任何边），而1表示一个完全连接的网络（即网络中任意两个不同的节点之间都存在边），通常真实网络的密度总是小于0.5(Mayhew and Levinger, 1976)。高密度值意味着网络中的节点之间存在较多的连接，表明词与词之间的连接很紧密；相反，低密度值表明网络中的连接较少，词与词之间的连接较为稀疏。

第二十三届中国计算语言学大会论文集，第712页-第725页，太原，中国，2024年7月25日至28日。
卷1：主会论文

2.3.2 度分布

假设一个有N个节点的网络中，边都是无向且无权的，那么节点i具有的边的数量则是该节点的度k_i。节点的度分布$p(k)$指的是网络中随机选择的节点恰好为k度的概率，如果度分布$p(k)$满足幂律分布($p(k) \propto k^{-\gamma}$，γ为常数)，则网络被称为无标度网络(scale-free network)。通常，真实网络的度分布都服从幂律分布，而随机网络都服从泊松分布(Barabási and Albert, 1999)。

2.3.3 平均路径长度

在无向无权的网络中，两个节点之间的距离d为连接二者的最短路径的边的数目(刘涛et al., 2005)，平均路径长度L则是网络中任意两个节点之间距离的平均值。

$$L = \frac{2\sum_{i<j} d_{ij}}{N(N-1)}$$

其中，d_{ij}表示节点i和节点j之间的最短距离，N为网络的节点数。平均路径长度短表示网络中任意两个节点之间可以通过较少的中间节点相互到达，网络具有较高的全局连通性；反之，则表示网络的连通性较差。平均路径长度是衡量网络是否具有小世界属性的重要指标之一(Albert and Barabási, 2002)，绝大多数真实网络的平均路径长度要小于随机网络(Watts and Strogatz, 1998)。

2.3.4 聚集系数和度相关平均聚集系数

聚集系数C是用来描述网络中节点之间聚集程度的系数。聚集系数可以反映连接到同一个节点上的两个节点相互连接的概率(Newman, 2010)。假设节点i通过k_i条边和其他的k_i个节点相连，那么这k_i个节点都可以视为节点i的邻接点。如果i的所有邻接点之间又都互相连接，那么他们之间存在$k_i(k_i\text{-}1)/2$条边。如果将M_i看作是i的所有邻接点之间存在的实际边数，那么M_i与$k_i(k_i\text{-}1)/2$的比值则是节点i的聚集系数C_i(刘涛et al., 2005)。

$$C_i = \frac{2M_i}{k_i(k_i-1)}$$

一个网络的聚集系数C则是整个网络中所有节点聚集系数的平均值(**?**)。聚集系数的值介于0和1之间。0表示网络中没有节点形成紧密的群组，而1表示网络是完全聚集的，即每个节点的所有邻居节点都彼此直接相连。聚集系数也是小世界效应的重要衡量标准之一(Albert and Barabási, 2002)。通常，在相同节点数和边数的情况下，真实网络的聚集系数要远大于随机网络的聚集系数(Watts and Strogatz, 1998)。

度相关聚集系数$C(k)$（Degree-dependent clustering coefficient）则是聚集系数的扩展，指的是所有度数为k的节点的平均聚集系数。在研究中，可以通过该指标分析网络的度分布和聚集系数的关系，以揭示网络的层级结构。当度相关聚集系数$C(k)$遵循幂律分布（$C(k) \propto k^{-\gamma}$，γ为常数），这表明网络往往呈现出层级化的拓扑结构特征(Ravasz and Barabási, 2003)。这种层级结构可被视为网络的一种组织模式(刘海涛, 2017)。幂律分布意味着高度节点较少参与形成三角结构（三元闭包），通常都是低度节点倾向形成三元闭包构成较为紧致的子网络，子网络间的连接相对较少，而高度节点则将这些子网络连接成一个连通的网络。

2.3.5 同配系数

同配系数（assortativity coefficients）是用来量化网络中节点的连接倾向的一个指标，可以用来反映网络中节点连接的同质性。基于节点度的同配系数可以衡量高度节点是否倾向于与其他高度节点相连，以及低度节点是否倾向于与其他低度节点相连。公式如下：

$$r = \frac{M^{-1}\sum_i j_i k_i - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}$$

其中，M是边的总数，j_i和k_i是连接到第i条边两端节点的度。同配系数的取值范围在-1和1之间，当同配系数为正值（r >0）时，表示网络具有同配性，即类似度数的节点更可能相连；同

配系数为负值（r <0）时，网络具有异配性，即高度数的节点倾向于与低度数的节点相连；同配系数接近0意味着网络中的连接模式更接近随机，没有明显倾向于同配性或异配性(Newman, 2002)。真实世界的许多网络存在同配现象(Newman, 2003a)；而大多数语言网络则呈现出异配性(刘海涛, 2017)。

3 结果

3.1 网络基本概貌

根据上文所述的网络构建方法，绘制了6个水平的汉语中介语词同现网络图，见图2。

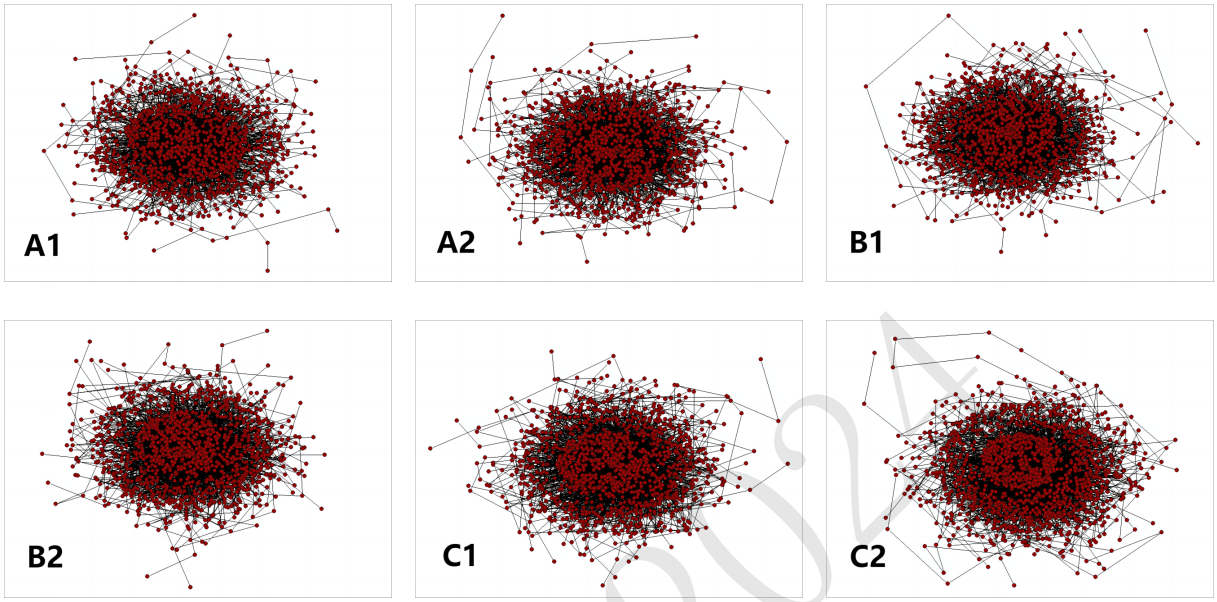


Figure 2: 汉语中介语词同现网络图

图2中的点表示网络的节点，即词型；节点与节点之间的线则表示网络的边，即两个词之间的同现关系。通过图2可以直观地观察到中介语词同现网络的复杂性。然而，由于节点和边的数量庞大，直接从图示角度分析网络结构变得相当困难。因此，必须借助相应的网络数据进行深入的分析。

表4提供了6个汉语中介语词同现网络和汉语母语词同现网络的基本信息，包括网络的节点数、边数、词数、型例比（TTR）和密度。其中，节点数就是网络的词型数（word types），词数表示网络的词例数（word tokens）。

网络	A1	A2	B1	B2	C1	C2	母语
节点数	1193	1244	1225	1290	1342	1563	1926
边	3518	3631	3830	3876	3938	4069	4262
词数	5020	5014	5018	4987	5048	5021	4981
TTR	0.238	0.248	0.244	0.259	0.266	0.311	0.387
网络密度	0.0049	0.0047	0.0051	0.0047	0.0044	0.0033	0.0023

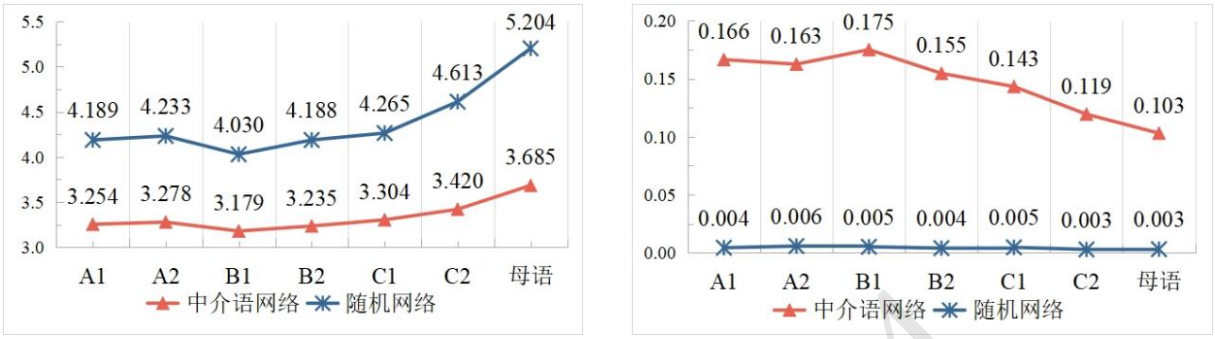
Table 4: 母语语料与中介语语料抽样信息

在相同的文本规模下，TTR可以有效衡量文本的词汇丰富程度(Richards, 1987)，由表4可知，随着学习者能力水平的提升，汉语中介语词同现网络的节点数和TTR的值都在逐渐增长。学习者语言水平的提升，书面写作中会使用更加丰富的词汇。不过，这种增长并非线性，节点数和TTR的值都在A1、A2和B1水平下产生了波动，而B1到C2则是线性增长。除此之外，6个水平的网络节点数和TTR的值都未超过母语者网络。

网络密度的变化与TTR正好相反，数值随着学习者能力水平的提升而波动下降，中介语网络6个水平的网络密度均高于母语网络。这表明，中介语网络要比母语网络更加紧密，而且从低水平到高水平，网络密度趋于稀疏，并逐渐向母语网络靠拢。

3.2 网络的小世界属性

平均路径长度和聚集系数是衡量网络是否具有小世界属性的重要参数。当网络同时具有较高的聚集系数和较短的平均路径长度时，这种网络就是一种小世界网络(Watts and Strogatz, 1998; Albert and Barabási, 2002)。中介语网络与随机网络的平均路径长度和聚集系数详情见图3。



(a) 平均路径长度　　(b) 聚集系数

Figure 3: 小世界属性统计结果

由图3可以清晰的看出，与随机网络相比，同一水平、相同规模下中介语网络都拥有更短的平均路径长度和更大的聚集系数，6个中介语网络均具有小世界属性。小世界属性表明，中介语中的词汇通过同现关系形成了高度结构化的网络。这种结构化表明词与词之间的同现搭配不是随机分布的，而是按照某种内在的逻辑和规律相互关联。

小世界属性揭示了中介语系统内部语言要素是紧密联系的。这种结构可能反映了学习者在二语习得过程中，对词汇和句法结构内在逻辑的适应和理解。Corominas-Murtra et al(2009)曾发现儿童一语者在24个月时，句法图的全局拓扑会从树状结构急剧转变成小世界网络，出现一种“句法涌现”。而Jiang et al(2019)则发现，英语二语学习者句法网络并没有像一语者那样出现“句法涌现”，这表明学习者习得一语句法和习得二语句法的方式是不同的，二语学习者会依靠母语和目的语之间的句法相似性来接近目的语。本研究也有类似地发现，汉语中介语词同现网络中，最低水平（A1）的中介语网络已经出现了小世界属性。即使是在非母语的使用中，低水平的汉语学习者也能有效地构建词汇间有意义的同现搭配，形成一个紧凑而高效的网络。

除此之外，我们还发现平均路径长度和聚集系数的波动与节点数和TTR类似，从B1到C2水平都是是线性增长或下降，而波动主要发生的在A1、A2和B1水平。

3.3 网络的无标度属性

网络研究中，随机网络的度分布通常都服从泊松分布（Poisson distribution），而真实网络的度分布通常都服从幂律分布（Power-law distribution），当一种网络的度分布服从幂律分布时，通常把这种网络称作无标度（scale-free）网络(Barabási, 2002)。为了探究汉语中介语词同现网络是否具有无标度属性，本研究提取了6个中介语网络的度数，并在双对数坐标轴中观察度的分布服从何种分布，详情见图4。

由图4可知，6个网络度分布的幂律拟合决定系数R^2均在0.9以上，这表明6个中介语网络的度分布均服从幂律分布，汉语中介语词同现网络是无标度网络。无标度网络是一种特殊类型的网络，由于度分布遵循幂律分布，网络中存在少数几个高度连接的节点，大多数节点则只有少数的连接(Barabási and Albert, 1999; Newman, 2003b)，即无论学习者处于低水平还是高水平，中介语网络中始终存在一些高度连接的中心枢纽节点。无标度网络的特性也反映了中介语系统的适应性和潜在的演化路径。随着学习者语言能力的提高，新的词汇可能成为新的中心枢纽节点，原有的网络结构可能发生调整，以适应更复杂的表达需求。

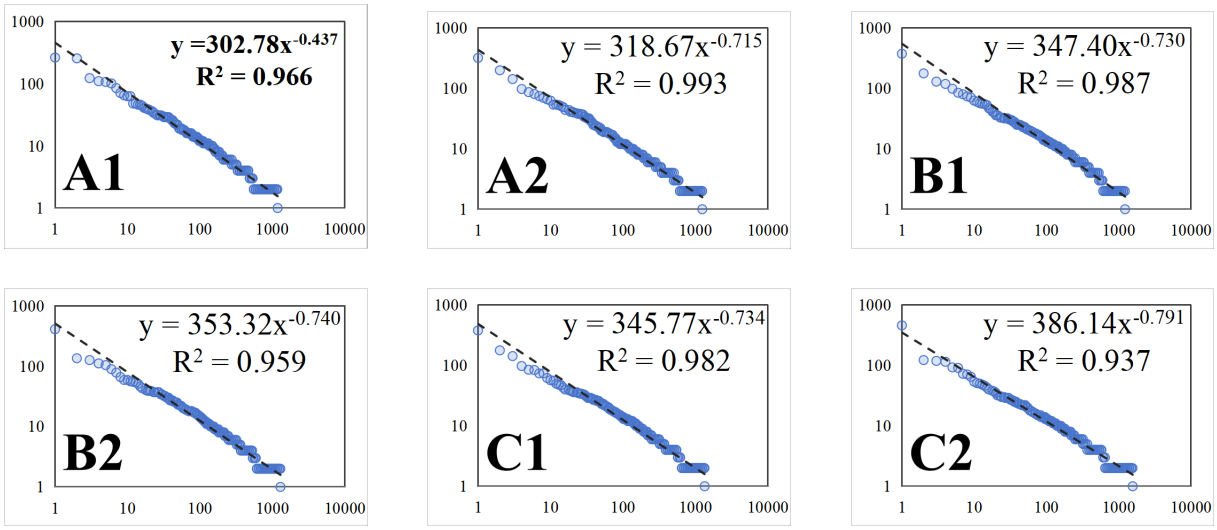


Figure 4: 网络度分布与幂律拟合结果

	A1		A2		B1		B2		C1		C2	
词序	节点	度数	节点	度数	节点	度数	节点	度数	节点	度数	节点	度数
1	我	264	的	321	的	376	的	409	的	380	的	461
2	的	256	我	200	我	178	我	135	我	178	了	123
3	是	123	是	143	了	130	是	126	了	143	我	119
4	了	110	有	98	是	119	了	111	是	98	是	115
5	她	107	他	87	很	99	有	104	很	85	和	93

Table 5: 母语语料与中介语语料抽样信息

表5展示了6个中介语词同现网络各自度数最高的5个节点。由表5可知，“我”“的”“是”“了”“她”“有”“他”“很”“和”是中介语中最常见的中心枢纽节点。其中，“我”“的”“是”三个节点在6个网络中均出现。不过，中心枢纽节点的位置也并非一成不变。从A1到C2可以明显地观察到代词“我”“她”“他”的地位在下降，助词“的”“了”的地位在上升，其中“的”是A2到C2，5个网络最中心的枢纽节点。

3.4 同配系数

同配性和异配性是网络中节点连接的两种不同的方式。一般学者合作网络、电影演员合作网络、公司董事合作网络等都具有同配性，即高度节点和高度节点连接，低度节点和低度节点连接；而因特网、万维网、神经网络等则具有异配性，即高度节点通常和低度节点连接(Newman, 2002)。图5展示了中介语网络和母语网络以及其对应的随机网络的的同配系数。

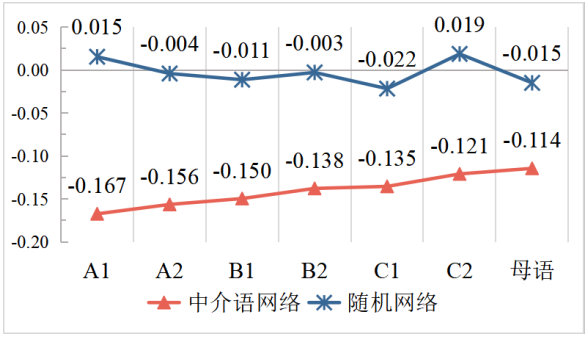


Figure 5: 同配系数

由图5分析得知，六个中介语网络的同配系数均呈现负值，揭示出不同层次的汉语中介语词汇共现网络均展现出异配性特征。这表明在中介语网络中，高度节点与低度节点的连接模式更为常见。与此形成对比的是，六个随机网络的同配系数则趋近于0，其值在0的附近波动，表明随机网络中节点的连接方式较为无序，没有表现出明显的同配或异配倾向。值得注意的是，尽管所有中介语网络都显示出异配性，同配系数的数值却随着学习者水平的提高而逐渐上升（绝对值降低），中介语的同配系数逐渐向母语者靠拢。这表明，在初级水平的网络中，异配性特征更加明显，而在高级水平的网络中，这一特征则相对减弱。

3.5 层级结构

网络的层级结构（Hierarchical organization）是一种特定的网络拓扑，其中节点按照不同的层次或级别进行组织，形成了明显的上下级关系。当度相关聚集系数$C(k)$满足幂律分布时，网络往往呈现出层级结构。为了探究词同现网络是否呈现层级结构，本研究计算了网络中每一个度数的平均聚集系数，并在双对数坐标轴中观察度相关聚集系数的分布服从何种分布，详情见图6。

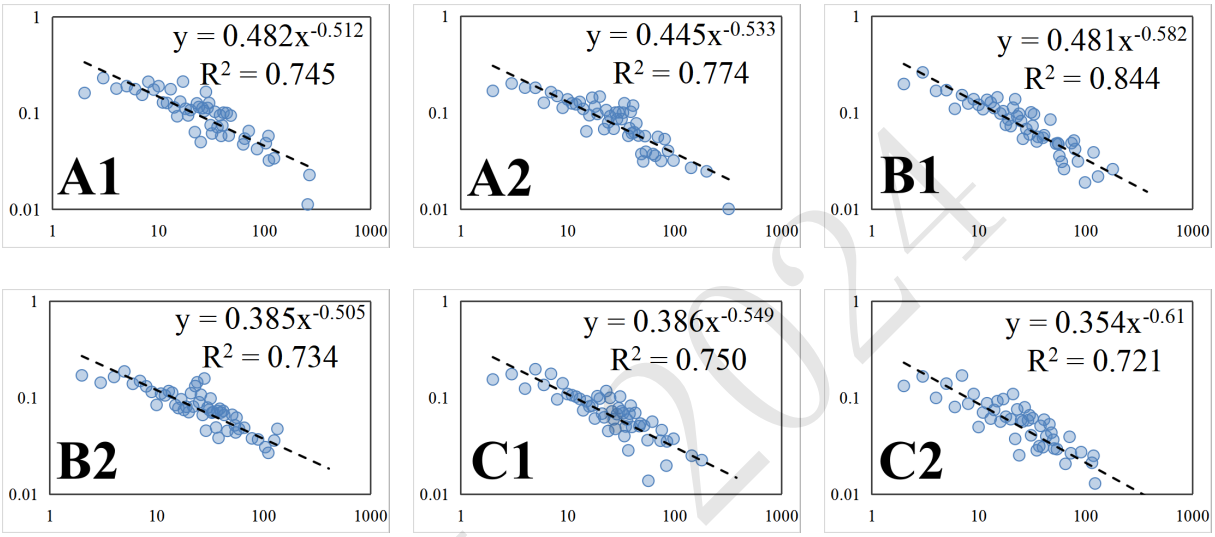


Figure 6: 度相关聚集系数拟合结果

由图6可知，六个网络度相关聚集系数的幂律拟合决定系数R^2均超过了0.7，拟合结果在可接受范围内。六个中介语网络均呈现出层级结构。这表明，学习者在语言使用方面即便处于初级水平，词汇之间的搭配也已经较为紧密，呈现出明显的组织和结构。这可能说明学习者在初级水平已经形成对目的语的系统性认识，建立了基本的词汇和语法体系。

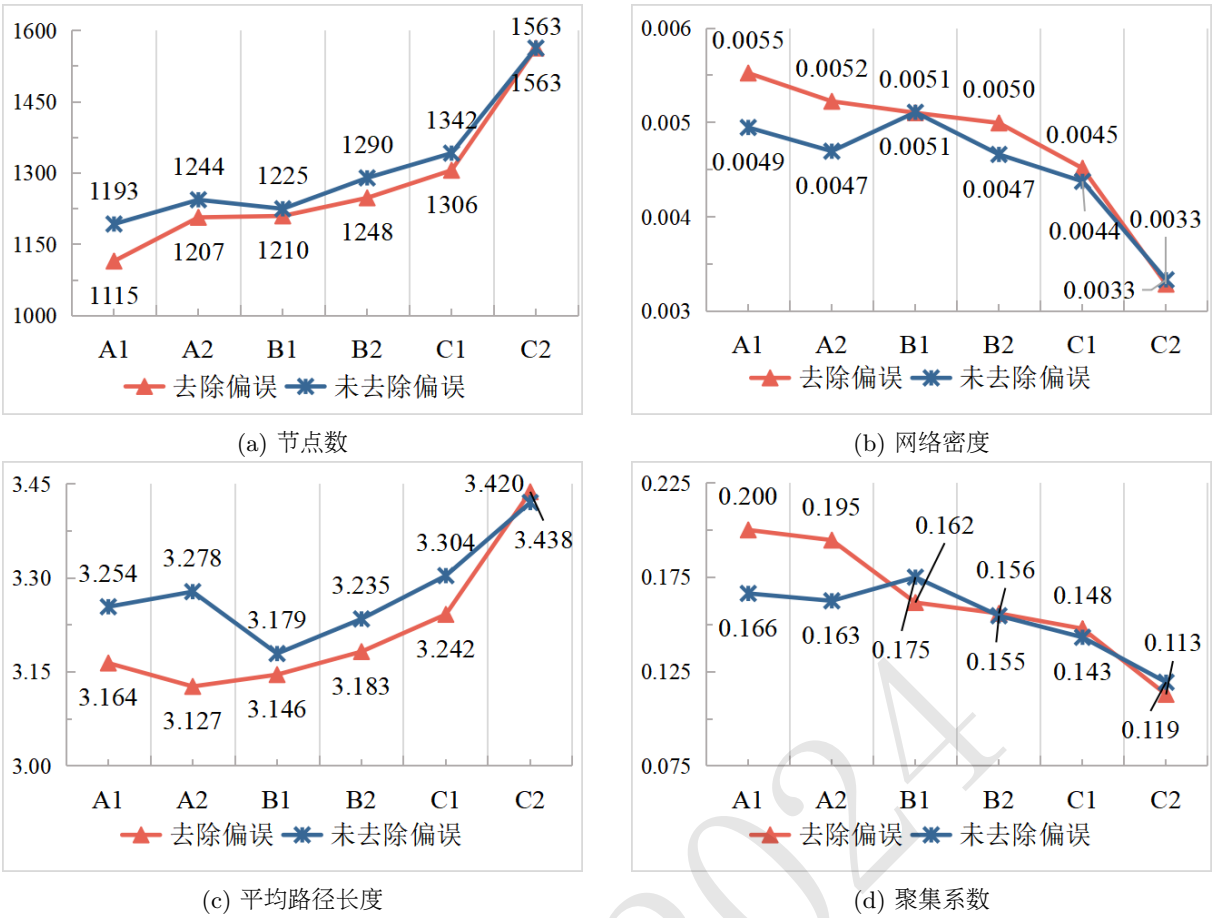
3.6 纠正语言错误后的中介语网络

由表4和图3、5可知，六个不同能力水平的网络在节点数、网络密度、平均路径长度以及聚集系数等关键参数上，并未显示出一致的线性变化趋势。在A1、A2和B1等级上，参数数值出现了波动。这种规律性波动可能与学习者在语言使用中的错误有关。鉴于此，本研究对所有网络中的语言错误进行了人工校正。中介语网路中错误数量见表6。

网络	A1	A2	B1	B2	C1	C2
错误数量	549	374	327	235	144	72

Table 6: 中介语网络中语言错误的数量

由表可知，网络规模相同，学习者的能力水平越低，网络中包含的语言错误就越多。为了探究这些语言错误对中介语网络产生了怎样的影响，我们统计了纠正语言错误后的节点数、网络密度、平均路径长度和聚集系数。纠正错语言误后的网络参数见图7。



(a) 节点数　(b) 网络密度

(c) 平均路径长度　(d) 聚集系数

Figure 7: 纠正语言错误后的网络参数

由图7可知，经过人工纠正网络中的语言错误后，节点数、网络密度和聚集系数这三项指标的开始表现出线性变化趋势。尽管平均路径长度的数据并未完全遵循线性变化，但在A1、A2、B1水平，这项指标的波动幅度已有所降低。可见，网络中的语言错误实际对网络的变异产生了影响。

4　讨论

4.1　学习者词汇同现搭配的动态演变

从网络分析的数据中，可以清晰的观察到日本汉语学习者词汇同现搭配的动态演变过程。前文分析中提到，中介语词同现网络表现出小世界属性和异配性特征。从A1到C2水平，中介语网络的平均路径长度增加，聚集系数下降，异配性特征减弱。这三个参数有效反映学习者的写作词汇丰富度和句法水平的演变过程。在A1、A2和B1水平，学习者的词汇量和句法手段都很有限，表达不够灵活自由。为了传达正确的信息，学习者只能在有限的词汇间进行搭配和组合，节点和节点之间就会产生更多的连接，形成紧密连接的群组，即网络中存在许多相互紧密连接的三角结构（三元闭包），从而导致网络具有更大的聚集系数。并且，网络之间有限的搭配和组合也非常依赖高度节点的参与，高度节点通常都是学习者最早掌握并熟知的词，通过迂回的表达使用高度节点，学习者才可以在写作中传达出更多的信息。这种紧密的节点连接和对高度节点的依赖也意味着任意两节点之间可以通过更少的中间节点互相到达，从而导致较低的平均路径长度。随着学习者能力水平的提升，其掌握的词汇越来越多且越来越丰富，使用的句法结构也更加多样，写作表达更加灵活、自由，不再依赖于有限的词汇或固定搭配，尤其是对高度节点的依赖，这使得网络中节点间的连接更加分散，网络的聚集系数逐渐减小，同配系数逐渐上升，以及平均路径长度逐渐增加。这些参数的变化共同揭示了学习者在逐步接近目标语言过程中，从简单到复杂的语言使用发展轨迹，反映了学习者在不断拓展其语言表达的广度和深度。

第二十三届中国计算语言学大会论文集，第712页-第725页，太原，中国，2024年7月25日至28日。
卷1：主会论文

4.2 学习者语言习得的认知机制

根据前文的分析可知，相较于母语者，二语学习者的中介语网络呈现出更紧密的连接模式，网络连接更加依赖高度节点，且中介语网络的参数趋近而不超越母语网络。这一现象可能反映了学习者在二语习得过程中采取的策略。

中介语网络具有无标度属性，而无标度网络通常通过“优先连接”（Preferential Attachment）机制来增加新的节点(Barabási and Albert, 1999)，即新节点倾向于与高度节点节点建立连接。“优先连接”会导致网络出现异配性特征。这种网络特性可能反映了学习者在学习过程中的一种自然倾向：将新学习到的词汇与已经熟悉的、使用频率较高的词汇相结合。这可以被理解为认知过程中的一种“锚定”(Tversky and Kahneman, 1974; **?**)。通过将新知识与既有知识关联，学习者能够促进知识的巩固和新知识的内化。然而，这种“锚定”也可能成为障碍，导致学习者过度依赖已有知识，使得其语言水平难以完全达到母语水平。

在二语习得领域，锚定可能与学习者减轻认知负荷的习得模式有关。认知负荷是指在一个特定的作业时间内施加于个体认知系统的心理活动总量(Sweller, 1988)，其假设人的认知结构由工作记忆和长期记忆组成，信息只有先被工作记忆注意和处理后，才能存储在长期记忆中。工作记忆是大脑中负责临时存储和处理信息以执行认知任务的记忆系统，与语言理解、产出和习得等活动密切相关(Bergsleithner, 2010; Wilhelm et al., 2013)。工作记忆理论强调，人们在信息处理过程中可利用的工作记忆容量是有限的；因此，有效的信息组织与处理机制对学习者尤为关键。在学习新词汇时，二语学习者需要依赖工作记忆来暂时存储和处理信息。因此，他们倾向于将新的词汇与已经存在于长期记忆中、容易连接的高频词汇进行搭配编码。这样做可以减轻认识负荷，使得这些词汇更容易在工作记忆中被保持和处理，从而提高学习的效率。一旦这种搭配信息被编码，它就被存储在大脑的长期记忆中。当需要使用长期记忆中的搭配信息时，大脑检索和提取的负担也会更小。即便学习者的语言能力得到提升，这种以节省认知负荷为导向的学习模式仍在其语言习得过程中占据主导地位。例如在C2水平上，学习者所表现出的异配性特征明显超过母语者。这一现象表明，节省记忆负担在二语习得的全过程中都发挥着重要作用。同样地，中介语网络呈现出的层级结构特性，亦反映了学习者减轻认知负荷的学习模式。即使在A1水平，中介语网络已展现出明显的层级结构，该结构提供了一种高效的信息组织模式，使学习者能够通过分层组织词汇的同现搭配，有助于降低语言学习者的认知负担，优化认知资源的使用。

4.3 语言错误在网络变异中产生的影响

汉语中介语的语言错误，通常按性质被划分为四大类，分别是遗漏、误加、误代和错序(鲁健骥, 1994)。在写作中，又可以从语言层面上被分成字错误、词错误和句错误(鲁健骥, 1992)。

在中介语网络中，字错误可视为是一种节点内部的错误。这类错误包括错字、别字、漏字、多字等。字错误通常会在网络中引入新的节点。例如，在A2网络中，“时候”这个词，虽然多数学习者正确书写了“时候”，但仍有少数学习者将其错写为“候”。在未考虑书写错误的情况下，“时候”和“候”被视为了不同的节点，从而导致网络节点数量的增加。观察图7（a）也能发现，通过修正这类语言错误，网络的节点数有所减少，纠正后的网络比原网络具有更低的节点数。而中介语网络又是一个复杂自适应系统，语言错误产生的这些新节点，通常又都是低度的，低度节点的涌现，又降低了网络的密度。并且这些低度节点之间相互通达的距离拉长，导致平均路径长度被拉长。同时，低度节点通常都是网络的边缘节点，无法与邻接点形成三元闭包，进而降低了网络的聚集系数。日本汉语学习者在写作中常出现字错误，部分原因在于汉语和日语共用汉字作为文字系统，但两种语言中的汉字并不完全一致。这些易混淆的汉字成为日本汉语学习者常犯字错误的一个原因。例如，汉语中的“年龄”与日语的“年”，汉语的“东京”与日语的“京”都不同，这类错误在网络中增加了大量低度节点，引起网络结构的变异。

词错误不仅影响节点和边的数量，还可能导致节点间的错误连接，包括错词、缺词和多词三种情况。错词是指学习者使用了不恰当的词汇；缺词则是学习者在句子中漏写了必要词；多词则是学习者在书写时多写了不必要的词。错词和多词在网络中都会引入不正确的节点连接，这不仅提高了某些节点的度数，还可能人为地增加了局部聚集系数，因为这些错误连接可能形成了实际上并不存在的三元闭包。此外，错词和多词错误通过引入额外的边，也有可能提升了网络的整体密度。这类错误连接还可能导致两个原本不直接相关的节点之间形成较短的路径，进而减少了网络的平均路径长度。然而，缺词对网络结构的影响则可能与之相反，网络中缺少

了原本应当存在的节点，打断了原有的正确连接，因而很有可能导致网络密度的下降，局部聚集系数的降低以及平均路径长度的增加。总而言之，词错误对中介语网络结构的影响是多方面的，它们通过改变节点和边的数量及连接方式，进一步影响网络的统计特性和拓扑结构。这对于理解学习者语言能力的发展和中介语的特性有重要意义。

汉语和日语的语序有很大的不同。日语是黏着语，修饰谓语的词语在句中的位置有较大的自由度(崔立斌, 2001)；而汉语是孤立语，缺乏严格意义上的形态变化，语序则比较固定(叶蜚声、徐通锵, 2015)。因此，受语言迁移的影响，语序错误成为日本汉语中介语中最常见的句错误类型之一(崔立斌, 2001)。在网络分析中，语序错误可被视为关系错误，即边的错误。语序错误对语言网络结构的影响较为复杂，且不易直接从网络的统计特征中捕捉。这类错误一般不会直接增加或减少网络中的节点与边的总数，而是将节点间正确的同现连接改变成错误的同现连接。错误连接的形成以及正确连接的消失，二者共同改变了网络的拓扑结构，这可能会在网络的某些区域造成结构扭曲，从而增加了对语言网络理解的难度。错误的上下文建立连接，可能会导致母语者对中介语内容的理解产生偏差，这也解释了为什么语序错误对作文的可读性影响较大。

值得注意的是，尽管偏误造成了中介语网络的变异，但它们并未改变网络的根本属性，这一点在偏误数量最多的A1网络中尤为明显。A1网络包含的549个偏误，占整个网络词例规模的10.9%。即便如此，A1网络仍然展现出小世界属性、无标度属性以及层级结构特性。这表明，尽管偏误的比例相对较高，它们并未根本影响网络的基础属性，这反映了中介语网络在面对语言偏误时，具有一定的鲁棒性。

5 结语

本研究通过深入分析汉语中介语词同现网络，揭示了汉语学习者在汉语二语习得过程中的词汇搭配模式和网络结构变化。本文发现：低水平学习者的中介语网络已经表现出了小世界属性、无标度分布、异配性和层级结构特征，这些特性随着学习者语言水平的提升而逐渐变化。网络分析结果显示：低水平学习者的网络更加紧密并且高度依赖高度节点，反映了学习者减轻认知负荷的习得模式。随着语言水平的提升，学习者开始探索更广泛的词汇组合，网络结构变得更加分散，聚集系数下降，异配性特征减弱，而平均路径长度增加，表明学习者的词汇使用和句法结构变得更加复杂和成熟。此外，语言错误对中介语网络结构的变异也产生了深刻的影响，这些错误不仅增加了网络的节点数，而且影响了网络的密度和聚集系数。

总的来说，汉语中介语词同现网络研究为我们提供了一个全新的视角，可以帮助我理解二语学习者在汉语习得过程中展现的语言习得特征与策略。未来的研究，我们期望通过对比分析不同国别背景下学习者的中介语网络特性及其演变的差异，来揭示多国别背景下学习者的语言发展路径与习得机制，从而为汉语二语习得研究提供更加广泛的实证支持。

参考文献

R Albert and A L Barabási. 2002. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

A L Barabási and R Albert. 1999. Emergence of scaling in random networks. *science*, 286(5439):509–512.

A L Barabási. 2002. *The new science of networks*. Cambridge MA. Perseus.

JM Bergsleithner. 2010. Working memory capacity and l2 writing performance. *Ciências & Cognição*, 15(2):2–20.

J Borge-Holthoefer and A Arenas. 2010. Semantic networks: Structure and dynamics. *Entropy*, 12(5):1264–1302.

EJ Briscoe. 1998. Language as a complex adaptive system: co-evolution of language and of the language acquisition device. In *Proceedings of eighth computational linguistics in the Netherlands Conference*, volume 4.

L Cameron and D Larsen-Freeman. 2007. Complex systems and applied linguistics. *International journal of applied linguistics*, 17(2):226–240.

RFI Cancho, RV Solé, and R Köhler. 2004. Patterns in syntactic dependency networks. *Physical Review E*, 69(5):051915.

J Cong and H Liu. 2014. Approaching human language with complex networks. *Physics of life reviews*, 11(4):598–618.

B Corominas-Murtra, S Valverde, and R Solé. 2009. The ontogeny of scale-free syntax networks: phase transitions in early language acquisition. *Advances in Complex Systems*, 12(03):371–392.

L Danon, AP Ford, T House, et al. 2011. Networks and the epidemiology of infectious disease. *Interdisciplinary perspectives on infectious diseases*.

J Davidsen, H Ebel, and S Bornholdt. 2002. Emergence of a small world from local interactions: Modeling acquaintance networks. *Physical review letters*, 88(12):128701.

NC Ellis and D Larsen-Freeman. 2006. Language emergence: Implications for applied linguistics—introduction to the special issue. *Applied linguistics*, 27(4):558–589.

NC Ellis and D Larsen-Freeman. 2009. Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Language learning*, 59:90–125.

M Garg and M Kumar. 2018. The structure of word co-occurrence network for microblogs. *Physica A: Statistical Mechanics and its Applications*, 512:698–720.

SA Greenberg. 2009. How citation distortions create unfounded authority: analysis of a citation network. *Bmj*, 339.

Y Hao, X Wang, M Wu, et al. 2021. Syntactic networks of interlanguage across l2 modalities and proficiency levels. *Frontiers in Psychology*, 12:643120.

J Jiang, W Yu, and H Liu. 2019. Does scale-free syntactic network emerge in second language learning? *Frontiers in Psychology*, 10:418102.

M Levine and EH Davidson. 2005. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences*, 102(14):4936–4942.

J Li and J Zhou. 2007. Chinese character structure analysis based on complex networks. *Physica A: Statistical Mechanics and its Applications*, 380:629–638.

W Liang, Y Shi, CK Tse, et al. 2009. Comparison of co-occurrence networks of the chinese and english languages. *Physica A: Statistical Mechanics and Its Applications*, 388(23):4901–4909.

H Liu and J Cong. 2013. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 58:1139–1144.

H Liu. 2008. The complexity of chinese syntactic dependency networks. *Physica A: Statistical Mechanics and Its Applications*, 387:3048–3058.

BH Mayhew and RL Levinger. 1976. Size and the density of interaction in human aggregates. *American Journal of Sociology*, 82(1):86–110.

MEJ Newman. 2002. Assortative mixing in networks. *Physical review letters*, 89(20):208701.

MEJ Newman. 2003a. Mixing patterns in networks. *Physical review E*, 67(2):026126.

MEJ Newman. 2003b. The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

MEJ Newman. 2010. *Networks: an introduction*. Oxford University Press.

Council of Europe. 2018. *Common European Framework of Languages: Learning, Teaching, Assessment*. Cambridge University Press.

GA Pagani and M Aiello. 2013. The power grid as a complex network: a survey. *Physica A: Statistical Mechanics and its Applications*, 392(11):2688–2700.

E Ravasz and AL Barabási. 2003. Hierarchical organization in complex networks. *Physical review E*, 67(2):026112.

B Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

J Scott. 2000. *Social Network Analysis: A Handbook*. SAGE Publications.

Larry Selinker. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(1-4):209–232.

J Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.

A Tversky and D Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

S Wasserman and K Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press.

DJ Watts and SH Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

DJ Watts. 2003. *Six degrees: The science of a connected age*. WW Norton & Company.

O Wilhelm, A Hildebrandt, and K Oberauer. 2013. What is working memory capacity, and how can we measure it? *Frontiers in psychology*, 4:00433.

刘海涛. 2017. 计量语言学导论. 科学出版社.

刘海涛. 2022. 依存关系与语言网络. 科学出版社.

刘涛, 陈忠, and 陈晓荣. 2005. 复杂网络理论及其应用研究概述. 系统工程, (06):1–7.

叶蜚声、徐通锵. 2015. 语言学纲要. 北京大学出版社.

崔立斌. 2001. 日本学生汉语学习的语法错误分析与汉日语言对比. 语言文字应用, (04):20–24.

张宝林、崔希亮. 2022. "全球汉语中介语语料库"的特点与功能. 世界汉语教学, 36(01):90–100.

王士丛. 2021. 基于词同现网络的汉语中介语系统性实证分析. Ph.D. thesis, 华侨大学.

韩笑, 张亮, 张华, et al. 2021. 复杂网络视角的汉语二语口语句法复杂度发展研究. 世界汉语教学, 35(03):377–391.

鲁健骥. 1992. 偏误分析与对外汉语教学. 语言文字应用, (01):69–73.

鲁健骥. 1994. 外国人学汉语的语法偏误分析. 语言教学与研究, (01):49–64.