文章编号：1003-0077(2014)01-0009-10

语言网络研究进展

韩 普[1]，王东波[2]，路高飞[3]，苏新宁[3]

（1. 南京邮电大学 管理学院，江苏 南京 210023；
2. 南京农业大学 信息科学技术学院，江苏 南京，210095；
3. 南京大学 信息管理学院，江苏 南京 210093）

摘 要：语言网络作为一个新的研究领域，其研究正在迅速崛起，目前已经吸引了不少领域的研究者们的关注。该文首先简要介绍了语言网络的特点、常用的统计特征以及相关的网络模型；其次，根据语言构成单位以及当前语言网络研究热点，将语言网络分为语音网络、共现网络、依存句法网络、概念语义网络，并详细介绍了各类语言网络研究的主要进展。最后总结了语言网络研究的现状并给出了展望。

关键词：语言网络；小世界现象；无尺度分布

中图分类号：TP391 **文献标识码**：A

Research and Progress in the Language Network

HAN Pu[1], WANG Dongbo[2], LU Gaofei[3], SU Xinning[3]

(1. School of Management, Nanjing University of Posts & Telecommunications, Nanjing Jiangsu 210023, China;
2. College of Information and Technology, Nanjing Agricultural University, Nanjing Jiangsu 210095, China;
3. School of Information Management, Nanjing University, Nanjing Jiangsu 210093, China)

Abstract: As a new research field, the study of language network is developing rapidly. It has aroused great attention from researchers in different areas. Firstly, a briefly introduction is delivered to illustrate the characteristics of language network, statistical properties and the related network models. Secondly, based on the composition of language and the hot topic of network, language network is divided into phonetic network, co-occurrence network, syntactic dependency network, semantic and concept network. Besides, the main research content of language network are described in detail. Finally, the drawbacks and advantages of language network study are summarized.

Key words: language network; small-world phenomenon; scale-free distribution

1 引言

1998 年，*Nature* 上发表了 Watts 和 Strogatz 有关小世界网络的论文[1]，1999 年，*Science* 发表了 Barabasi 和 Albert 的随机网络的论文[2]，两篇文章在全球科学领域产生了巨大影响，被认为是复杂网络研究的里程碑。从数学的角度上讲，复杂网络起源于图论，数学界称 1736 年是图论历史元年，因为这一年瑞士数学家 Euler 发表了图论的首篇论文《哥尼斯堡七桥问题无解》。传统条件下，图论研究的顶点数量往往比较少，现代信息技术的出现，使得图论得到进一步的发展，借助现代信息技术，可以处理拥有几万甚至几十万节点的真实网络。大规模真实网络是人类社会发展需要解决的问题，正是源于社会的需求，复杂网络得到社会学、生物学、医学、物理学、经济学、信息科学、数学、计算机科学、交通等学科领域研究者的关注[3-4]。

在复杂网络的研究中，语言网络作为一个新的研究方向，正在悄然兴起。语言和文字是人类文明的起源，也是人类文明出现的两大标志，作为人类智慧的结晶，也是除了化石之外，最能体现悠久文明和灿烂文化的方式之一。据推测，人类目前有数千种语言[5]，传统语言学一般将其划分为 9 大语系。受

收稿日期：2012-02-07 **定稿日期**：2012-07-09

基金项目：863 计划项目(2011AA01A206)；国家自然科学基金(71273126)

到地域和文化的影响，同一语言也存在着分化现象。虽然语言种类繁多，但不同语种之间存在着一定的联系，目前的相关研究尚不能对以下问题进行解释，几千种语言之间是否存在共性？不同语言中的规律和渊源如何挖掘？一些小语种语言正在消失，所蕴含的人类智慧如何保留？1949年，哈佛大学语言学家Zipf发现了语言学中的Zipf定律[6]，这一定律最初在英语中发现，但随后的相关研究表明，其他语言一定程度上也符合Zipf定律[7-10]，虽然在部分语言中呈现的并不完美[9-10]。对语言研究来说，Zipf定律无疑是一个重大发现，它描述了词频和词序存在着一定联系，揭示了语言学中的静态规律，但如果将单词打乱，词频和词序依然可以满足Zipf定律，所以这个定律并不能解释人类语言更为复杂的问题。在语言学界，语言是一种网络的观点已经被普遍接受[11-12]，由于语言的特点，语言不仅是一种网络，还是一种复杂网络[13]。Cancho和Sole首次用复杂网络的方法研究了英语同现词网络。随后，不同语种中由不同语言单位及其关系构成的语言网络受到了关注。由于语言网络的跨学科特点，该领域吸引了一批语言学家、物理学家、生物学家和数学家参与其中。从已有的研究来看，语音、语素、词汇、短语在不同语言中构成的网络几乎均具有真实网络的一般统计特性，多数网络在整体上呈现出了典型的小世界特征和无尺度现象，与社会网络、生物网络、生态网络具有类似的特征。总的来说，目前语言网络的研究已经取得了一定的进展。本文将从语言网络的特点、常用统计特性、相关模型、语言网络的分类和研究进展进行论述。

2 语言网络常用统计特征

语言网络是复杂网络的子集，在语言网络研究中常借鉴复杂网络的研究方法。一般来说，度、平均最短路径长度、聚集系数以及中介度是语言网络常用的统计特征。

度：度是对节点而言，节点 i 的度即与该节点连接的其他节点的数目。语言网络通常是有向网络，根据节点的指向关系，度又分为出度和入度。节点度是语言网络最常用的统计参数，度分布是衡量一个网络无尺度现象的重要特征。

平均最短路径长度：在复杂网络中，节点 i 与 j 的距离 $d(i, j)$ 实际上就是连接节点 i 和节点 j 所需的最短路径长度。大部分真实网络都具有较小的平均最短路径长度 $<d>$，在语言网络中，$<d>$ 表示从一个语言节点到另一个语言节点所需要的平均最短路径长度，该参数是小世界网络判断的重要参数之一，常常用来和随机语言网络进行对比。

聚集系数：在图论中，聚集系数是图中点倾向于集聚在一起的程度的一种度量。对于语言网络，该参数呈现了与一个语言节点相连的其他节点中相互直接连接的概率。网络聚集系数可分为基于全局的和局部的，通常情况下，聚集系数是指全局平均聚集系数。该参数和平均最短路径长度一起用来判断小世界网络。

中介度：该概念源于分析社会网络中个体的重要性，1977年由Freeman提出[14]，他认为，如果一个节点处于多对节点之间，该节点的度可能会较低，但这个度较低的点可能会起到重要的中介作用，是网络中重要的节点。中介度衡量了一个节点位于其他节点之间的程度，表示其他节点对其依赖的程度。在语言网络中，陈芯莹和刘海涛认为，中介度测量的是一个点在多大程度上位于网络中其他点的“中间”，一个度数相对比较低的点可能起到重要的“中介”作用，因而处于网络中心[15]。一个节点中介度测量的是该节点对应的行动者在多大程度上成为“掮客”或者“中间人”，能在多大程度上控制其他节点。一个节点的中介度越大，表明大量语句将通过它，它的作用就越重要。

3 语言网络相关模型

语言网络具有哪些特征，属于什么类型，与其他类型的网络有哪些不同，这是语言网络研究首先要关注的基本问题。在语言网络研究中，多种网络被证明具有小世界模型和无尺度模型的特征，为了判断语言网络的类型，往往会与其他网络模型等进行比较。这里仅列出语言网络研究中常涉及到的几个模型。

随机网络模型：该模型是随机图论在网络中的进一步发展。随机网络是在给定一个概率 p 的情况下，对网络中任意两节点间的可能连接，都尝试以概率 p 进行连接。经典的随机网络模型是Erdös和Rényi提出的ER随机网络模型。真实的语言网络模型并不是ER模型，但在语言网络研究中，为了界定语言网络的类型，突出语言网络的特征，在整体特征统计分析时，往往与ER随机网络进行比较。客观世界中，一个真实网络具有小世界现象的一个体现是其最长路径长度 $D \approx D_{rand}$。小世界现象是真实

网络的一个重要特征，但真实网络与 ER 随机网络的一个重要区别是聚类系数 $C \gg C_{rand}$。

小世界模型：该模型是一个总称，当一个网络满足较高的聚集系数和较短的平均最短路径等条件时，便可以称为小世界网络。在语言网络中，小世界网络一般是指 1998 年由 Watts 和 Strogatz 在 *Nature* 中提出的基于人类社会网络的网络模型。他们最早生成了具有高聚集系数和最短路径长度的网络，该网络也称 WS 小世界模型。语言网络大都符合 WS 小世界模型，大多数节点只需经过少量的边便可到达。在聚集系数上，与随机语言网络相比，真实语言网络的聚集系数较高。

无尺度网络模型：无尺度网络是物理学领域的一个专业词汇，统计物理学家习惯于把服从幂律分布的现象称为无尺度现象，相应的网络称为无尺度网络。度分布是判断无尺度网络的重要特性，在大量的真实网络实验中，度分布呈现出无尺度现象，度分布一般对两边取 log 做图。其分布可用函数 P(k) 来描述，P(k) 表示的是一个随机选定的节点的度为 k 的概率。即 P(k) 为网络中度为 k 的节点占节点总数的比例，见式(1)。

$$P_k = \sum_{k'=k}^{\infty} P(k') \tag{1}$$

为了减小度分布曲线尾部噪音的干扰，也可以采用累积度分布函数 P_k 表示累积度的分布。大量真实语言网络被证明具有幂律度分布的现象，是一种无尺度网络。换言之，语言网络具有成长性和优先连接性，可以将分散的节点组织起来，形成稳定有意义的系统。

4　语言网络分类及研究进展

在复杂网络研究基础上，语言网络研究已经取得了一定进展。由于语言网络具有典型的跨学科特点，其研究分散在多个学科中，如语言学、数学、物理学、生命科学和信息科学。如何全面了解语言网络的当前研究成果和研究进展，对语言网络进行合理分类是必要的。目前语言网络并没有统一认可的分类，从不同的角度，可将语言网络划分为不同的类型。根据网络是否有向，可分为有向语言网络和无向语言网络；按照是否有权重，可分为加权语言网络和无权语言网络；按照网络构建来源是否真实语料，可分为静态语言网络和动态语言网络，如基于词典资源的静态语言网络，基于真实文本语料的动态语言网络。在当前多种语言网络研究基础上，从语言单位构成并结合目前语言网络主要关注方向，本文将语言网络划分为语音网络、共现网络、依存句法网络、语义概念网络，对于没有包含在 4 种网络中的，称为其他语言网络。目前语言网络繁杂，本分类中前四种网络可以涵盖大部分的研究，对于部分关注较少，或者仅在某一语言中存在的语言网络，即在 4 种语言网络之外的网络，本文一并称为其他语言网络，具体见图 1。

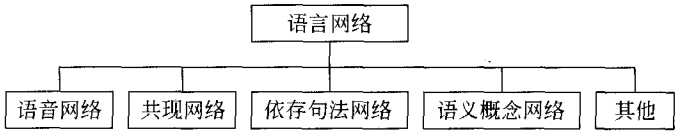

图 1　语言网络结构图

将语言网络进行合理的划分对语言网络研究具有重要意义，首先，通过对语言网络研究的系统梳理，有助于研究者全面了解语言网络当前的研究现状。其次，对于不同领域研究者，可以结合自己的研究方向，选择语言网络的一个或几个子领域，有针对性的深入研究。本文将对以上几类语言网络分别进行详细介绍。

4.1　语音网络

语音系统是人类重要的交流系统，在沟通交流中扮演着重要角色。从语言的观点来看，语音是最微观的范畴。音节是听觉能感受到的最自然的语音单位，音素是最小的语音单位或最小的语音片段，是音节的组成部分。虽然目前世界上有几千种语言，但音素数量却要小的多，不同语言的发音差异较大。语音系统在整体上有什么特点？是否有共性？对于这些问题，研究者从复杂网络的角度对音节、音素等语音网络进行了探究。

Medeiros 和 Corso 等基于葡萄牙语词典和作家作品全集，构建了葡萄牙语的音节网络[16]，网络节点为葡萄牙语音节，节点的连接以两个音节是否可以组成词为依据，统计参数表明该网络具有较高的聚集系数和较短的平均距离，该网络的幂指数 $\gamma \approx 1.4$，葡萄牙语音节的增长符合优先增长模型。

Peng和Minett等基于普通话词典和粤语词典，分别构建了普通话和粤语的基本音节网络和音调音节网络[17]，该方法与Medeiros采用的方法类似，以音节为节点，以两个音节是否可以组成汉语中的词建立音节之间的边。如“火车”汉语拼音为“huo3 che1”，粤语拼音为“fo2 ce1”，汉语音调音节网络中“huo3”和“che1”为节点，其相邻连接“huo3”和“che1”为网络的边，粤语音调音节网络构造与汉语类似，这些网络都表现出了随机网络所不具有的，但真实网络所具有的特征，度分布符合无尺度分布，具有较高的聚集系数，表明汉语音节网络是一种小世界网络和无尺度网络。Arbesman和Strogatz等基于词典研究了英文、中文等6种语言的音位网络[18]，发现音位网络具有与其他网络不同的特点，在度分布上介于指数分布和幂律分布之间。于水源、刘海涛利用汉语字典、汉语词典和两组真实语料，分别从字、词和句子的角度，以汉语音素为节点，相邻音素构造有向边，如“甘”包含三个音素“k”、“a”和“n”，共包含两个有向边“k→a”、“a→n”，构建了6种汉语音素网络[19]。结论发现音素网络有相当高的度和更短的平均路径，音素的度分布符合指数分布，但有权音素网络度分布符合无尺度分布，表明语音网络是一种稳定的网络系统。

语音网络主要以静态网络为主，动态语音网络研究较少。通过多组语音网络的研究，可以发现语音网络整体上的特点，多种语言之间呈现出了比较接近的特征，但与字词等其他类型语言网络有不同的特征，尤其是在度分布上。可以认为语音网络是一种特殊结构的网络，这种结构保证语音系统是一种高效并且有效的人类交流系统[19]，是人类语音在进化过程中逐步演变的结果。通过语音网络研究，有助于认识语音系统的组织结构，了解人类在语音上的认知机理以及语音交流系统的原理。

4.2 共现网络

共现网络是基于真实语料而构建的网络，不同语料构建的网络会有所差异。共现网络具有动态性，属于典型的动态网络。按照共现网络节点的构成，还可以进一步划分为字共现网络和词共现网络。词共现网络不论是在表意文字还是表音文字中均可构建，字共现网络存在于汉语等表意文字中。较早采用复杂网络方法构建的语言网络是英文词共现网络[11]。共现网络构造比较方便，尤其是对于英文等不需要分词的语言，非常容易构建词共现网络，不需要大量的语言学知识支持，只需考虑共现关系，相关的研究也比较多。

对于共现关系，也有不同的理解，最简单的共现是邻接关系，也可以将共现理解为在一个句子中同时出现。Cancho和Sole认为，在一个句子中出现的词是有关系的，多数共现关系是有语法联系的，最相关的词一定是距离最近的。他们基于BNC语料库，将同现的距离控制在2以内，构建了英语的共现词网络[11]，该网络平均最短路径在2.6左右，与随机网络相比，表现出明显的无尺度特性和小世界效应。Dorogovtsev和Mendes认为，相互连接的词可以用复杂网络来描述，并且根据句子中词的共现关系，提出了一个语言演变的模型[12]，该模型将语言视为词之间的自组织网络。Choudhury和Chatterjee等对涵盖了3大语系的7种语言[20]（英语、法语、德语、孟加拉语、爱沙尼亚语、印地语、泰米尔语）构建了词共现网络，通过整体拓扑特征进行深入比较，揭示了7种语言网络的共同特征，并进一步研究了共现网络的谱特征。

在中文词共现研究中，刘知远和孙茂松在1 300万词次的《人民日报》语料和5 000万字左右的人工分词语料库基础上构建了汉语的词共现网络[21]，得到汉语词共现网络的平均最短路径在2.63～2.75之间，聚类系数远大于相同参数下的随机网络，揭示了汉语在词共现网络上的小世界效应和无标度特性，表现出了与英语共现词网络类似的性质。Zhou和Hu等在1998年1月份的《人民日报》语料基础上，采用不同方法构造了两种汉语词无向同现网络[22]，一种是邻接距离为1的网络，一种是只要两个词汇在一个句子中同时出现，则认为两个词节点存在连接的网络，并且考虑了不同词性的情况，结果两个网络均呈现出小世界效应、无尺度特征、层次结构和负相关性，在整体特征上和其他语言网络相似。

和英语等表音文字相比，汉语是表意文字，在构建语言网络上有更多选择，在没有分词的情况下，还可以构成字共现网络。Peng和Minett等基于词典资源，根据汉语词汇中的共字关系构建了汉字网络[17]，由汉字构建的网络表现出明显的高聚集系数和无尺度特征。Liang和Shi等对散文、小说、科普文章、新闻报道4种体裁的中文和英文语料，分别构建了英文词共现网络、中文字共现网络和词共现网络[23]，从复杂网络角度揭示了3类语言网络的共性和个性，其共同之处是均满足无标度特征和小世界现象，不同之处在于从某种程度上英文的表达要比

中文更为简洁。Liang 和 Shi 等还对中国历史上不同历史时期的汉字网络进行了对比研究[24]，发现 99.6%的汉字网络具有无尺度特征度分布，95.0%的汉字网络有小世界的现象。Sheng 和 Li 构建了英文词共现和中文字共现的有权网络[25]，语料分别来自 George Orwell 英文版小说《一九八四》和中文版的《毛泽东传记》，结果发现两个网络不仅呈现出无尺度等共同特征，还呈现出显著的不同，中文字共现网络中高权重连接要高于英文词共现网络。

此外，词共现网络还被用来研究语言的演化，Ke 和 Yao 基于英语儿童对话语料，采用词共现方法构建了不同儿童的语言网络[26]，从网络视角研究了儿童语言的发展。

构建词共现网络需要满足一个重要假设，即 Cancho 和 Sole 在构建 BNC 英语词共现网络时的假设[11]，在一个句子中以邻接关系出现的两个词汇是有一定联系的。词共现网络的每一个节点都是有意义的语言单位，通过调节词共现的距离可以构造一个句子内部词汇之间的连接，虽然很难将词共现称为句法网络，但邻接词之间的确有一定的意义。字共现网络主要以汉语为代表，汉语是典型表意文字，具有独特和优美的结构，有强大的组合能力，古汉语中一个字往往可以表达一个完整的含义，但现代汉语由一个字表示完整词意的比较少，多是由组合词来表示词意。汉语字共现网络的构建可以像英文词共现网络一样，不需要分词处理，这对于汉语研究来说，所构建的网络更为客观，往往可以用来探索词汇的形成以及字词的演化。

4.3　依存句法网络

依存语法理论的创立者，法国语言学家 Tesnière 认为句子是一个有机的整体，词和邻近词会产生联系，这些联系构成了句子框架，并认为“谓语”中的动词是句子的中心，不受其他成分支配，其他成分与动词直接或间接地产生联系。图 2 是依存句法的示例。

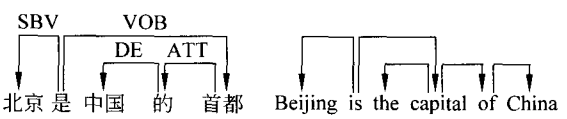

图 2　依存句法中英文示例

在图 2 中，箭头代表句法上的一种支配关系，支配者在箭头起点，被支配者在箭头终点。Cancho 和 Sole 等给出了依存句法网络（SDN）的描述[27]，SDN 是一种有向网络，每个单词构成网络的节点，网络的有向边由存在依存关系的词进行连接。Cancho 在依存句法基础上，构造了 3 种欧洲语言（德语、罗马语、捷克语）依存句法网络[28]，从度分布、层次组织、中心性、聚集系数和负相关性等几个统计特性进行了分析，发现 3 种语言的句法网络具有与其他语言网络类似的特性，并且在一些细微的模式上表现出同质性。

刘知远、郑亚斌和孙茂松利用清华大学 100 万词的句法标注树库，在依存句法基础上，构造了汉语依存句法有向网络[29]，得到了汉语依存句法网络平均路径长度$<d>=3.8$，聚集系数 $C=0.13$，出度和入度的累积度分布均具有无尺度特征。刘海涛对 20 种语言的依存句法网络进行了统计，发现相邻接的词只有 50%左右是语法相关的[30]，并认为用依存句法来构造语言句法网络是最合适的[30]，用邻接词构建的语言网络没有充分的语言学知识支持，缺乏合理的解释。刘海涛在新闻联播和实话实说两种不同体裁的汉语依存树库基础上[31]，以依存句法为基础[32]，构造了不同体裁的汉语语言依存无向网络，该网络中，每个词作为节点，每个支配关系作为网络的边。通过对网络直径、平均最短距离、聚集系数和幂律等参数的统计分析，表明两个汉语依存网络均属小世界网络，其度分布符合无尺度特征，平均最短路径长度在 3 左右，累积度分布的幂律指数分别为 2.40 和 2.18，与刘知远等对汉语依存句法网络研究结果非常接近。刘海涛还对中文、英文等 15 种语言的依存句法网络进行聚类研究[33]，得到平均最短路径长度在 2.755—3.938 之间，幂律指数在 1.077—1.353 之间。虽然 15 种语言网络均是小世界网络，但聚集系数存在显著差异，通过选择网络的 7 组特征，对 15 种语言进行了聚类，发现英语和汉语网络整体上比较接近。

基于依存句法的复杂网络研究已经不仅仅局限于网络的整体特征，目前已有部分研究深入到网络内部。Cancho 和 Capocci 等利用谱方法对依存句法网络中的名词、动词、形容词和副词等节点进行了聚类[34]，发现句法功能相同的词汇会被聚到同一个簇。Čech和 Mačutek 等认为动词在句法依存网络中有重要作用[35]，他们对 6 种语言构建了依存句法网络，探索了网络结构中动词的作用，认为句法对依存句法网络的拓扑特性有显著影响。此外，由于汉语与英语等其他语言的差异性，虚词在汉语表达中有重要的作用，陈芯莹和刘海涛研究了汉语复杂网

络的中心节点[15]，以汉语中两种体裁的语料构建依存句法网络，研究发现，在汉语依存句法网络中，“的”是全局中心节点，“了”是局部中心节点，“在”接近全局中心节点，汉语词的入度比出度在句法结构的完整性上更为重要。Čech和 Mačutek 分别以单词原形作为节点和词形还原后词根作为节点比较了捷克语依存句法网络的不同，发现两种不同的句法网络在整体特征上存在差异[36]。

依存网络句法结构本身简便，与共现网络构建相比，依存网络的构建需要依存句法标注，显得稍微复杂。和共现网络相比，依存句法能够较好展现词与词之间的句法关系。虽然基于词共现的语言网络比较易于构建，但却忽略了词与词之间的句法和语义的关系。基于依存句法构建的语言网络比基于词共现构建的语言网络更具有语言学特征，依存句法网络更容易获得语言学领域的认可。目前依存句法网络研究和其他语言网络研究一样大多停留在宏观层面上，需要进一步的深入探索。另外，依存句法网络也存在一些问题，一方面，依存句法网络需要依存句法分析，单纯的依存句法损失了节点的顺序关系，不利于语言的生成；另一方面依存句法构建的网络和人脑的认知网络是否最为接近，还有待进一步探究和证明。此外，句法网络和词共现网络在整体的特性上也有很多相同之处，其原因也有待于进一步探索。

4.4 语义概念网络

语义概念网络是从语义层面上构建的较为深入的语言网络。根据网络构建资源的不同，语义概念网络可分为静态语义概念网络和动态语义概念网络。静态语义概念网络利用概念词典资源构建，动态语义概念网络基于真实标注语料构建。静态语义网络较为常见，该类型网络的一个典型特征是静态性，其构建资源并不是真实语料。根据词典资源的不同，还可以进一步划分，基于同义词词典可以构成同义词网络，基于概念词典可以构成概念网络。词典资源便于获取并且精确度相对也比较高，相关的研究较多。

Sigman 和 Cecchi 基于 Wordnet 概念词典，构造了基于 WordNet 中名词语义网络[37]，该网络以词典中的名词为节点，以名词之间的 4 种连接关系（上位关系 hypernymy，反义关系 antonomy，部分关系 meronomy，一词多义关系 polysemy）作为语义网络的边，研究发现 WordNet 本身就是一个自组织系统，遵从无尺度分布，并发现一词多义对构建整个语义网络有重要作用。Motter 和 de Moura 等基于 Moby II 同义词词典，构建了英文概念网络[38]，该网络以单词为节点，以单词之间是否有同义关系构建网络的边，发现该网络具有较高的聚集系数（C＝0.52）和较短的平均路径长度（＜d＞＝3.16），具有典型的小世界特征，并且幂律分布呈现两个区间，在度分布高区间呈现出无尺度现象。Holanda 和 Pisa 等在 Motter 的基础上，进一步研究了同义词词典的构成[39]。Steyvers 和 Tenenbaum 等基于 WordNet、Roget 同义词词典和 Free Association Norms 词典分别构建了 3 种语义网络[40]，从最短路径、稀疏性、度分布等 5 种网络特性上对 3 个语义网络进行了对比分析，发现通过不同方式构建的语义网络，均呈现出真实网络的特征，具有小世界性和无尺度特征，根据语义网络呈现出的特点，作者还提出了一个简单语义网络增长模型。Tang 和 Zhang 等构建了基于 HowNet 的汉语语义网络[41]，发现基于 HowNet 的中文语义网络具有与 WordNet 和 Roget 词典网络类似的特征，具有较短的平均路径长度和较高的聚类系数，属于典型的小世界网络，具有无尺度现象，但在具体参数上与 WordNet 并不完全相同，存在一定差异。

基于词典的语义网络是静态的，所反映的现象并不完全是语言在真实交流过程中的呈现，但由于动态语义标注语料较困难，动态语义概念网络的研究较少。刘海涛通过对真实语料进行语义角色标注，构造一种节点为实词、连接为语义或论元关系的网络[42]，研究了汉语的动态语义概念网络的整体特征。虽然研究结果表明汉语动态语义网络也是小世界和无尺度的，但在一些特征上与依存句法网络和静态语义网络有所不同。

与共现网络和句法网络相比，语义网络是一种更为复杂的网络。静态语言网络反映了概念之间的语义关系，如同义关系、上下位关系等。静态语言网络可以从一定角度上通过揭示这些语义关系来研究人脑中知识网络的形成，对语义词典的构建和人类认识的探索有一定帮助。基于真实语料的动态语义概念网络，反映的是在真实环境中人类语言交流中的语义关系，可以用来研究语义产生的机理，深入了解动态的概念交流网络。

4.5 其他语言网络

尽管语言网络类型较多，但相关研究主要集中

在前面提到的 4 种网络上。除此之外，还有一些语言网络，关注度较少，或者仅存在于某一语言中。例如，汉语中的字结构网络，这在英语等表音文字中是不存在的。根据汉字的构成，Li 和 Zhou 对新华字典中 6 652 个汉字进行了拆解，构造了汉字的部首网络[43]，如“按”可以拆分成“扌”和“安”两个部首节点，由于两个部首可以组成汉字，那么这两个节点之间存在连接，研究揭示了汉字部首网络具有与其他真实语言网络同样的特性。另外，根据汉语词组的组成，Li 和 Wei 构建了汉字词组网络[44]，该网络将词组作为网络的节点，若两个词组节点中出现同一个汉字就认为它们有一条连接，如“网球”、“网络”、“络绎不绝”便可以构建 3 个节点两条边的词组网络，研究发现汉字词组网络的平均最短路径和聚类系数与英语单词网络类似，到达另一个词组的平均距离为 3，具有典型小世界特性。此外，王建伟和荣莉莉对清华紫光数据库中两个字组成的词构建了中文字网络[45]，他们以选取的 7 440 个汉字作为网络中的节点，以词中相邻汉字为网络的边，研究表明中文字共现网络具有真实网络的统计特性（$\gamma=1.15$，$C=0.451\,6$）。

通过对语言网络研究的系统梳理，我们发现，从语言最基本单位音素到句法结构，均可构建相应的语言网络。从各种语言网络的研究结果来看，依据不同方法、不同资源构建的语言网络几乎均属于小世界网络并且具有无尺度特征，与其他复杂网络具有类似的整体特征，但在具体特征参数上，存在着差别，这些共性和个性可以总结如下：

首先，语言是人类智慧的结晶，语言网络具有与随机网络不同的特征。通过多种语言网络的研究表明，无尺度特性和小世界现象在语言网络中普遍存在。语言网络的无尺度特征表明，在节点数量庞大的各种语言网络中，发挥着重要作用仅有少部分节点。小世界现象表明，语言网络和社会网络一样，一个节点到另外一个节点的最短距离往往很短。

其次，各种语言网络在整体上呈现出类似的特性，但不同语言网络之间存在着差别，如部分语音网络的度并不完全符合幂律分布，汉语音素无权网络的度呈现指数分布[19]。在语言网络其他统计特征上，也存在显著差异，如在凝聚度和最短路径方面，和静态语义概念网络相比，动态语义概念网络凝聚度偏低，平均最短路径较长，所组成的网络显得更为松散。对于动态语言网络，不同的体裁、语种构建的网络也有所区别，这些都表明语言网络不仅可以从整体上衡量语言的特性，还可以用来研究语言的个性化和相似性。

从音素、音节、字、词、短语、句法到语义、概念，语言网络研究层次在逐渐加深，但对于人类语言中的复杂问题依然没有进行很好的解释，哪种语言网络更贴近人类在语言交流时的语言系统，语言表达中词汇究竟是如何组织的，静态语义概念在人类大脑中如何存储，目前的语言网络研究还不能回答这些问题。

5　语言网络研究展望

作为复杂网络的一个子领域，语言网络刚刚出现 10 年左右的时间，已经在国际上产生了一定影响力的研究，受到了物理学、语言学、信息科学等多个领域的关注。总的来说，语言网络研究进展可以总结为以下几点。

1）语言网络研究开创了语言学研究新方向

作为一门以经验为基础的学科，语言学在 19 世纪中叶开始成为一项独立的研究，它是以其自身特征、规律作为学科对象进行研究的一门学科。语言学的研究方法主要以定性、定量或定性结合定量为主，复杂网络为语言学研究提供了一个全新的视角，借助现代信息技术，将语言作为一个系统，从整体和局部挖掘语言的规律，呈现语言节点之间的动态连接性，是对当前以字、词、短语、句子和篇章范畴的语言学研究的深化。

2）当前时期是语言网络研究的黄金时机

语言学规则是通过语言学专家根据经验和内省的知识总结，存在着一定的局限性。面对浩瀚的语言文本，只能窥一面而不能知全貌，信息技术可以为超级复杂网络的运算提供便利途径。此外，网络上大量的电子资源为语言网络研究提供了丰富语料来源。

3）语言网络已经取得了一定研究成果

从已有研究来看，语言网络研究已经发现了之前所没有关注的研究领域。将人类语言作为一个整体系统，揭示了语言作为一个有机系统具有真实网络的特征，发现了语言的一些共性，如语言网络中的核心节点在整个网络的构成中发挥着重要作用；超越单个以句子为单位的分析；一个语言节点对整个语言网络都有一定的作用；语言网络不同于随机网络；语言网络的邻接节点发生连接的概率要远大于随机网络等。

4）语言网络有别于其他网络

虽然目前的研究从多种角度揭示了语言网络具有和大部分真实网络一样的特征，但不能忽视语言网络与其他网络的不同之处。如引文网络是一种独特的网络，一种典型的不连通网络，从时间上说，被引文献节点一般只能出现在引文的时间之前。语言是一种有向网络，如果构建的是无向语言网络，这样就忽视了词的先后顺序的问题，而词的先后顺序是影响语言生成机制的重要因素。所以在研究网络共性的同时，不能忽视语言网络的个性，这些个性特征对于语言的识别和区别均是关键问题。

5）语言网络研究是一门跨学科的研究

语言网络研究属于典型的跨学科研究，不是一个学科所能解决的问题，需要将语言学、物理学、数学、信息科学、计算机科学、认知科学等多个学科知识融合起来。目前来自物理学、数学、计算机科学的研究者对该领域进行了关注，语言学领域的研究者还比较少。不同学科的关注视角也有所不同，物理学、数学注重网络机理研究，语言学偏重于从定性的角度进行研究。目前这些领域的研究还是基本处在孤立的状态，未能真正实现多学科的交叉融合。语言网络研究的时期已经来临，迫切需要多学科领域的研究人员进行协作研究，解决目前语言中还难以回答的问题。

6）构建合适的语言网络

究竟采用什么样的方式构建语言网络是合适的，这个问题又回到了语言网络的本质问题上，采用复杂的方式还是采用简便的方式，虽然在依存句法关系中，仅有50%左右的连接属于邻接词，但依存网络就是在真实交流系统中，反映在人脑中的语言网络吗？可以直接用于失语症患者的治疗吗？如果不是真实网络，那么到底有多接近呢？经验语言学还回答不了这些问题，笔者认为，从认知的语言角度，如果能结合真实环境下的人脑所建构的语言复杂网络，应该有更大的应用前景。要想更深入研究人脑中的语言网络，需要认知语言学和心理学领域的结合，而不仅仅是局限于网络整体的研究，而应将更多的研究着眼于局部细节。

7）语言网络研究还有待深入

目前的语言网络研究主要还停留在整体层面，针对语言网络内部结构的深入研究还非常少，目前语言学界等领域的研究者已经意识到该问题，逐渐将目光投向网络内部。

语言网络不是一个泛泛的理论研究，相关研究已经应用于信息检索[46]、机器翻译[47]、词义消歧[48]、自动文摘[49]、关键词提取[50]、情感分析[51]、失语症患者治疗研究[52]等领域。语言网络的研究才刚刚起步，我国学者已经紧随这一潮流，目前在语言网络领域中已经占有一席之地，尤其是在汉语语言网络领域。汉语作为最古老的语言之一，也是目前使用人数最多的语言，其研究不仅可以解决汉语语言中的问题，还对英语、日语等其他语言研究有重要的启发。我们期待在各学科领域的全力协作下，语言网络研究能取得一定进展。

参考文献

[1] Watts D J, Strogatz S H. Collective dynamics of small-world networks[J]. Nature, 1998, 393: 440-442.

[2] Barabasi A L, Albert R. Emergence of scaling in random networks[J]. Science, 1999, 286: 509-512.

[3] 汪小帆，李翔，陈关荣. 复杂网络理论及其应用[M]. 北京：清华大学出版社，2006.

[4] 陈关荣. 复杂网络及其新近研究进展简介[J]. 力学进展，2008, 38(06)：653-662.

[5] Crystal D. The Cambridge Encyclopedia of Language [M]. London: Cambridge University Press, Cambridge, UK, 1997.

[6] George K. Zipf. Human Behaviour and the Principle of Least-Effort [M]. London: Addison-Wesley, Cambridge MA, 1949.

[7] Jayaram B D, Vidya M N. Zipf's Law for Indian Languages[J]. Journal of Quantitative Linguistics, 2008, 15(04): 293-317.

[8] Tuzzi A, Popescu I-I, Altmann G. Zipf's Laws in Italian Texts [J]. Journal of Quantitative Linguistics, 2009, 16(04): 354-367.

[9] 游荣彦. Zipf 定律与汉字字频分布[J]. 中文信息学报，2000, 14(03)：60-65.

[10] Wang D, Li M, Di Z. True reason for Zipf's law in language[J]. Physica A, 2005, 358(02): 545-550.

[11] Cancho R F I, Sole R V. The Small World of Human Language[C]//Proceedings of the Royal Society of London Series B-Biological Sciences, 2001, 268 (1482)：2261-2265.

[12] Dorogovtsev S N, Mendes J F F. Language as an evolving word web[C]//Proceedings of The Royal Society of London. Series B, Biological Sciences, 2001, 268(1485): 2603-2606.

[13] 刘海涛. 语言网络：隐喻，还是利器？[J]. 浙江大学学报(人文社会科学版)，2011, 41(02): 170-180.

[14] Freeman L C. A Set of Measures of Centrality Based on Betweenness[J]. Sociometry, 1979(40): 35-41.

[15] 陈芯莹,刘海涛. 汉语句法网络的中心节点研究[J]. 科学通报，2011,56(10):735-740.

[16] Medeiros Soares M, Corso G, Lucena L. The network of syllables in Portuguese[J]. Physica A,2005, 355(02): 678-684.

[17] Peng G, Minett J W, Wang W S Y. The networks of syllables and characters in Chinese[J]. Journal of Quantitative Linguistics. 2008,15(03): 243-255.

[18] Arbesman S, Strogatz S H, Vitevitch M S. The Structure of Phonological Networks Across Multiple Languages[J]. International Journal of Bifurcation and Chaos,2010,20(03): 679-685.

[19] Yu S, Liu H, Xu C. Statistical properties of Chinese phonemic networks[J]. Physica A,2011, 390(07): 1370-1380.

[20] Choudhury M, Chatterjee D, Mukherjee A. Global topology of word co-occurrence networks: Beyond the two-regime power-law[C]//Association for Computational Linguistics, Beijing,2010,162-170.

[21] 刘知远,孙茂松. 汉语词同现网络的小世界效应和无标度特性[J]. 中文信息学报,2007,21(06): 52-58.

[22] Zhou S, Hu G, Zhang Z, et al. An empirical study of Chinese language networks[J]. Physica A, 2008, 387(12):3039-3047.

[23] Liang W, Shi Y, Tse C K,et al. Comparison of co-occurrence networks of the Chinese and English languages[J]. Physica A, 2009, 388(23): 4901-4909.

[24] Liang W, Tse C K, Huang Q, et. al. Study on the co-occurrence of character networks in Chinese essays from different periods[J]. Science in China Ser. F, 2011,accepted.

[25] Sheng L, Li C. English and Chinese languages as weighted complex networks[J]. Physica A, 2009, 388(12): 2561-2570.

[26] Ke J, Yao Y. Analyzing language development from a network approach[J]. Journal of Quantitative Linguistics, 2008,15(01):70-99.

[27] Cancho R F I, Solé R V, Köhler R. Patterns in Syntactic Dependency Networks[J]. Physical Review E, 2004. 69(05): 051915.

[28] Cancho R F I. The Euclidean distance between syntactically linked words[J], Physical Review E, 2004, 70(05): 056135.

[29] 刘知远,郑亚斌,孙茂松. 汉语依存句法网络的复杂网络性质[J]. 复杂系统与复杂性科学，2008,5(2): 37-45.

[30] Liu H T. Dependency Distance as a Metric of Language Comprehension Difficulty[J]. Journal of Cognitive Science,2008, 9(02):159-191.

[31] Liu H T. The complexity of Chinese syntactic dependency networks[J]. Physica A, 2008, 387(12): 3048-3058.

[32] 刘海涛. 依依存语法的理论与实践[M]. 北京：科学出版社，2009.

[33] 刘海涛. 语言复杂网络的聚类研究[J]. 科学通报，2010, 55: 2667-2674.

[34] Cancho R F I, Capocci A, Caldarelli G. Spectral methods cluster words of the same class in a syntactic dependency network[J]. International Journal of Bifurcation and Chaos, 2007, 17(07):2453-2463.

[35] Čech R, Mačutek J, Žabokrtsky Z. The role of syntax in complex networks: Local and global importance of verbs in a syntactic dependency network[J]. Physica A, 2011, 390(20):3614-3623.

[36] Čech R, Maéutek J. Word form and lemma syntactic dependency networks in Czech: a comparative study [J]. Glottometrics, 2009,19: 85-98.

[37] Sigman M, Cecchi G A. Global organization of the Wordnet lexicon[C]//Proceedings of the National Academy of Sciences of the United States of America, 2002. 99(03): 1742-1747

[38] Motter A E, de Moura A P S, Lai Y C, et al. Topology of the conceptual network of language[J]. Physical Review E,2002, 65(06):065102.

[39] Holanda A J, Pisa I T, Kinouchi O, et al. Thesaurus as a complex network[J]. Physica A , 2004,344(03-04):530-536.

[40] Steyvers M, Tenenbaum J B. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth [J]. Cognitive Science, 2005,29(01):41-78.

[41] Tang L, Zhang Y G, Fu X. The Statistic Properties of Chinese Semantic Network in HowNet[C]//Proceedings of NLP-KE'05, 2005,58-61.

[42] Liu H T. Statistical properties of Chinese semantic networks[J]. Chinese Science Bulletin, 2009, (16): 2781-2785.

[43] Li J Y, Zhou J. Chinese character structure analysis based on complex networks[J]. Physica A,2007, 380 (01):629-638.

[44] Li Y, Wei L, Li Wei, et al. small-world patterns in Chinese phrase networks[J]. Chinese Science Bulletin, 2005, 50(3): 286-288.

[45] 王建伟，荣莉莉. 基于复杂网络理论的中文字字网络的实证研究[J]. 大连海事大学学报，2008, 34(4): 15-18.

[46] Veronis J. Hyperlex: lexical cartography for information retrieval[J]. Computer Speech & Language, 2004:18(03): 223-252.

[47] Amancio D R, Antiqueira L, Pardo T A S, et al. Complex networks analysis of manual and machine translations [J]. International Journal of Modern

Physics C,2008, 19 (04):583-598.

[48] Tsatsaronis G, Varlamis I, Nørvåg K. An experimental study on unsupervised graph-based word sense disambiguation [C]//Proceedings of Computational Linguistics and Intelligent Text Processing, 11th International Conference, CICLing2010, Iasi, Romania, March 21-27, 2010: 184-198.

[49] Antiqueira L, Oliveira Jr O N, Costa, et al. A complex network approach to text summarization[J]. Information Sciences,2009,79(05), 584-599.

[50] 赵鹏,蔡庆生,王清毅,等.一种基于复杂网络特征的中文文档关键词抽取算法[J].模式识别与人工智能.2007, 20(06):827-831.

[51] 余传明,周丹.情感词汇共现网络的复杂网络特性分析[J].情报学报,2010,29(05):906-914.

[52] 江钟立,林枫,孟殿怀.复杂适应性系统理论在言语认知康复中的应用前景[J].中国康复医学杂志,2006, 21(2):183-185.

韩普(1983—),男,博士,讲师,主要研究领域为中文信息处理、信息分析。
E-mail: hanpu0725@163.com

路高飞(1988—),男,硕士研究生,主要研究领域为信息分析,信息检索。
E-mail: lugaofei.cool@163.com

王东波(1981—),男,博士,讲师,主要研究领域为自然语言处理与文本挖掘。
E-mail: wangdongbo0102@gmail.com