汉语词同现网络的小世界效应和无标度特性

刘知远，孙茂松

（清华大学计算机科学与技术系，清华信息科学与技术国家实验室，北京 100084）

摘　要：人类语言的某些重要方面可以通过复杂网络来刻画。本文基于不同规模和类型的语料库，建立了汉语词同现网络，并从复杂网络的角度对这些网络进行了系统的实验考察。实验结果表明汉语词同现网络具有复杂网络的两个基本性质：(1)网络的平均最短路径为 2.63-2.75，聚合系数远大于相同参数下的随机网络，这揭示了汉语同现网络的小世界效应；(2)网络中词的度大体上呈幂律分布，表明汉语同现网络具有无标度特性。本文还对实验中所得到的汉语核心词典进行了定量分析。

关键词：词的同现；复杂网络；小世界；无标度；核心词典

中图分类号：TP391

Chinese Word Co-occurrence Network:

Its Small World Effect and Scale-free Property

Liu Zhi-yuan, Sun Mao-song

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084)

Abstract：Some perspectives of human languages can be characterized by complex network analysis. In this paper, word co-occurrence networks for the Chinese language are automatically constructed based on very large manually word-segmented Chinese corpora with different size and style at first. Then systematic observations on these networks are made from the complex network's point of view. Experimental results show that these networks display two important features of complex networks: (1) The average distance between two words is 2.63-2.75, and the clustering coefficient is much greater than that given by a random network with the same parameters, exhibiting a typical small-world effect; and (2) The degree distributions of these networks generally obey the power-law, i.e., the scale-free property. Related factors of kernel lexicons derived from these experiments are also discussed in a quantitative way.

Keywords: word co-occurrence; complex networks; small world; scale-free; kernel lexicon

1.引言

自然界和人类社会中的大量复杂系统，如生态网、神经网、Internet、社会网络等，越来越成为相关学科的研究热点。复杂网络理论则试图在这些互不相同的复杂网络之中找出它们的共性规律。

20世纪的后40年里，Erdos和Renyi[1]建立的随机图理论(ER模型)一直是人们研究复杂网络结构的基本理论。假设图中有 N 个节点，每对节点以概率 p 连接，则约有 $pN(N-1)/2$ 条边，这样就构成了一个ER模型。虽然人

收稿日期：

基金项目：国家自然科学基金 (项目号: 60573187, 60621062 和 60520130299)。

作者简介：刘知远(1984-)，男，博士生，主要研究方向为自然语言处理。孙茂松(1962-)，男，博士生导师，主要研究方向为信息检索、人工智能和自然语言处理。

们不断对ER模型进行扩展以使其更接近真实网络[2]，但由于大多数实际的复杂网络并不是随机连接的，ER模型作为复杂网络的模型，无疑存在着较大缺陷。

几乎与此同时，人们还开展了对“小世界”效应的实验研究，其中最著名的是六度分离推断[3]。1998年，Watts和Strogatz[4]将小世界模型引入对复杂网络的研究，称为WS模型。稍后Newman和Watts[5]对该模型进行了改进，建立了NW模型。设网络中的节点在一个圆上等距排列，每个节点与左右各3个节点相连，该网络被称为最近邻耦合网络。WS模型在一个最近邻耦合网络中，以较小的概率 p 随机选取节点对添加链接，同时以概率 q 删除部分原有链接。NW模型则仅以概率 p 随机添加链接，但不删除原有链接。这两个小世界模型本质上是一样的，它们都反映了实际复杂网络的一个性质，即大部分节点只与它们的邻近节点相连，同时也有某些节点可与非邻近节点直接相连。

通常从如下两个角度观察小世界效应：平均最短路径长度和聚合系数[2, 4, 6, 7]。

平均最短路径长度是网络中两节点之间的平均距离。具有小世界性质的网络的平均最短路径会很短，远小于网络规模[4]（这也是“小世界”命名的原因）。设平均最短路径为 L，网络中节点个数为 N，网络节点的平均度为 $\langle k\rangle$，对“小世界”网络，则有：

$$L \approx \ln(N)/\ln(\langle k\rangle) \tag{1}$$

随机图模型和小世界模型在这一点上对复杂网络的刻画都比较恰当。

一个节点的聚合系数反映了其相邻节点所构成集合的聚集程度。整个网络的聚合系数 C 是每个节点 i 的聚合系数 C_i 的平均值($0 \le C \le 1$)。对一个有 N 个节点的网络，在极端情况下，当网络所有节点均为孤立节点时，$C=0$；当网络所有节点为全耦合节点时，每个节点与其余 $N-1$ 个节点均有连接，$C=1$。对一个包含 N 个节点的ER随机图网络，当 N 很大时，有：

$$C \approx \langle k\rangle / N \tag{2}$$

即其聚合系数远小于 1。而大规模的实际复杂网络表现出显著的聚合效应[8-10]。显然，随机图模型在这一点上存在不足。WS 和 NW 小世界模型则可较好地反映这种聚合效应，即聚合系数 C 虽然小于 1，但比 $O(N^{-1})$ 要大得多。

对复杂网络进行考量的另一个重要方面是节点的度分布。ER 模型和 WS、NW 模型给出的度分布近似为 Poisson 分布[2]。但大量研究表明，实际复杂网络的度分布明显不同于 Poisson 分布，而更接近于幂律分布（无标度分布）：

$$\Pr(k) \propto k^{-\gamma} \tag{3}$$

其中 $\Pr(k)$ 是度为 k 的节点出现在网络中的概率，γ 为常数。Barabasi 和 Albert [11]认为实际复杂网络具有两个重要性质：(1)增长性，即网络规模不断扩大同时其自身在不断演化；(2)优先连接性，即新的节点更倾向于与那些具有更高连接度的节点连接，表现出“马太效应”。这两个性质导致了复杂网络中节点的度分布服从幂律分布，存在少量度相对很高的节点，但绝大多数节点的度相对很低(即存在所谓的“长尾”)。在此基础上，他们提出了无标度网络模型—BA 模型。在此外，还有一些研究致力于构造能够揭示幂律分布本质的模型[12]。

以上进展引发了众多学科开始以复杂网络理论的角度研究问题。人们发现很多领域都存在着复杂网络现象，如生态网[10]、软件工程[13]、基因工程[14]等。

语言学、心理学、生物学等领域共同研究的对象——语言，是人类文明的结晶，是一个经过漫长演化而来的复杂系统[15]。Sole认为语言在各个层次上都体现了复杂网络的性质，包括语音、词法、句法和语义[16]。国内外已经在词同现网络、词法网络、句法网络和语义网络等方面开展了相应的研究。

Cancho和Sole[17]基于规模为10^7个词次的英语国家语料库(BNC)[1]，构造了一个英语词同现网络，发现该网络表现出小世界效应和无标度特性：(1)平均最短路径小于3；(2) 无论网络的规模如何变化，网络度分布曲线基本呈幂律分布。Cancho和Sole[18]还根据句子中词与词之间的句法结构关系构造了句法网络，发现其也具有复杂网络的基本特征。Motter和Moura等[19]根据概念间的相似性构建了英语的概念网，该网络含有超过3000个概念，也表现出复杂网络的基本特征。

为了深入研究语言网络，Dorogovtsev和Mendes[20]通过优先连接算法建立模型，来模拟人类语言的进化(词由少到多，网络由简单到复杂)，称为DM语言模型。该算法每向网络增加一个新词，进行两个操作：(1)新词按照已有词的度的大小与已有词连接起来，度越大，与该词连接的概率越大；(2)已有词之间按照它们的度的乘积以一定的概率连接。该模型较好地拟合了实际的人类语言结构，同样呈现出小世界网络的性质。进一步地，该模型认为人类语言存在一个核心词典，为该语言的使用者所共用，其规模不随语言的进化而显著变化，约为10^3量级。核心词典在词同现网络中表现为两个斜率不同的度分布线段。其中属于第二段的词，度较高，构成了核心词典；而第一段则为特定领域所使用的词。Cancho和Sole[17, 21]指出，英语的核心词典约含5000词，它们的度与词频以0.66的幂律相关。

针对汉语语言网络，已有一些初步的研究工作。韦洛霞等[22, 23]根据一个基本词语集，构造了汉语词法网络（两个词语如果包含同一个汉字，则认为其间存在连接，如“法治”和“法网”），实验表明该网络的拓扑结构表现出复杂网络的基本性质。唐璐等[24]在两个通行的大型语义网络HowNet和WordNet上，考察了汉语的语义网络，显示其同样具有复杂网络的性质。而关于汉语词同现网络及其性质的定量研究，迄今为止尚未见报道。本文将在大规模语料库的支撑下，定量考察汉语在词同现网络上的小世界效应和无标度特性，并进而得到汉语的核心词典。相关算法刻意地完全采用了前人的经典算法，以便同英语取得的实验结果作相应的比照。

2．汉语词同现网络的构造及相关概念

汉语词法网络的构造主要基于一个静态的基本词语集。而词同现网络的构造则应基于动态的大规模语料库。对汉语而言，这个语料库显然需经过分词处理。

词同现网络的构造算法十分简单[17]：语料库所包含的每一个词型（word type），对应着词同现网络中的一个节点（每一个节点在人脑中可映射为独立的认知实体，这样去考察节点之间的同现关系，才更有意义）。如果在一个句子中，两个词之间在n阶Markov链的条件下存在同现关系，则认为网络中相应的两个节点之间存在一个连接。对语料库中的所有句子进行上述处理，便可构造出词同现网络。

语言工程的实践表明，n阶Markov链中的n取2比较合适，因为句子中两个词的邻接同现是最常见的，如“香港回归”的“香港”和“回归”、“清华大学”的“清华”和“大学”。同时存在大量的间隔1个词的同现，如“在书桌上”的“在”和“上”，“我的家”的“我”和“家”等）。虽然也存在一些间隔大于1的相关词对，但如果在模型中考虑此种远距离关联，则会引入大量的无关词对，降低词同现网络对真实情况反映的准确性。采取这个策略，一方面可较充分地反映词与词之间的上下文制约关系，另一方面，又可使模型的复杂性得到较好的控制。

图1给出了一个根据上述算法由两句话生成的汉语词同现网络的简单示例。

[1] http://www.natcorp.ox.ac.uk/

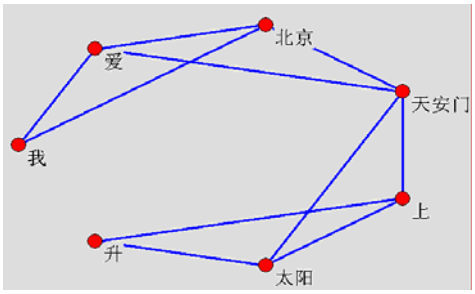

图 1 一个由“我爱北京天安门”和“天安门上太阳升”两句话生成的词同现网络。

一个词同现网络可以抽象为由词集 V 和边集 E 组成的无向图 $G=(V,E)$，其中词数 $N=|V|$，边数 $M=|E|$。网络中两个词 i 和 j 的距离 d_{ij} 定义为连接这两个词的最短路径长度。词 i 的度 $<k_i>$ 定义为与该词连接的其他词的数目。网络的平均最短路径 d 定义为任意两个词之间距离的平均值：

$$d=\frac{1}{N(N-1)}\sum_{i,j\in N,i\neq j}d_{ij} \tag{4}$$

如果词 i 与 j 之间存在连接，则设 $\xi_{ij}=1$。假设词 i 有 k_i 条边与其他词相连，那么这 k_i 个词定义为 i 的最近邻，其集合为 $\Gamma_i=\{j\,|\,\xi_{ij}=1\}$。词 i 的最近邻集合中词间的连接数为：

$$L_i=\sum_{j=1}^{N}\xi_{ij}\left[\sum_{k\in\Gamma_i,j<k}\xi_{jk}\right] \tag{5}$$

而在这 k_i 个词之间至多有 $k_i(k_i-1)/2$ 条边。所以词 i 的聚合系数 c_i 定义为：

$$c_i=\frac{L_i}{k_i(k_i-1)/2} \tag{6}$$

则该网络的聚合系数 C 为：

$$C=\frac{1}{N}\sum_{i=1}^{N}c_i \tag{7}$$

3．实验及其分析

本文实验利用了北京大学《人民日报（1998 年上半年）》1300 万字左右的人工分词语料库[2]和国家语委 5000 万字左右的人工分词语料库[3]（它们也是目前世界上规模最大、质量最高的汉语分词语料库）。前者可以按月份分割，用于在不同规模的汉语词同现网络上考察复杂网络性质及其平稳性；后者包含了各种题材、各个领域的文本，是较好的平衡语料库，可以更全面地考察汉语词同现网络的复杂网络性质。

本文设计了 4 组实验，用来生成词同现网络的语料库分别取北京大学《人民日报(1998 年上半年)》分词语料库的 1~2 月份、1~4 月份、1~6 月份和国家语委分词语料库，记作 CPD12，CPD14，CPD16 和 CYW。实验采用复杂网络分析软件 Pajek[4]进行数据分析。

实验结果见表格 1。其中 ENG 是英语词同现网络上的实验数据[17]，它采用了与本文相同的实验方法，在这里作为对照。ENG 实验从 750 万词的语料库得到含 460,902 个节点的词同现网络。而汉语 CPD16 实验从 730

[2] http://icl.pku.edu.cn/
[3] http://219.238.40.213:8080/
[4] http://vlado.fmf.uni-lj.si/pub/networks/pajek/

万词语料库最终只得到 124,334 个节点的词同现网络。从相同规模的语料库得到的英语词同现网络节点数明显多于汉语的词同现网络，主要原因是英语是一种屈折语言，其名词、动词等有各种屈折变化[25]，造成了网络中节点数目的激增。而汉语是一种孤立语言，缺少严格意义上的形态变化[26, 27]。

表格 1 中 C_{random} 和 d_{random} 分别是相同参数下(N 和 $\langle k \rangle$ 相同)的随机网络中的聚合系数与平均最短路径。可以看到 $C >> C_{random}$，而 $d \approx d_{random}$。汉语词同现网络与英语词同现网络一样，平均最短路径远小于网络规模而聚合系数非常高，具有明显的小世界效应。这说明，虽然大量的词(几万甚至以十万计)储存在人脑中，但是人们在这个词网络中，可以只用很短的路径就能从一个词到达另一个词。也就是说在交流中，当使用了某个词，可以通过有限步很快地跳转到另外一个词。这样，语言网络一方面很好地保证人们在交流时的速度，另一方面能够从规模上保证人们在交流时用词的丰富性[17]。

从 CPD12，CPD14 和 CPD16 可以明显看出平均度 $\langle k \rangle$ 随着语料库规模增大。这可以被看作某种语言进化的过程：随着时间的推移和社会的发展，新产生的词被加入到语言中或者原来较少使用的词逐渐被关注，从而为人们所习用，例如从 CPD12 到 CPD14 新增加的词有“楼兰”、“报关员”、“彩色棉”等，从 CPD14 到 CPD16 新增的词有“帕格尼尼”、“公务车”、“核竞赛”等。伴随着这个过程，原有词的连接也会增加，如“市场”的度在各语料库中的变化为 1803(CPD12) -> 2842(CPD14) -> 3607(CPD16)，从一个侧面反映了 Barabasi 和 Albert 所阐述的复杂网络的增长性[11]。

表格 1 词同现网络的基本数据。其中 ENG 是英语词同现网络的实验数据，引自[17]。

实验	N	E	$\langle k \rangle$	C	C_{random}	d	d_{random}
ENG	4.61×10^5	1.61×10^7	70.13	0.437	1.55×10^{-4}	2.67	3.06
CPD12	0.71×10^5	0.10×10^7	28.44	0.535	3.99×10^{-4}	2.75	3.34
CPD14	1.03×10^5	0.18×10^7	34.18	0.556	3.32×10^{-4}	2.73	3.27
CPD16	1.24×10^5	0.24×10^7	38.99	0.569	3.14×10^{-4}	2.71	3.20
CYW	1.57×10^5	0.83×10^7	64.35	0.619	2.50×10^{-4}	2.63	2.99

图 2 列出了最短路径的分布。语料库中有部分孤立词与其他词没有连接（当这些词组成“独词句”时），会造成不可达词对，因此图中所列的节点对比例的和小于 1。平均最短路径的分布比较有规律，$d=2$ 和 $d=3$ 的节点对比例占了绝大多数，CPD12、CPD14 和 CPD16 中都超过 80%，CYW 中也占 78.5%之高。除不可达词对外，存在连接的词对的最短路径都比较小。路径中两个词的距离越短，说明它们之间的跳转越直接，也越容易，在人们交流的过程中越比较经常地一起使用，如路径“缉拿-凶犯”、“主隧-全长-公里”及“凶手-缉拿-凶犯”（这里“凶手”通过“缉拿”与“凶犯”产生同义关联）；反之，联系越松散，如路径“联系簿-警民-关系-群众”中的“联系簿”与“群众”。

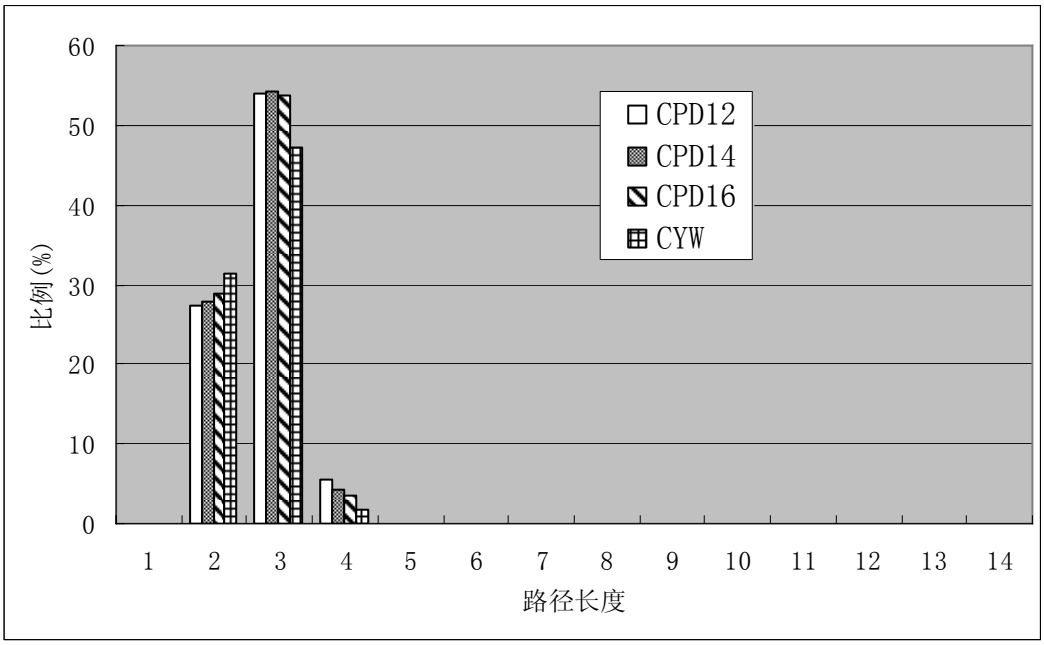

图 2：汉语词同现网络最短路径分布，即路径为 d 的数目及其比例。设词同现网络节点数为 N，路径长度为 d 的数目为 m_d，则比例为 $p_d = 2m_d/(N^2 - N)$。

网络节点的累积度分布曲线见图3。累计度分布是度不小于k的节点的分布概率：

$$P(k) = \sum_{j=k}^{\infty} \Pr(j) \tag{8}$$

当度分布曲线呈幂律分布时，其累积度分布曲线也呈指数值相差1的幂律分布[2]。根据式(3)可得：

$$P(k) = \sum_{j=k}^{\infty} j^{-\gamma} \propto k^{-(\gamma-1)} \tag{9}$$

可以看到四组实验结果都大体呈幂律分布，显示了无标度特性。

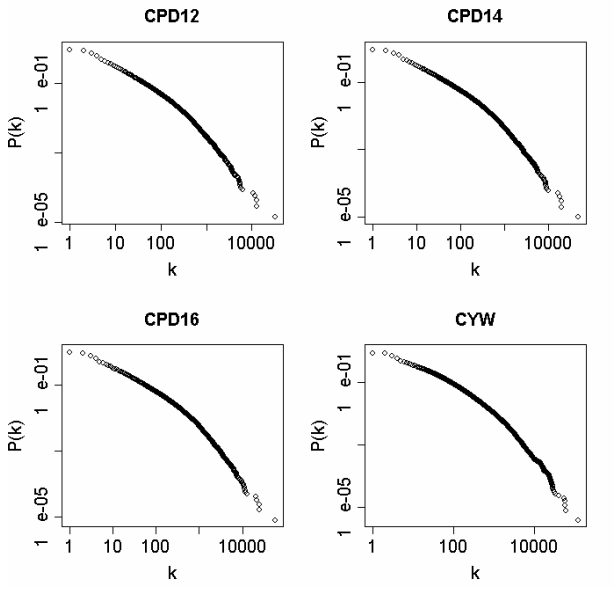

图 3　CPD12，CPD14，CPD16 和 CYW 的累积度分布曲线(log-log)。

如果对这些曲线进行更为细致的观察，则会发现其度分布并非一条直线，而是可以划分为两个斜率明显不

同的线段（英语词同现网络也存在类似的现象）。图 4 显示了对 CYW 累积度分布进行线性拟合的情况：以 $(k, P(k)) = (802, 0.0133511)$ 处为转折点，第一段斜率为 –0.51，第二段斜率为 –2.51，并根据式(9)可得第一段指数 $\gamma_1 = 1.51$，第二段指数 $\gamma_2 = 3.51$；在 CYW 生成的词同现网络中，度大于 802 的词数目为 $P(k) \times N \approx 3434$。其他三组实验的度分布也明显分为两个不同斜率的线段，实验数据见表格 2。

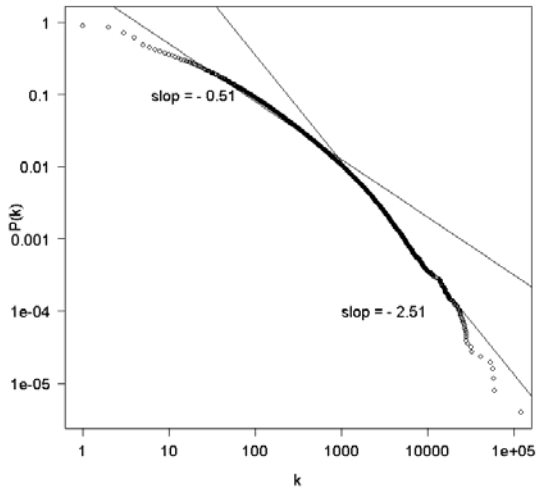

图 4 对 CYW 词同现网络累积度分布 (log10-log10) 的拟合曲线。

表格 2 四组实验的累积度分布曲线通过两个不同线段拟合的结果。其中 k 为转折点的横坐标。$N \times P(k)$ 为度大于 k 的词数。

实验	k	$P(k)$	γ_1	γ_2	$N \times P(k)$
CPD12	177	0.028	0.266	1.921	951
CPD14	240	0.025	0.267	2.066	1553
CPD16	318	0.022	0.301	2.214	1873
CYW	802	0.013	0.511	2.511	3434

心理学实验表明，一个词在交流中出现频度越高，其语言产生的能力越强[28, 29]，即人脑能够更容易地使用这个词表达思想。图 5 显示 CYW 组实验中词频 f 与其度 k 之间存在相当强的相关性，即 $f \propto k^{\gamma}$ ($\gamma > 0$)。这表明一个词的度越高，一般来说，其语言产生的能力也就越强。

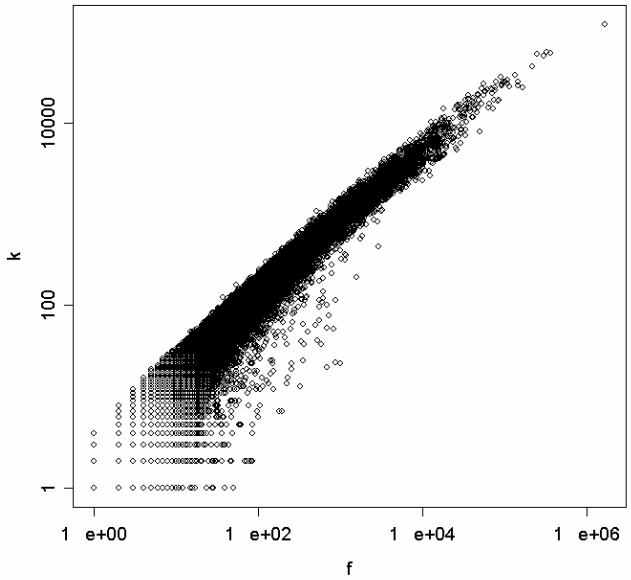

图 5 CYW 中词频 f 与其度 k 的相关性分析。

由表格 2 还可见，$N \times P(k)$ 即核心词典的规模约为 10^3 量级，基本符合 DM 语言模型对核心词典规模的推论（注意：英语的核心词典约含 5000 词[17, 21]）。

表格 3 给出了四组实验所得的核心词典(KLi, i 对应 CPD12，CPD14，CPD16 和 CYW)与从相应语料库产生的规模相同的词频表（FLi）、核心词典相互之间以及与人工建立的《普通话三千常用词表》(PTL)[30]的比较结果。表中数字为相同词条的数目。不同语料库下的核心词典与相同规模的词频表的比较，符合前文词频 f 越高则度 k 越趋高的结论。CPD12，CPD14 和 CPD16 核心词典之间的比较表明绝大部分词条是相同的，说明核心词典具有一定的稳定性；由于 CPD12、CPD14、CPD16 与 CYW 语料库来源不同，它们之间的核心词典存在较大差异。各核心词典与《普通话三千常用词表》进行比较，大部分词条出现在该表中。而它们之间存在一定差别的主要原因是：(1)《普通话三千常用词表》是人工整理的词表，以人的主观感觉为主要判断依据，与词频的定量分析有一定出入。表格 4 显示了四组实验中核心词典、《普通话三千常用词表》及两者的交集对语料库的词次(word token)覆盖率。此外，CPD16 核心词典对 CYW 语料库的词次覆盖率为 61.7306%，CYW 核心词典对 CPD16 语料库的词次覆盖率为 71.5804%，两核心词典交集对 CPD16 语料库的词次覆盖率为 66.2198%，对 CYW 语料库的词次覆盖率为 61.2254%，各核心词典对语料库的覆盖率明显高于《普通话常用三千词表》。在这一点上，核心词典显示了其定量分析的长处。(2)CYW 和 CPD16 的核心词典依赖于该语料库的来源、规模和分词标准等因素，因此只能是一定意义下的“汉语核心词典”。

表格 3 四组实验所得的核心词典(KLi)与从相应语料库产生的规模相同的词频表（FLi）、核心词典相互之间以及与《普通话三千常用词表》的比较。表中数字为相同词条的数目。

比较对象	KL:CPD12	KL:CPD14	KL:CPD16	KL:CYW
	951	1553	1873	3434
FLi	851	1364	1658	3151
KL:CPD12	-	951	951	906
KL:CPD14	951	-	1151	1388
KL:CPD16	951	1151	-	1597
KL:CYW	906	1388	1597	-
PTL	610	892	1008	1719

表格 4 四组实验中核心词典(KLi)、相同规模词频表(KLi)、《普通话三千常用词表》(PTL)及其交集对语料库的词次覆盖率。

覆盖率	CPD12(%)	CPD14(%)	CPD16(%)	CYW(%)
KLi	60.02	67.87	70.49	75.38
FLi	60.93	68.73	71.30	75.73
PTL	56.71	56.89	56.95	62.01
KLi∩PTL	48.06	51.73	52.73	59.46

在核心词典中，“的”、“和”、“在”、“了”、“是”、“为”、“有”、“这”、“他”、“我”和“人”等词的度最高。这些词或者是虚词，用以粘着成句，或者是具有强烈语法作用的实词。。它们中的相当一部份对于句子的理解似乎没有太大的直接贡献，而一旦缺失这些词，句子将变得支离破碎。这也反映了小世界网络的一个特性：如果随机地去掉网络中的节点，该网络仍然可以保持较好的连接性，而如果一旦去除的是高连接度的节点，整个网络将破裂成为若干孤立的网[31]。某些失语症患者就表现为功能词缺失、不能正确组合语句、语句不完整、缺少长句和复杂句[32]。

4．结论

本文基于大规模语料库，通过实验揭示了汉语在词同现网络上的小世界效应和无标度特性，并对汉语核心词典进行了初步研究。英语与汉语虽然存在显著差异（前者为印欧语系，后者为汉藏语系），但在词同现网络上表现出了类似的复杂网络性质。这一方面验证了 DM 语言模型对汉语的有效性，另一方面也从一个侧面印证了复杂网络的普适性。

参考文献：

[1] Erdös P, Renyi A. On the Evolution of Random Graphs[J]. Publ Math Inst Acad Sci, 1960. 5: 17-61
[2] Newman M E J. The Structure and Function of Complex Networks[J]. Siam Review, 2003. 45(2): 167-256
[3] Milgram S. The Small-World Problem[J]. Psychology Today, 1967. 2: 60-67
[4] Watts D J and Strogatz S H. Collective Dynamics of 'Small-World' Networks[J]. Nature, 1998. 393(6684): 440-442
[5] Newman M E J, Watts D J. Renormalization Group Analysis of the Small-World Network Model[J]. Physics Letters A, 1999. 263(4-6): 341-346
[6] Dorogovtsev S N, Mendes J F F. Evolution of Networks[J]. Advances in Physics, 2002. 51(4): 1079-1187
[7] Albert R, Barabasi A L. Statistical Mechanics of Complex Networks[J]. Reviews of Modern Physics, 2002. 74(1): 47-97
[8] Strogatz S H. Exploring Complex Networks[J]. Nature, 2001. 410(6825): 268-276
[9] Jeong H, Tombor B, Albert R, et al. The Large-Scale Organization of Metabolic Networks[J]. Nature, 2000. 407(6804): 651-654
[10] Montoya J M, Sole R V. Small World Patterns in Food Webs[J]. Journal of Theoretical Biology, 2002. 214(3): 405-412
[11] Barabasi A L, Albert R. Emergence of Scaling in Random Networks[J]. Science, 1999. 286(5439): 509-512
[12] 胡海波, 王林. 幂律分布研究简史[J]. 物理, 2005. 34(12): 889-896
[13] Valverde S, Cancho R F, Sole R V. Scale-Free Networks from Optimal Design[J]. Europhys Lett, 2002. 60(4): 512-517
[14] Sole R V, Pastor-Satorras R. Complex Networks in Genomics and Proteomics[A]. Bornholdt S, Schuster H G, eds. In: Handbook of Graphs and Networks[C]. Berlin Wiley-VCH, 2003. 145-167
[15] Steels L. Language as a Complex Adaptive System[A]. Schoenauer M, ed. In: Proceedings of ppsn-vi, Lecture Notes in Computer Science[C]. Berlin: Springer-Verlag, 2000. 17-26
[16] Sole R V, Murtra B C, Valverde S, et al. Language Networks: Their Structure, Function and Evolution[J]. Trends in Cognitive Sciences, 2006
[17] Cancho R F I, Sole R V. The Small World of Human Language[J]. Proceedings of the Royal Society of London Series B-Biological Sciences, 2001. 268(1482): 2261-2265
[18] Cancho R F I, Sole R V, Kohler R. Patterns in Syntactic Dependency Networks[J]. Phys Rev E, 2004. 69(5): 051915
[19] Motter A E, de Moura A P S, Lai Y C, et al. Topology of the Conceptual Network of Language[J]. Phys Rev E, 2002. 65(6): 065102
[20] Dorogovtsev S N, Mendes J F F. Language as an Evolving Word Web[J]. Proceedings of the Royal Society of London Series B-Biological Sciences, 2001. 268(1485): 2603-2606
[21] Cancho R F I, Sole R V. Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited[J]. Journal of Quantitative Linguistics, 2001. 8(3): 165-173
[22] 韦洛霞, 李勇, 康世勇, 等. 汉语词组网的组织结构与无标度特性[J]. 科学通报, 2005. 50(15): 1575-1579
[23] 韦洛霞, 李勇, 李伟, 等. 汉字网络的 3 度分隔与小世界效应[J]. 科学通报, 2004. 49(24): 2615-2616
[24] 唐璐, 张永光, 付雪. 语义网络的结构:我们怎样学习语义知识[J]. 东南大学学报（英文版）, 2006. 22(3): 413-417
[25] Matthews P H. Morphology[M]. Cambridge: Cambridge University Press 1991
[26] 吕叔湘. 汉语语法分析问题[M]. 北京: 商务印书馆. 1979
[27] 朱德熙. 语法讲义[M]. 北京: 商务印书馆. 1982
[28] Griffin Z M, Bock K. Constraint, Word Frequency, and the Relationship between Lexical Processing Levels in Spoken Word Production[J]. Journal of Memory and Language, 1998. 38(3): 313-338
[29] 张清芳, 杨玉芳. 言语产生中的词汇通达理论[J]. 心理科学进展, 2003. 11(1): 6-11

[30] 郑林曦. 普通话三千常用词表[M]. 1987, 文字改革出版社: 北京
[31] Albert R, Jeong H, Barabasi A L. Error and Attack Tolerance of Complex Networks[J]. Nature, 2000. 406(6794): 378-382
[32] 施琪嘉, 高素荣, 刘锡民, 等. 说汉语的失语患者言语表达的初步分析[J]. 中国康复医学杂志, 2004. 19(1): 11-14