科学通报 2011 年 第 56 卷 第 10 期：735 ~ 740
论 文 www.scichina.com csb.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

汉语句法网络的中心节点研究

陈芯莹[①②], 刘海涛[①*]

① 浙江大学外国语言文化与国际交流学院, 杭州 310058;
② 中国传媒大学应用语言学研究所, 北京 100024
* 联系人, E-mail: lhtzju@gmail.com

2010-12-16 收稿, 2011-01-19 接受
国家社会科学基金资助项目(09BYY024)

摘要 以两种语体的汉语依存句法树库为基础, 根据词频及分布率统计结果, 选取 3 个汉语虚词作为研究对象. 对提取的 3 个虚词节点进行了节点度数、点出度、点入度、接近性、内接近性、外接近性、中间度等网络特征的统计, 并将这 3 个节点从网络中移除, 对比分析网络前后的节点数、平均度、平均路径长度、网络直径、孤立节点数、最大范围、密度等网络特征的变化. 结果表明, 3 个虚词均是网络的中心节点, 但地位各有不同, 它们对网络整体结构的影响也有较大区别. 本研究不仅为汉语虚词的研究提供了新方法, 也为复杂网络中的节点特性研究提供了新的思路.

关键词
复杂网络
中心节点
语言网络
虚词

语言系统是一种复杂的网络结构体[1], 因此采用复杂网络来研究语言是很有必要的一项尝试[2]. 国内外有关语言复杂网络研究的成果不少[2~10]. 尽管语言网络的构造原则各有千秋, 但大部分研究都偏重于对各种网络共性的探讨, 例如小世界和无尺度特征等. 这种偏重共性的研究, 会给人“天下网络一般黑”[11]的感觉, 这使得语言结构的差异也被淹没在语言网络的整体特征当中. 为了从语言网络中挖掘出语言结构的个性, 需要进一步深入网络结构的内部, 更多地去考察分析网络中局部结构以及节点的情况.

考察网络局部结构特征的首选切入点是网络的中心节点, 因为中心节点的存在是网络表现出小世界和无尺度特征的重要原因. 就语言网络而言, 什么是它的中心节点? 这些节点在语言网络中又起着什么样的作用呢?

人类目前正在使用的语言有 6800 多种[12]. 尽管语言的类型不同, 构建语言网络的方法也有多种选择[2~10]. 但对于语言学家而言, 语言网络只是研究语言的手段, 而非目标[8]. 用网络的方法去发现和解释语言结构与现象, 才是语言学家的真正目的. 所以, 要构建有语言理论依据的句法和语义网络. 由于不同语言在句法层面表现出的区别特征相较于语义层面更明显一些, 所以构建句法网络可能是更好的选择. 对于汉语句法网络来说, 虚词很可能就是网络的中心节点, 是我们的研究对象. 这是因为汉语是孤立语, 实词缺乏表示语法意义的形态变化, 虚词(和语序)便成了表达语法功能的主要手段, 显得尤为重要[13].

本研究不同于以往关注语言结构共性与整体特性的研究, 将目光投向了网络结构的内部, 关注网络的中心节点, 试图从宏观与微观相结合的角度挖掘更多的网络结构个性. 据我们所知, 国内外现有的语言网络研究当中还没有对中心节点的相关研究. 从语言学的角度来说, 本研究为汉语虚词的研究提供了一种新的方法. 从复杂网络的角度看, 由于语言网络的节点具有定义明确、特性便于描述和量化的特点, 所以汉语句法网络的中心节点研究可以从理论上较好地解释复杂网络结构的特点及构成. 这对于复杂网络节点特性的研究及网络结构动力的研究都有一定意义.

英文引用格式: Chen X Y, Liu H T. Central nodes of the Chinese syntactic networks (in Chinese). Chinese Sci Bull (Chinese Ver), 2011, 56: 735–740, doi: 10.1360/972010-2369

1 资源与方法

为了减少语体对研究结果的影响, 我们选用“实话实说”(以下简称 SHSS)和“新闻联播”(以下简称 XWLB)两类语料作为研究资源. 我们先通过词频及分布率的统计比较, 确定了具体的 3 个虚词作为研究对象. 然后考察这 3 个虚词节点的网络特征值, 分析其在网络结构中的中心节点地位. 然后分别将 3 个节点从网络中移除, 对比分析网络前后的特征值变化. 结果表明, 3 个虚词均是网络的中心节点, 但地位各有不同, 它们对网络整体结构的影响也有较大区别.

我们首先关注的是虚词出现的频率. 一般认为, 词的频度统计是计量语言学的基础[14]. 但频率标准是有局限性的. 因为频率统计的准确程度与所选取的语言材料的容量有密切关系, 因此频率的准确程度具有相对性[15]. 由于本研究所使用的两个语料库均不到 2 万词, 为保证研究结果的准确性, 我们仅将位列前 50 位的虚词作为研究对象. 统计 SHSS 和 XWLB 中出现频率最高的 50 个词并整理其中所有虚词的数据可以得到表 1.

统计结果显示, SHSS 中, 词频处于前 50 的虚词有 5 个, 包括 3 个助词和 2 个介词; XWLB 中, 词频处于前 50 的虚词有 9 个, 包括 3 个助词、2 个连词和 4 个介词.

表 1 虚词词频列表 a)

XWLB			
Rank1	Rank2	Freq	Word
1	1	930	的 u
2	2	273	和 c
3	3	223	在 p
4	4	202	了 u
5	11	81	对 p
6	15	64	等 u
7	26	48	从 p
8	30	45	为 p
9	35	43	并 c
SHSS			
Rank1	Rank2	Freq	Word
1	1	1051	的 u
2	6	429	了 u
3	21	124	在 p
4	43	73	着 u
5	48	66	把 p

a) Rank1 为虚词词频序号, Rank2 为词频序号

时代的不同, 地域的不同, 语言材料容量的不同, 语言材料是书面语言还是口头语言, 这些材料都会影响到词频, 所以我们不能只以频率标准作为选择词汇的唯一标准. 一个词在一定篇数的语言材料的样本中出现在多少篇数中, 也是衡量该词重要与否的标准. 这个标准, 叫做分布率标准[15]. 根据统计结果, 频率在前 50 位的虚词在两个树库中均有分布的有 3 个: 的、了、在 p(作介词). 而根据《现代汉语频率词典》的显示, “的”、“了”、“在”分列词频的第一、二和六位, 是词频最高的 3 个虚词[16]. 因此, 我们将研究对象定为“的”、“了”、“在 p”(以下简称为 A, B, C)3 个虚词节点.

为了观察和分析这 3 个节点的网络特性, 我们在 SHSS 和 XWLB 两个依存树库的基础上构建了两个汉语依存句法网络并分别统计了 3 个节点的节点度数(all degree)、点出度(out-degree)、点入度(in-degree)、接近性(all closeness)、内接近性(in-closeness)、外接近性(out-closeness)、中间度(betweenness). 网络构建方法参见文献[8], 这里不再赘述. 为了消除统计误差并方便对比, 我们将测量结果的最大值设为 1, 对数据进行标准化处理.

度数指与一个点直接相连的其他点的个数. 一个点的度数就是对其“领域”规模大小的一种数值测度. 如果某点度数高, 则称该点居于中心. 但由于度数的测量仅仅根据与该点直接相连的点数, 忽略间接相连的点数, 因此, 测量出来的度可以称为“局部中心度”.

一个点的点入度指直接指向该点的点数总和; 点出度指该点所直接指向的其他点的总数.

如果一个点与其他许多点的距离都很短, 这样的点与网络中许多其他点都“接近”. 一个点的接近性与它到其他各点的距离之和(即距离和)是反向的. 一个点与其他点的距离和大, 其接近性就小.

在一个有向网络中, 根据出入不同方向计算出来的“接近性”也有所不同. 这样, 可以分别测算出一个有向网络的“内接近性”和“外接近性”.

中间度概念测量的是一个点在多大程度上位于网络中其他点“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. 一个点的中间度测量的是该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他人.

表 2 为用网络分析软件 PAJEK[17]统计出的数据

表 2　节点 A, B, C 的网络参数

网络特征	A		B		C	
	XWLB	SHSS	XWLB	SHSS	XWLB	SHSS
节点度数	964	830	133	234	222	131
标准化节点度数	1	1	0.13797	0.28193	0.23029	0.15783
点入度	504	405	0	0	88	61
标准化点入度	1	0.81984	0	0	0.17460	0.12348
点出度	460	425	133	234	134	70
标准化点出度	1	1	0.28913	0.55059	0.29130	0.16471
接近性	0.50188	0.55770	0.35197	0.43941	0.40977	0.44158
标准化接近性	1	1	0.70130	0.78790	0.81647	0.79178
内接近性	0.37375	0.39885	0	0	0.26484	0.28302
标准化内接进行	0.91600	0.84731	0	0	0.64907	0.60124
外接近性	0.21871	0.27586	0.15682	0.22887	0.18254	0.21486
标准化外接近性	1	1	0.71705	0.82963	0.83465	0.77887
中间度	0.32098	0.27229	0	0	0.02750	0.01365
标准化中间度	1	1	0	0	0.08569	0.05012

结果.

为了考察中心节点(包括整体中心节点与局部中心节点)在整个网络中所起的作用, 我们分别将 A, B, C 这 3 个节点从网络中剔除, 并统计原始的 XWLB, SHSS 网络和 3 个去节点网络的节点数(number of vertices)、平均度(average degree)、平均路径长度(average path length)、网络直径(diameter)、孤立节点数(number of isolated vertices)、最大范围(domain)、密度(density)等几个网络特征数据, 观察节点删除前后的变化(表 3).

平均度指某具体网络中, 每个节点平均具有的节点度数. 网络所有节点度数之和与节点数之比即是平均度.

平均路径长度指网络中任意两点间的平均最短路径:

$$\langle d\rangle = \frac{1}{\frac{1}{2}N(N+1)}\sum_{i\geq j} d_{ij}, \tag{1}$$

式中 N 代表网络节点数, d_{ij} 指节点 i 与节点 j 的距离, 可以用两点间最短路径所包含的边数来表示.

网络直径指网络中最长路径值. 例如, 在SHSS的依存句法网络中, 最长的路径是节点 101 与节点 708 的路径, 这两点间的距离为 9, 是网络中存在的最长的距离.

孤立节点数指度数为 0 的节点数量.

范围指某节点通过链接可达到的节点数目. 最大范围指所有节点范围中的最大值.

密度概念描述了一个图中各个节点之间关联的紧密程度:

$$\frac{2L}{n\times n}, \tag{2}$$

其中 L 代表网络中实际存在的边数, n 表示网络中的节点数.

2　结果与讨论

参考表 2 和 3 中平均度的数据可知, A 无论是在 XWLB 还是 SHSS 网络中, 度数、点出度、接近性、外接近性和中间度都是 1, 即所有节点的最高值. 它的高度数、高入度、高出度特性, 说明它在网络中是一个局部中心节点. 接近性为 1, 说明它与其余各点的距离和最小. 具有最高的中间度, 这说明它的整体中心度最高, 是整个网络的最“中心”节点.

B 在 XWLB 和 SHSS 中均具有高度数特点, 说明它是一个局部中心节点. 中间度为 0, 所以整体中心度低, 并不是网络的整体中心节点, 仅能作为一个局部中心节点存在. B 节点的一个显著特点是, 无论在 XWLB 还是在 SHSS 网络中, 它的点入度、内接近性均为 0.

表 3 网络数据对比 a)

网络		节点数	平均度	平均路径长度	直径	孤立节点数	最大范围	密度
XWLB	完整	4011	6.15	3.58	12	0	4010	0.00153
	去 A	4010	5.67	3.93	12	42	3928	0.00141
	去 B	4010	6.09	4.56	20	0	4009	0.00076
	去 C	4010	6.04	4.59	20	17	3990	0.00075
SHSS	完整	2601	8.56	3.05	9	0	2600	0.00327
	去 A	2600	7.92	3.25	10	57	2521	0.00303
	去 B	2600	8.38	3.95	13	0	2599	0.00161
	去 C	2600	8.46	3.96	13	5	2590	0.00163

a) 去掉 A, C 之后网络中存在孤立节点, 所统计的平均路径长度、直径均指非孤立节点的值

C 同样具有高度数、高点入度、高点出度的特点, 是一个局部中心节点. 整体说来, 它比 B 更加靠近网络的整体中心.

韦洛霞等人[7]曾提出, 按字频选择汉字是导致词组网幂律度分布的重要原因, 且汉语词组网的组织结构服从自然界普遍存在的最小省力原则. 而 Liu[8]则发现, 汉语句法网络中, 词的度分布也是符合幂律的, 汉语的句法结构同样符合最小省力原则, 但词的度与词频并不是一致对应的, 不是词频高词的度数就一定高. 对比表 1 和 2 的数据, 可发现词频与节点的网络中心地位也不是一一对应的, 并非词频越高在网络结构中的中心地位就越高. 究竟什么原因造成这一现象, 值得深入研究探讨. 但就现在的数据和了解来推测, 这很有可能与节点的接近性和不可替代性有关. 虽然 B 只是一个局部中心节点, 但它的外接近性大, 它与邻居节点间的距离较短. 另外, 去掉 B 之后, 网络的密度降低了, 平均路径长度及直径均增加了, 而且这些变化都比较明显. 这说明, B 在缩短某些部分节点间的距离上有着不可替代的功用. 这也是为什么虽然 B 的整体中心度低于 C, 但出于最小省力原则, 它的频率仍比 C 高.

表 3 显示, 无论是 XWLB 还是 SHSS 网络, 在去掉 A 后, 网络的平均度、最大范围、密度均降低了, 而平均路径长度及孤立节点数则增加了. XWLB 网络在去 A 后直径没变, SHSS 网络在去 A 后直径变大.

平均度降低是因为 A 的度数远大于原始网络的平均度, 去 A 后自然会降低网络的平均度.

去 A 之后, 那些仅能依靠 A 而进入句法网络的节点就变成了孤立节点. A 具有最高的中间度, 是网络中最重要的“中间人”. 因此, 去 A 之后有相当一部分节点被孤立了, 这个现象在口语体的 SHSS 网络中更加明显, 有超过 2%的节点由于 A 的缺失而被孤立.

除了部分节点会被完全孤立外, 还有些节点会三两成对地游离在大多数节点组成的网络之外. 因此, 我们需要统计最大范围. 根据表 3, XWLB 去 A 后的最大范围为 3928, 这就是说在去 A 后的网络中, 最大的一个子网络是由连通的 3929 个节点组成的, 有 82 个节点游离在这个大成分之外, 无法链接到大部分节点, 其中的 42 个节点更是处于完全孤立状态, 无法与任何其他节点链接. SHSS 去 A 后的最大范围为 2521, 有 79 个节点游离在这个大成分之外, 无法链接到大部分的节点, 其中的 57 个节点成了孤立节点.

去 A 后, 两个网络的平均路径长度和密度均降低了, 但由(2)式可知, 网络节点数是影响这两个参数的要素, 因此, 我们无法判断 A 本身对数据的影响程度.

SHSS 网络在去 A 后直径变大, 说明通过 A 节点能够缩短网络中部分节点间的距离. 而 XWLB 网络在去 A 后直径并没变, 我们认为这可能是由于语体的不同而造成的.

无论是 XWLB 还是 SHSS 网络, 在去掉 B 后, 网络的平均度、最大范围、密度都有所降低, 平均路径长度及直径则均有所增加. 孤立节点数依然为 0 没有变化.

平均度降低是因为 B 的度数远大于原始网络的平均度, 去 B 后自然会降低网络的平均度.

B 的中间度为 0, 并不是网络的中心节点. 在去掉 B 之后, 没有节点因此而被孤立, 也没有小的节点对游离在大网络之外, 整个网络中的所有节点仍然是连通的.

虽然我们仍然无法判断 B 本身对平均路径长度和密度数据的影响有多大, 但可将它与 A 做一对比. 通过对比可以看出, B 对平均路径长度和密度的影响都比 A 大. 特别是密度, 去掉 B 后的网络密度仅为原始网络密度的一半左右. 由于密度概念描述了一个网络中各节点之间关联的紧密程度, 因此我们认为 B 具有使网络中部分节点联系得更加紧密的能力, 且 B 的这种能力中不可替代的部分较 A 要多. A 作为局部和整体中心节点一定也具备这种能力, 去 A 后密度变化较小可能是因为 A 的这种能力会部分地被一些其他节点所替代. 例如, A 能够使 *x*, *y*, *z* 节点联系更加紧密, 但可能同时有其他节点也能使这些节点联系更加紧密, 所以在去掉 A 后, 节点间紧密程度并没有变化.

去 B 后直径变大, 这说明通过 B 节点能够缩短网络中部分节点间的距离. 与 A 相比, 去 B 后直径的增加要更加明显. 我们认为, 这也可能是 B 缩短节点间距离的能力中不可替代的部分较 A 要多所致.

无论是 XWLB 还是 SHSS 网络, 在去掉 C 后, 网络的平均度、最大范围、密度均降低, 平均路径长度、直径、孤立节点数则增加了.

平均度降低是因为 C 的度数远大于原始网络的平均度, 去 C 后自然会降低网络的平均度.

C 具有较高的中间度, 但 C 受语体影响比较大, 它在 XWLB 网络中更加接近整体中心. 去 C 之后有一部分节点被孤立了, 这个现象在类书面语体的 XWLB 网络中更加明显. 因此, 我们认为 C 在 XWLB 网络中对其他节点有更强的控制力.

XWLB 去 C 后的最大范围为 3990, 有 20 个节点游离在这个子网络之外, 无法链接到大部分的节点, 其中的 17 个节点更是完全孤立无法与其他任何节点链接. SHSS 去 C 后的最大范围为 2590, 有 10 个节点游离在这个子网络之外, 无法链接到大部分的节点, 其中的 5 个节点更是完全孤立无法与其他任何节点链接.

数据也显示, C 对平均路径长度和密度的影响与 B 相当, 都比 A 的影响要大. 与 B 类似, 去掉 C 后的网络密度仅为原始网络密度的一半左右. 因此我们认为, C 同样具有使网络中部分节点联系得更加紧密的能力, 且 C 的这种能力中不可替代的部分与 B 相当, 较 A 要多.

去 C 后直径变大, 这说明通过 C 节点能够缩短网络中部分节点间的距离. 与 A 相比, 去 C 后直径的增加更加明显, 但与去 B 后的直径是一致的. 因此我们认为, C 缩短节点间距离的能力中不可替代的部分与 B 相当, 较 A 要多.

A, B, C 三个节点虽然同为中心节点, 但其地位也有很大差别. A 是整个句法网络的最中心节点. B 是非常明确的局部节点, 中间度为 0. C 是局部中心节点, 整体中心度介于 A 与 C 之间. 在分别去除这三个节点后, 数据反映出了网络的不同变化. 其中, 最显著的特点是去除 B 后未造成任何孤立节点以及游离节点. 我们认为这与 B 的点入度、内接近性和中间度为 0 相关. 其中, 点入度是根源. 点入度为 0 意味着这一节点没有支配其他节点的能力, 它只能依附在其他节点之上. 假设它的某一邻居节点只能通过它和其他节点联系起来的话, B 就必须具备入度, 而这与 B 的真实特性相悖, 因此, 去掉 B 后不会有孤立节点和游离节点的出现. 同理可以推断, 之所以去掉 A 和 C 之后会出现孤立节点和游离节点, 是因为它们的入度不为 0. 至于虚词的节点入度是否为 0, 则是由词节点本身所具备的配价能力所决定的[18].

3 结论

本研究结果说明, 在依存句法中, 词的入度(即支配能力)比出度(即被支配能力)在维持句法结构的完整性上更重要. 这与传统句法研究中中心词是维持句法结构完整的重要组成部分的观点相符. 网络的研究方法为这一观点提供了宏观、可量化的数据支持. 本文也证明, 汉语的句法结构具有鲁棒性. 即使在去掉最中心节点的情况下, 仍能保持绝大部分节点的连通性. 而从复杂网络的角度来看, 研究表明在网络结构的研究中, 我们应重视节点自身的特性. 虽然宏观的网络数据很重要, 但网络的构建动力、它们的形成、发展和变化的背后是节点的个性. 节点根据自身的不同特性而自发地连接, 最终形成了小世界、无尺度的网络. 节点的个性才是决定各种网络结构的根源.

此外, 研究中我们还发现, 3 个虚词受语体的影响不同. 那么, 对语体敏感的虚词有哪些? 它们的网络统计特征能否作为语体研究的一个参数? 这些问题都值得进一步研究和探讨.

致谢 感谢本文所有树库标注者们的辛勤工作. 感谢刘望提供的程序支持. 本研究得到中国传媒大学“211 工程”三期重点学科建设项目的部分资助.

参考文献

1 Hudson R. Language Networks: The New Word Grammar. Oxford: Oxford University Press, 2007
2 刘海涛. 语言复杂网络的聚类研究. 科学通报, 2010, 55: 2667–2674
3 刘海涛. 语言网络: 隐喻, 还是利器? 浙江大学学报(人文社会科学版), 2010, doi: 10.3785/j.issn.1008-942X, 2010.10.041
4 Ferrer i Cancho R, Solé R V, Köhler R. Patterns in syntactic dependency networks. Phys Rev E, 2004, 69: 051915
5 Yu S, Liu H, Xu C. Statistical properties of Chinese phonemic networks. Physica A, 2011, 390: 1370–1380
6 Li J, Zhou J. Chinese character structure analysis based on complex networks. Physica A, 2007, 380: 629–638
7 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. 科学通报, 2005, 50: 1575–1579
8 Liu H. The complexity of Chinese dependency syntactic networks. Physica A, 2008, 387: 3048–3058
9 刘海涛. 汉语语义网络的统计特性. 科学通报, 2009, 54: 2060–2064
10 Liu H, Hu F. What role does syntax play in a language network? Europhys Lett, 2008, 83: 18002
11 刘宏鲲, 张效莉, 曹崑, 等. 中国城市航空网络航线连接机制分析. 中国科学 G 辑: 物理学 力学 天文学, 2009, 39: 935–942
12 Grimes B F. Ethnologue: Languages of the World. 14th ed. Dallas, TX: SIL International, 2000
13 黄伯荣, 廖序东. 现代汉语(增订三版). 北京: 高等教育出版社, 2002
14 刘源, 梁南元. 汉语处理的基础工程——现代汉语词频统计. 中文信息学报, 1986, 1: 17–25
15 冯志伟. 计算语言学基础. 北京: 商务印书馆, 2001. 75–76
16 北京语言学院语言教学研究所. 现代汉语频率词典. 北京: 北京语言学院出版社, 1986
17 de Nooy W, Mrvar A, Batagelj V. Exploratory Social Network Analysis with Pajek. Cambridge: Cambridge University Press, 2005
18 刘海涛, 冯志伟. 自然语言处理的概率配价模式理论. 语言科学, 2007, 3: 32–41

Central nodes of the Chinese syntactic networks

CHEN XinYing[1,2] & LIU HaiTao[1]

[1]*School of International Studies, Zhejiang University, Hangzhou 310058, China;*
[2]*Institute of Applied Linguistics, Communication University of China, Beijing 100024, China*

Based on two syntactic dependency treebanks built with two different styles of Chinese, a statistical study is conducted regarding word-frequency and distributions. We extracted three grammatical words as the research objects and analyzed their network features, including all degree, out-degree, in-degree, all closeness, in-closeness, out-closeness and betweenness. Then these three nodes were removed from the networks. We recorded and compared the network features of the two original networks and the three networks from which one node is respectively removed, including the number of vertices, average degree, average path length, diameter, the number of isolated vertices, domain and density. The results show that all three function words are central nodes of the Chinese syntactic networks but have different status. Their influence to the overall structure is also quite different. The research not only provides a new method for the study about Chinese grammatical words but also provides a new way of thinking the node characteristics in the complex network.

complex network, central node, language network, grammatical word

doi: 10.1360/972010-2369