汉语词组网的组织结构与无标度特性

韦洛霞① 李 勇① 康世勇② 罗诗裕①

(①东莞理工学院计算机科学与技术系，东莞 523106；②烟台师范学院中文系，烟台 264025. E-mail: weilx@dgut.edu.cn)

摘要 定义了汉语词组的随机文本模型(Monkey 模型)，揭示了真实汉语词组网具有无标度特性，指出了节点的平均度与总节点数的比值($\bar{k}/N$)基本上是个常量. 通过模拟词组网络的演化，发现了按字频选择汉字是导致词组网幂律度分布的重要原因. 适当地调节汉字的选择概率可以使汉语的Monkey语言表现出与自然语言类似的统计特性，大k时涌现幂律度分布，且幂指数大约是6. 对比Monkey语言和真实汉语得出，人类能更好地运用汉字资源并以简洁的方式表达意图，从而证明了汉语词组网的组织结构服从自然界普遍存在的最小代价原则.

关键词 复杂网络 汉字网络 度分布 无标度 随机文本 **Monkey** 语言

语言是人类沟通的工具，与人类同步进化. 据估计，目前正在使用的主要语言有6800多种[1]. 每一种语言都是在一定的语法规则下将语素连接成词或句的. 人类对语言的研究由来已久，特别是计算机的使用，为语言的信息化提供了契机，同时也给语言的信息化带来了挑战.

注意到，同许多自然网络、人工网络一样[2,3]，语言网也是由海量相互作用的语素构成. 从不同模型出发，人们构造出了不同的语言网，利用统计物理学的方法对它们的宏观性质进行了研究，并得到了大量有意义的结果[4~7]. 其中引人注目的是英语网的小世界效应和无标度度分布的涌现特性. 度分布$P(k)$的这种行为表明，无论系统多复杂，从$P(k)$对k的log-log图上可以看出，二者始终线形相关(即表现为一条简洁的直线)[4~6].

汉语是世界上使用人口最多的语言之一[1]，它传承着上下 5000 年的中华文明. 深入研究汉语言的组织结构，对提高其信息化水平，焕发其活力，具有深远的意义. 汉字是图形化文字，常用字就有 3755 个. 它们之间通过相互作用形成词组，进而构成句子. 人们一般认为，汉语是一种比较复杂的自然语言(与英语等拼音语言相比)，要实现高效信息化的难度更大. 不过，我们可以利用复杂网络的方法，将词组看作节点，将语素在两个节点中的同时出现看作一条连接，构造出汉语词组网. 通过分析，发现了汉语词组网的小世界效应及 3 度分隔等特征[8].

进一步的问题是：汉语词组网是否也存在幂律度分布；是否也可以找到一个好的结构模型对其进行描述；字频是否会影响词组网的状态(如果是，将产生何种影响). 基于这些考虑，本文首先给出了词组网的度分布，建立了汉语的Monkey(猴子)模型，通过比较真实词组与模型词组，揭示了汉语词组的演化机制及其内部的组织结构. 我们发现(1)词组网也具有无标度特性，正是这个特性保证了汉语网络的健壮性. (2) 词长分布也服从幂律，平均词长大约是2.5. (3) 尽管词汇的生长存在不确定性，但节点的平均度与总节点数的比值($\bar{k}/N$)基本保持不变. (4) 字频不仅决定了个体词组的诞生，还决定了词组网整体的度分布状态. (5) 适当调节语素的选择概率，可以使Monkey语言表现出与人类自然语言相近的统计特性. 对比发现，在同样节点数的条件下，更多真实词组拥有较少的连接，而更多 Monkey 词组拥有较多的连接. 如果将节点看作是粒子，将连接数看作它们之间的相互作用，不难看出，词组网的建构同样遵循普遍适用的能量最低原则.

1 真实词组网的动力学特性

在语言网中，新词汇的不断诞生以及新相互作用的不断出现，为语言系统带来了随机性和多样性，也为人们认识语言结构带来了困难. 为了解释汉语的结构特点，有必要研究其演化过程的动力学性质. 度分布函数$P(k)$就是描述这个性质的重要测度，它的定义是任意节点具有k度的概率. 对于Erdös和Rényi定义的随机网络，度分布服从泊松规律，而对于其他更加复杂的真实网络，比如生物网、生态网、技术网、特别是语言网等，度分布服从幂律[2~6].

本文研究了三组语料，第一组有 10746 个词(常用词)，第二组有 47951 个词(来源于 Internet 的统计)，

第三组有 96234 个词(来源于其他汉语研究机构的统计). 图 1 给出了各组词的度分布. 从网络生长的角度看, 我们可以把后一组词看作是由前一组词生长得到的. 在生长过程中, 尽管节点数变化较大, 但是词组的度分布却表现出了定态特征.

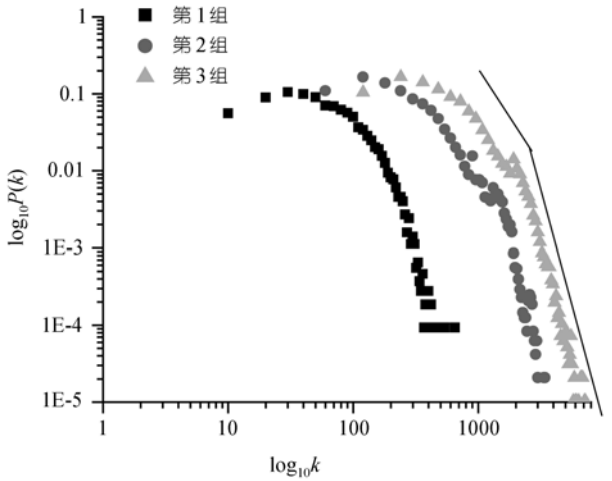

图 1 三组汉语词组的度分布. 大 k 时表现出幂律

从图 1 可以看出, 在大连接数时, 度分布为一条直线(幂律). 这说明, 无论词组网多复杂, 其幂律度分布仅需 1 个参数(斜率或幂指数)就可以描述. 此外, 这个幂律拖着一个长长的肥尾, 意味着随节点数的增多, 总会出现少数具有更高连接的节点, 而绝大多数节点仍然保持在低连接度水平. 计算表明, 对于这三组词: 当 $k>\bar{k}$ 时, 平均每度占有节点数分别只有 5.3, 5.4 和 7.5 个; 当 $k<\bar{k}$ 时, 平均每度占有节点数分别可达 97, 92 和 88 个. 结果指出, 词组的连接数总是倾向于保持在低度水平. 从图 1 还可以看出, 在小连接数时, 度分布表现为与指数相似的曲线. 从总体上看, 词组的度分布都表现出了非单纯幂律的特征. 我们认为在构建新词组时, 汉语不仅按字频优先选择语素, 还可能按其他方式选择, 例如随机选择语素等. 根据对英语网和其他复杂网络的研究, 我们推测, 按字频优先选择语素可能是大 k 时系统凸显幂律度分布的原因, 而按照其他方式选择语素则是小 k 时系统表现类似指数度分布的原因. 这种推测基本符合实际情况, 因为通常人们会选择组词能力更强的常用字构造新词, 例如“工人”、“中心”; 但有时新词的诞生也不遵守这个规则, 例如“徘徊”、“忐忑”.

图 2 给出了词长的概率分布 $P(\ell)$, 从它关于 l 的双对数图上可以看到一条陡峭的直线(即幂律分布), 直线的斜率分别是 5.92, 5.32 和 5.58. 值得注意的是 2 字词分别占总数的 77%, 68%和 64%, 2 字词和 3 字词占总数的 88%, 83%和 83%. 这说明人们习惯于用最简洁的语言传递信息, 充分体现出了蕴涵在语言中的最小代价原则.

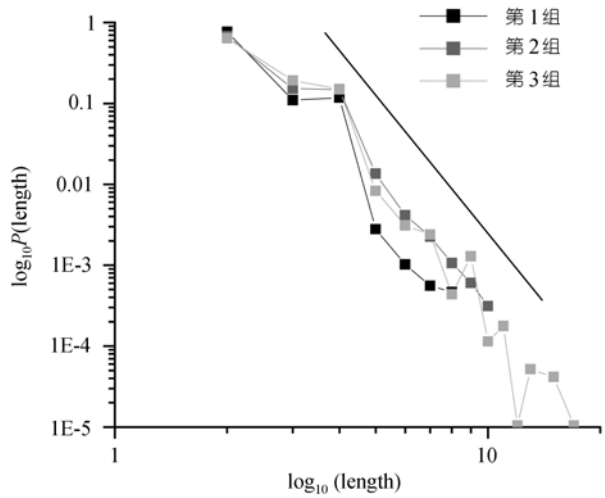

图 2 三组词组的词长分布. 在双 log 图上表现为一条(幂律)直线

2 建构模型

以等概率的随机方式键入英文字母, 如果空格出现的概率是 q, 那么每个字母被选中的概率就是 $(1-q)/26$. 这种产生文本的方式就是随机文本模型, 即所谓的Monkey语言[9]. 由于随机文本模型可用于模拟英语的结构和演化, 所以常被用来研究英语. 类似地, 我们也建立汉语的随机文本模型, 并用之研究汉语的演化. 除了前述的 3 组字频外, 我们还获得了 GB2312-80 规定的 6763 个信息交换用汉字频度. 模拟显示, 4 组频度都可以给出相同的结果.

将 6763 个汉字作为基本语素来构造词组网. 首先让我们建立基本的 Monkey 模型, 假设每个字拥有相同的选择概率 1/6763. 考虑到二字词占多数, 模拟仅包含二字词的网络生长, 得到了三组词汇的度分布(第一组有 10746 字, 第二组有 47951 字, 第三组有 96234 字), 它们都服从 Poisson 规律(如图 3 所示). 然后我们修正 Monkey 模型, 注意到每个汉字都有不同的含义, 组词能力也有很大差异, 等概率地选择汉字并不合适, 因此需要考虑字频因素. 基于这一点, 设定按字频随机选择汉字, 模拟二字词网并得到如图 4

所示的度分布. 可以看出, 词组的度分布背离了泊松规律, 表现出了幂律. 最后进一步修正 Monkey 模型, 我们还注意到真实词组的长度具有多样性, 仅仅模拟两字词也不恰当. 因此, 为了找到能更好地刻画自然语言的 Monkey 模型, 有必要将词长也作为构词的一个因素(词长的统计分布如图 2 所示). 即, 除了按字频优先选字外, 还要按词长频度随机决定新词的长度, 从而得到了图 5 所示的度分布.

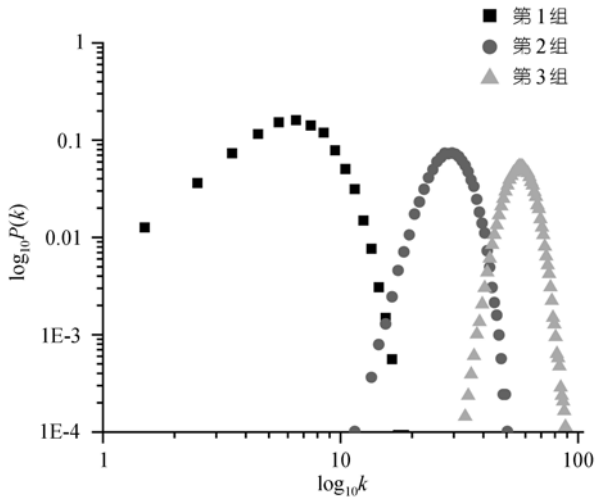

图 3　随机词组模型的度分布

等概率选字构造 2 字词, 得到的词组网是泊松分布

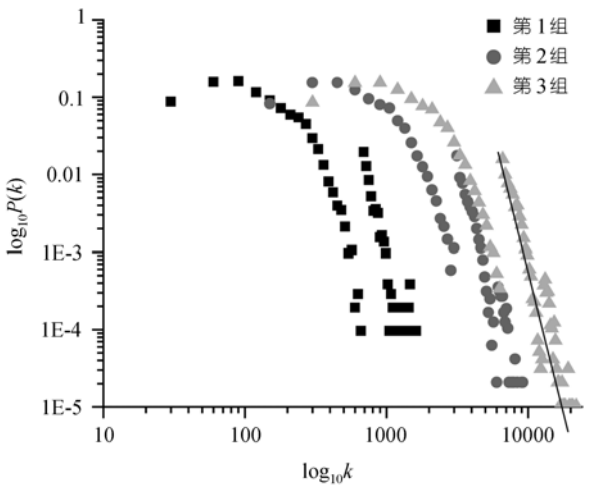

图 4　基于字频模型的度分布

按字频选字构造 2 字词, 在大 k 时看到了 power-law 斜线

为了更清晰地表明真实词组和 Monkey 词组之间性质的异同, 将二者放在同一个图中进行比较(如图 6). 可以看出, 大 k 时它们的度分布相似.

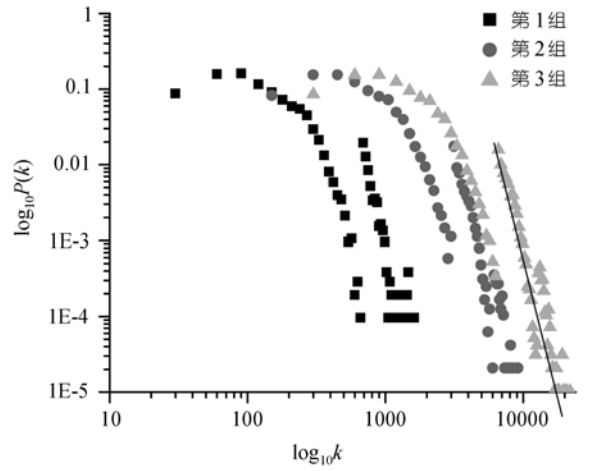

图 5　基于真实词组模型的度分布

考虑了字频、字长等因素, 模拟得到的分布曲线, 在大 k 时表现出 power-law, 且指数是 6.29

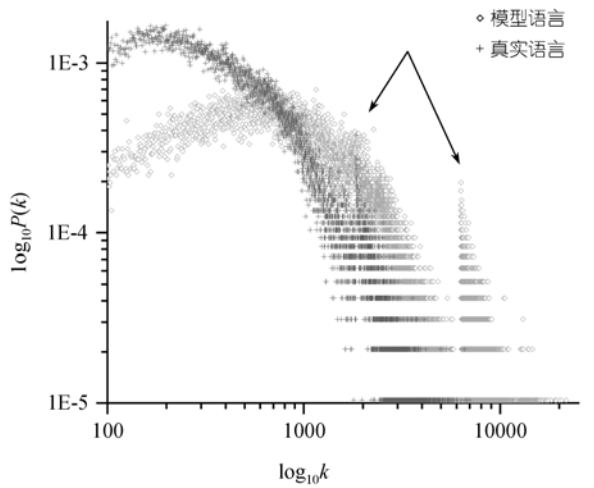

图 6　自然语言和模型语言的对比

图中十字叉是自然语言(较高、较窄), 菱形是 Monkey 词组(较低、较肥). 大 k 时二者度分布类似

3　讨论

等概率地选择汉字构造 2 字词, 每字被选中的机会都是 1/6763. 于是, 诞生一个 2 字词的概率就是 $P(2)=(1/6763)^2$. 以这种方式建构的词组网是具有泊松度分布的随机网络, 而随机网络不具备健壮性, 不能用于刻画自然语言的结构. 修改 Monkey 模型的组织方式, 使其按字频优先选字, 即假设语素的选择概率 $q(i)$)正比于它的频度, 就得到一个新词诞生的概率是

$$P(i,j,2)=q(i)q(j), \qquad (1)$$

其中 $i,j=\{1,2,\cdots,6763\}$. 这样的 2 字词网共有

45738169 个节点. 从图 4 看到, 词组网的度分布不再服从泊松规律, 随着节点数的增多它们表现出了定态特征, 特别是在大连接度时, 还表现出了幂律. 这个结果表明, 字频不仅决定了个体词组的诞生, 还决定了词组网整体的度分布状态.

在等概率的条件下, 产生每组 Monkey 词汇都用了几乎全部 6763 个字. 而按字频选择时, 即便是拥有最多词汇的一组随机文本(96234), 也只能利用到字频最高的 1969 个字. 其他 2/3 的字都被 Monkey 忽视. 对真实语料的研究显示, 拥有最少词的一组语料使用汉字 2613 个, 而拥有最多词的一组语料使用汉字 5605 个(表 1). 这表明人类运用汉字语素的能力远远超过 Monkey. 因此, 要能够更好地模拟真实汉语, 就有必要改进模型的组织方式. 在构建一个词组时, 汉字的字频决定了该汉字被选入词中的概率, 从而也决定了词节点诞生的概率. 由此看出, 在构词过程中字频扮演了重要角色. 此外, 还有其他因素也在起作用, 但这些因素对词组的形成影响较小, 这使我们有理由不去追究具体作用的细节, 而将它们整体上视为选字时的一个扰动. 于是, 任一汉字被选中的概率就是

$$p(i) = (1-\delta)q(i) + \delta \cdot 1/6763, \qquad (2)$$

表 1 三组词组网的性质

	第一组	第二组	第三组
节点数	10746	47951	96234
平均度	73	329	636
利用汉字数	2613	5072	5605
平均度/节点数	1/147	1/146	1/151
最高连接度	646	3374	7108
K<10 的节点数比例	0.0426	0.0098	0.0043
$k<\bar{k}$ 的节点数比例	0.64	0.66	0.60
最高连接度/平均度	8.849	10.255	11.176
度分布指数	6.50	6.29	5.45
词长分布指数	5.92	5.32	5.58
平均词长	2.36	2.53	2.56

其中δ是扰动因子. 当δ=0 时, 是完全按字频选字, 当δ=1 时, 是等概率地随机选字. 新词的诞生概率可表示为

$$P(i, j, 2) = p(i)p(j), \qquad (3)$$

取δ=0.1, 按照式(2)的概率选字组词, 产生了 25513 个词汇, 利用到了 4990 个汉字. 修改后的模型解决了原先所存在的选字过分集中的问题, 并且也没有引起 Monkey 词组度分布的改变. 图 4 给出了 3 组 Monkey 词汇的度分布, 从中可以看出, 大 k 时的行为服从幂律分布.

考虑到真实词组长度的差异, 在构词中除了按照式(2)随机选字外, 还要随机决定新词组的词长. 新词的长度ℓ, 是参照真实语料词长频度的分布随机获得的. 这样一个新词汇诞生的概率等于ℓ个 $p(k)$之积, 表示为

$$P(\ell) = \prod_{k=1}^{\ell} p(k), \qquad (4)$$

其中 $p(k)$ 由式(2)给出. 式(4)意味着等概率时, 概率的对数与词长成正比. 图 5 显示的结果与图 4 非常接近, 这是因为 2 字词占绝大多数(64%以上), 但长词的出现, 导致更多的高连接度节点的出现, 使图 5 的尾部向右移动, 从而使直线的斜率发生变化.

从图 6 可以看出, (1) Monkey 词组的最大连接度远大于真实词组的最大连接度. 这说明, Monkey 对个别汉字的过度使用, 导致了少数节点获得了过度冗余的连接. 而在真实语言的建构中, 当某些语素被反复使用后, 人们会选择其他语素来替代. (2) 尽管存在一些具有极高连接度的节点, 但在全局上, 真实词组仍保持在低连接度水平. (3) 按第三组的平均词长 2.55 计算, 能够产生的总词数为 6,384,940,166 个. 该组词汇看似已有足够多的节点(96234 个), 但同总词数相比它也仅仅是的一个很小子集(6 万分之一). (4) 还可以看出, 在大 k 时度分布表现出了震荡行为. 在英语等拼音语言中未发现这种情况. 此外还发现了其他有趣的现象; 例如, 由表 1 数据计算可得, 相邻两组词长的比分别是 1.072 和 1.012, 最高连接度与平均度之比分别是 8.849, 10.255 和 11.176. 后一组的 $k_{\max}/\bar{k}$ 等于前一组的 $k_{\max}/\bar{k}$ 加上后一组与前一组的词长比(8.849+1.072 = 9.921±0.03, 10.255 + 1.012 = 11.267±0.008). 这些问题是否代表了特定的意义, 是否表达了汉字结构的特殊信息, 我们今后将进一步的讨论.

4 结论

众所周知, 英语与汉语是在不同的语法规则下, 由不同的语素构成. 用复杂网络的方法描述他们时, 不仅节点的定义不同, 而且节点间相互作用方式的定义也不同. 于是, 人们自然就认为它们是不同结构的语言网, 它们会给出不同的统计性质. 然而, 我们对汉语词组网的研究却表明, 二者具有一系列相同的特征, 比如小世界效应以及幂律度分布[8].

本文通过对汉字网络动力学性质的描述，发现大k时凸现幂律，幂指数大约为 6(英语单词网是 1.5 和 2.7[5,6]，英语概念网是 3.5[4]). 尽管汉语系统中，词组的生长具有很强的不确定性，但是平均度与总节点数的比值($\bar{k}/N$)基本上是一个常量. 我们建立了汉语的Monkey模型并以之模拟人类的语言行为，发现字频不仅决定了个体词组的诞生几率，还决定了词组网整体度分布的状态，按字频优先选字构词是汉语词组网无标度行为的主因. 通过调节语素的选择概率，可以使Monkey语言表现出与人类自然语言相似的统计特性. 在汉语词组的建构中, Monkey仅取少数惯用的汉字，这种行为导致个别节点的连接度出现过度冗余. 而人类的行为模式是充分运用汉字资源并以简洁的方式来表达自己的意图. 这种行为方式的结果是使词组网保持在低连接态. 可见，汉语词组网的组织结构也服从自然界中普遍存在的最小代价原则.

参 考 文 献

1 Grimes B F, ed. Ethnologue: Languages of the World. 14th ed. Dallas(TX): SIL International, 2000

2 Montoya J M, Sole R V. Small world patterns in food webs. J Theor Biol, 2002, 214(3): 405~412[DOI]

3 Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the Internet topology. Comput Commun Rev, 1999, 29: 251~258[DOI]

4 Motter A E, de Moura A P S, Lai Y C, et al. Topology of the conceptual network of language. Phys Rev E, 2002, 65: 065102(R)

5 Ferrer-i-Cancho R, Solé R V. The small world of human language. Proc R Soc Lond B, 2001, 268(7): 2261~2265[DOI]

6 Ferrer-i-Cancho R, Solé R V, Köhler R. Patterns in Syntactic Dependency Networks. Phys Rev E, 2004, 69: 051915(1~8)

7 Albert R, Barabasi A-L. Statistical mechanics of complex networks. Rev Mod Phys, 2002, 74(1): 47~97[DOI]

8 韦洛霞, 李勇, 李伟, 等. 汉字网络的 3 度分隔与小世界效应. 科学通报, 2004, 49(24): 2615~2616

9 Casti J L. Bell curves and monkey language. Complexity, 1995, 1(1): 12~15

(2004-12-10 收稿, 2005-05-11 收修改稿)