科学通报 2013 年 第 58 卷 第 5-6 期：432 ~ 437
论 文 www.scichina.com csb.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS

基于平行词同现网络的语言聚类

刘海涛*, 丛进

浙江大学外国语言文化与国际交流学院, 杭州 310058
* 联系人, E-mail: lhtzju@gmail.com

2012-09-13 收稿, 2012-11-15 接受
国家社会科学基金(09BYY024, 11&ZD188)资助

摘要 考察了在语言精细分类中使用复杂网络以及在基于复杂网络的语言分类中使用平行词同现网络替代句法依存网络的可行性. 采用 12 种斯拉夫语言和 2 种非斯拉夫语言的平行文本, 构建了 14 个词同现网络. 通过这些网络的主要参数的恰当组合, 聚类分析能够将斯拉夫诸语言与非斯拉夫语言区分开来, 并能将 12 种斯拉夫语言正确地划分到各自的语支中去. 另外, 聚类也能反映某些斯拉夫语言在其语支内部的亲缘关系. 结果表明, 平行词同现网络能够被用于语言的精细分类, 而且在基于复杂网络的语言分类中可被用作句法依存网络的一种更为便捷的替代品.

关键词
词同现网络
斯拉夫语言
平行文本
语言分类
聚类分析

复杂网络无所不在, 几乎渗透到自然界与人类活动的各个方面[1]. 近年来, 复杂网络开始被运用于涉及人类语言的理论与应用性研究中[2]. 将语言作为一个系统的观念是现代语言学中最重要的假设之一[3]. 如果语言是一个系统, 那么其在系统层面的组织方式就无法通过注重语言结构细节的语言学传统方法来得到充分体现与描述. 而能够以整体观考察各种系统的复杂网络可以弥补语言学在方法上的欠缺. 语言的诸多方面及结构层面都能作为语言网络(以相关的语言单位为节点, 以它们之间的某种关系为边)来进行建模和描述[4~6].

对语言网络的定量分析可望成为语言学中不同领域的潜在研究方法. 语言分类就是一个有代表性的例子. 研究表明[7~9], 我们可以采用主要的复杂网络参数对句法依存网络(以词形为节点, 以它们之间的句法依存关系为边)进行聚类分析, 以达到对相应的语言进行分类的目的. 分类的结果能大致反映这些语言在语言谱系中的亲缘关系. 这种基于复杂网络的语言分类属于侧重语言结构特征的类型学分类[10]. Liu 和 Xu[8]的研究发现, 句法依存网络的复杂网络参数是语言的形态与句法特征在系统层面上的体现. 因此, 基于复杂网络的语言分类是对整体类型学[11]的一个重要贡献. 基于复杂网络的语言分类的研究[7~9]表明, 复杂网络的主要参数除了能揭示真实世界中网络的共性(例如在各种网络中普遍发现的统计特征, 包括小世界和无尺度属性)之外, 还能反映网络的多样性. 同时, 复杂网络在语言分类中的使用拓展了复杂网络的应用领域, 也拓宽了复杂网络研究的视野.

值得注意的是, 此前基于复杂网络的语言分类研究[7~9]一般只满足于将不同的语言大致划入各自的语族(如罗曼语族、日耳曼语族和斯拉夫语族), 并没有对各语族中的语言进行细分, 即目前尚无研究采用基于复杂网络的方法进行较为精细的语言分类(例如, 将同一语族的语言分入不同的语支). 如果能证明使用复杂网络可以得出精细的语言分类, 那么语言类型学将有可能更多地受益于这种基于复杂网络的方法, 复杂网络的应用也能被拓展到人文与社会科学中更为具体的领域中去. 就方法而言, 此前这些基于复杂网络的语言分类主要存在两大问题. 一方

引用格式: 刘海涛, 丛进. 基于平行词同现网络的语言聚类. 科学通报, 2013, 58: 432–437
英文版见: Liu H T, Cong J. Language clustering with word co-occurrence networks based on parallel texts. Chin Sci Bull, 2013, 58, doi: 10.1007/s11434-013-5711-8

面，在这些研究中，构建句法依存网络所用的语料在语义内容和语体方面的一致性难以保证．基于句法依存网络的语言分类的基本假设是：句法依存网络在拓扑结构上的异同(由其复杂网络参数体现)反映了相应的语言之间的异同[9]．所选语料在语义内容和语体上的不一致性虽然与语言之间的异同无关，但其仍然有可能影响相应的句法依存网络在拓扑结构上的异同，进而影响语言分类的结果．基于复杂网络的语言分类的更为适宜的语料是平行文本(即语义内容一致但语言不同的文本的集合，如某小说的原本及其不同语言的译本)，这种语料在语义内容和语体上是一致的．另一方面，句法依存网络的构建需要耗费大量人力、物力．句法依存网络是由句法依存树库转换而来的，而后者是通过对生语料进行句法依存标注得到的．尽管有自动化的标注方法可供使用，但是如果要获得能够满足语言学研究的标注精度，标注过程仍须以人工的方式逐词逐句完成．因此，即便句法依存网络能成为语言分类的一种有效方法，考虑到其构建过程的困难性，也难以将其运用到涉及较多语言的分类或语言类型研究中去．另外，句法依存标注方法的不同也可能影响句法依存网络的拓扑结构特征，进而影响语言分类的结果．因此，我们需要寻找一种更易获得的语言网络来作为句法依存网络的替代品．在所有其他类型的语言网络中，词同现网络[12]最有可能胜任这一角色(详细介绍见“方法与资源”部分)．鉴于以上两个问题，我们可以考虑在基于复杂网络的语言分类中采用基于平行文本的词同现网络(以下称“平行词同现网络”)作为对句法依存网络的一种可能的替代品．

本研究考察在语言精细分类中使用复杂网络以及在基于复杂网络的语言分类中使用平行词同现网络替代句法依存网络的可行性．我们在 12 种斯拉夫语言和 2 种非斯拉夫语言的平行文本的基础上分别构建了 14 个词同现网络，并通过其主要复杂网络参数的不同组合对这些网络进行聚类分析．对分类效果的评估是通过聚类结果与这些语言(尤其是12种斯拉夫语言)在语言谱系中的亲缘关系的比对来进行的．

1 方法与资源

词同现网络是由真实语料转换而来的．在本研究中，我们将“同现”定义为两个词形在句中的相邻关系．例如，在“John kicked the ball”中有三对相邻的词形，即 John kicked, kicked the 和 the ball．因此一个词同现网络可以表示为一个无向图 $G = (V, E)$，其中 V 是节点的集合，表示语料中所有不同的词形；而 E 是边的集合，表示词形在组句时形成的所有不同的相邻关系．因此，如果两个词形在至少一个句子中存在相邻关系，那么其对应的节点 u，$v \in V$ 将被一个边 $e \in E$ 所连接．根据这一定义，我们可以从真实语料中提取词形在组句时形成的所有不同的二元组，并将该二元组的集合转换为词同现网络．词同现网络可以通过自动的方式来构造．使用词同现网络的一个主要优势在于它的无歧义性，因为同现关系可以被明确地定义并且能够以理论中立的方式从语料中提取出来．图 1 展示了一个按照以上定义所构建的词同现网络(材料取自史迪芬·平克《语言本能》的第一章)．若无特别说明，以下文中提到的词同现网络均指按照以上定义所构建的网络类型．

一个词同现网络和一个句法依存网络——假设它们均基于相同的真实语料——仅在边的类型上有所不同．前者的边表示词形在句中的相邻关系，而后者的边表示词形在句中的句法依存关系．对诸多不同语言的研究数据表明[13]，一个句法依存关系在较大概率上(一般在 50%以上)存在于两个相邻的词形之间．这意味着词同现网络与基于相同真实语料的句法依存网络在拓扑结构上具有较高的相似性，因为二者的边存在显著的重合．例如，图 1 中的词同现网络的中心节点一般为虚词，这与句法依存网络的情况[14,15]是一致的．因此，词同现网络在语言网络研究中可以作为句法依存网络的一个可能的替代品．一个词同现网络的复杂网络参数可以被用作与之对应的句法依存网络的相同参数的一种方便的近似估计，能在系统层面上大致反映一种语言的形态和句法特征．

本研究构建的词同现网络所基于的平行文本包括 14 种语言：俄语、白俄罗斯语、乌克兰语、捷克语、斯洛伐克语、波兰语、上索布语、塞尔维亚语、克罗地亚语、斯洛文尼亚语、保加利亚语、马其顿语、英语和汉语．14 种语言中有 12 种为斯拉夫语言，分别属于三个语支，即东斯拉夫语支(俄语、白俄罗斯语和乌克兰语)、西斯拉夫语支(捷克语、斯洛伐克语、波兰语和上索布语)和南斯拉夫语支(塞尔维亚语、克罗地亚语、斯洛文尼亚语、保加利亚语和马其顿语)[16]．这些平行文本系小说《钢铁是怎样炼成的》(Kak

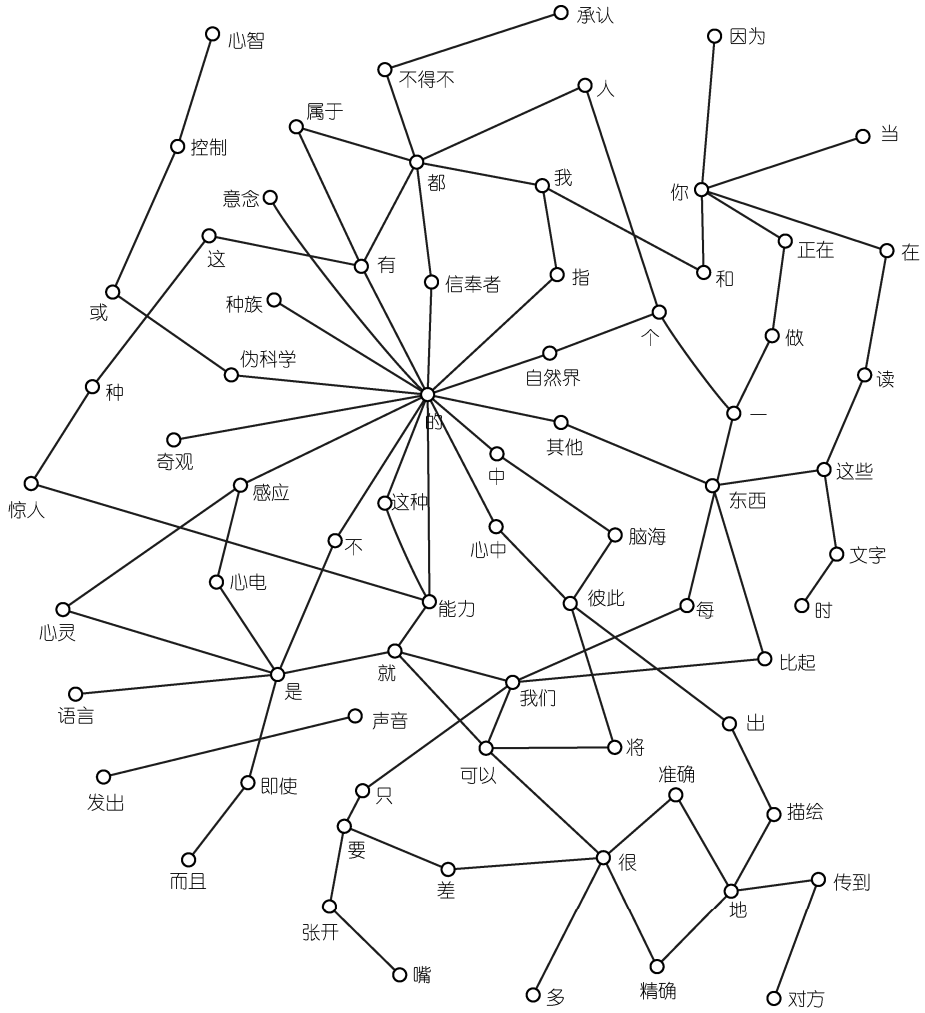

图 1 一个汉语的词同现网络

zakaljalas' stal')的俄语原著(N. A. 奥斯特洛夫斯基著于 1932~1934 年期间)和其他 13 种语言的译本. 其中 12 种斯拉夫语言的文本系来自 Emmerich Kelih 所建的斯拉夫语平行语料库(详细介绍见文献[17]), 而英语和汉语的文本是我们自行从这两种语言的译本中获得的. 由于这 14 种语言中有 12 种同属于斯拉夫语族, 并分属不同的斯拉夫语支, 这为检验使用平行词同现网络进行语言精细分类的效果提供了条件.

我们采用复杂网络分析平台 Cytoscape 的插件之一 NetworkAnalyzer[18]计算了 14 个词同现网络的 10 个复杂网络参数. 这些复杂网络参数是: 平均度($\langle k\rangle$)、平均路径长度(L)、聚集系数(C)、网络中心度(NC)、直径(D)、网络异质度(NH)、与 $P(k)$(度分布)拟合最佳的幂律的指数(γ_1)、与 $P(k)$拟合最佳的幂律的决定系数(R^2_1)、与 $\overline{k}_{nn}(k)$ (相邻节点平均度的分布)拟合最佳的幂律的指数(γ_2)以及与 $\overline{k}_{nn}(k)$ 拟合最佳的幂律的决定系数(R^2_2)(对这些参数及其应用的详细介绍见文献[9,19]).

以上这些参数足以呈现一个复杂网络的拓扑结构特征的概貌, 例如它是否为小世界或无尺度网络. 聚类分析在语言分类中的使用至少可以追溯到 Altmann 和 Lehfeldt[20]的研究. 基于这些复杂网络参数的不同组合, 聚类分析被用于 14 个词同现网络. 这些参数在参与聚类之前都经过标准化. 聚类分析采用离差平方和法和曼哈顿距离. 根据此前基于复杂网络的语言分类研究的经验[7~9], 我们选取$\langle k\rangle$, L, C 和 NC 的组合作为基准集. 其他的参数组合系通过在基准集的基础上添加其他参数得到. 共有 64 个参数

组合在聚类分析中得到检验.

2 结果与讨论

按照前面介绍的方法, 我们得到了 14 个词同现网络的主要参数, 结果见表 1.

对分类效果的评估通过聚类结果与这些语言在语言谱系中的亲缘关系的比对来进行. 由于 14 种语言大都为斯拉夫语言, 我们侧重于考察聚类结果如何反映 12 种斯拉夫语言之间的亲缘关系. 评估分类结果的基本标准是 12 种斯拉夫语言必须首先聚类, 其次再与 2 种非斯拉夫语言聚类. 换言之, 聚类结果必须能将 12 种斯拉夫语言与 2 种非斯拉夫语言区分开来. 如果满足这一标准, 我们再考察 12 种斯拉夫语言是否被正确地分入各自的语支当中.

在被检验的 64 个复杂网络参数组合中, 有 15 个组合的聚类结果能将斯拉夫语言与非斯拉夫语言区分开来, 并将 12 种斯拉夫语言正确分入各自的语支中. 图 2 展示的是这些结果中的一个, 是由基准集加 D, R_1^2, γ_2 和 R_2^2 的组合得出的. 图 2 较好地呈现了斯拉夫语族的细分情况, 12 种斯拉夫语言都被准确地划分到了各自的语支中. 另外, 聚类也能反映某些斯拉夫语言在其语支内部的亲缘关系. 例如, 尽管塞尔维亚语和克罗地亚语使用不同的书写系统, 但一般认为它们是同一种语言[16]. 如图 2 所示, 塞尔维亚语和克罗地亚语在其语支内以 1.70 的距离被聚为一类. 保加利亚语和马其顿语之间的亲缘关系也得以反映(距离为 3.57). 对斯拉夫语言分类的这一结果要稍好于 Kelih[17]基于相同的斯拉夫语平行语料库、通过考察斯拉夫语言中的型例关系而得出的结果. 后者仅仅得出了一个 12 种斯拉夫语言的序列, 能够反映它们之间亲缘关系的远近, 但无法体现它们应如何分类. 该分类结果与采用包括词汇统计学[21]在内的其他方法所得到的结果大致具有可比性.

聚类分析也涉及到英语和汉语这 2 种非斯拉夫语言. 如图 2 所示, 英语和汉语作为一个聚类与 12 种斯拉夫语言作为另一个聚类之间的距离为 39.33, 而英语和汉语之间的距离为 3.34. 这一结果不仅反映了英语和汉语作为非斯拉夫语言与 12 种斯拉夫语言之间的差异, 也反映了英语与汉语之间的相似性. 二者的相似性在此前基于真实语料的研究中[7,22]也有发现.

本研究所用方法的自动化程度较高, 而对人工参与的要求较低. 例如, 该方法无须考虑不同语言的书写系统. 而且书写系统的差异被证明不会影响到语言分类的结果. 在 12 种斯拉夫语言当中, 俄语、白俄罗斯语、乌克兰语、塞尔维亚语、保加利亚语和马其顿语使用西里尔字母, 而其他 6 种语言则使用拉丁字母. 然而如图 2 所示, 这些语言在书写系统上的差异对其分类并无影响. 这也引起我们对语言与书写系统之间关系的思考. 例如汉语从其特殊的书写系统来看, 与英语的差异似乎非常大. 然而, 从本研究以及文献[7,22]中的结果来看, 二者的差异实际上比

表 1 14 种语言的词同现网络的主要参数

	$\langle k \rangle$	L	C	NC	D	NH	γ_1	R_1^2	γ_2	R_2^2
白俄罗斯语	4.819	3.797	0.100	0.114	17	5.833	1.232	0.742	0.451	0.794
保加利亚语	5.690	3.354	0.186	0.144	11	6.767	1.159	0.711	0.525	0.855
汉语	8.684	2.944	0.283	0.354	9	6.113	1.180	0.755	0.534	0.930
克罗地亚语	5.353	3.479	0.151	0.127	13	6.574	1.212	0.712	0.505	0.847
捷克语	4.945	3.627	0.119	0.157	13	6.696	1.257	0.75	0.500	0.873
英语	9.043	2.964	0.299	0.297	10	5.499	1.157	0.743	0.533	0.883
马其顿语	6.206	3.225	0.220	0.170	10	6.698	1.138	0.724	0.546	0.841
波兰语	4.983	3.628	0.118	0.112	14	6.351	1.229	0.720	0.475	0.824
俄语	4.504	3.891	0.091	0.109	17	5.972	1.268	0.748	0.444	0.757
塞尔维亚语	5.348	3.485	0.147	0.126	15	6.543	1.213	0.707	0.515	0.832
斯洛伐克语	5.166	3.592	0.128	0.137	14	6.255	1.235	0.747	0.477	0.836
斯洛文尼亚语	5.367	3.406	0.164	0.192	13	7.400	1.192	0.738	0.565	0.787
乌克兰语	4.865	3.814	0.096	0.076	16	5.433	1.254	0.764	0.424	0.737
上索布语	5.347	3.550	0.131	0.161	14	6.359	1.239	0.741	0.466	0.822

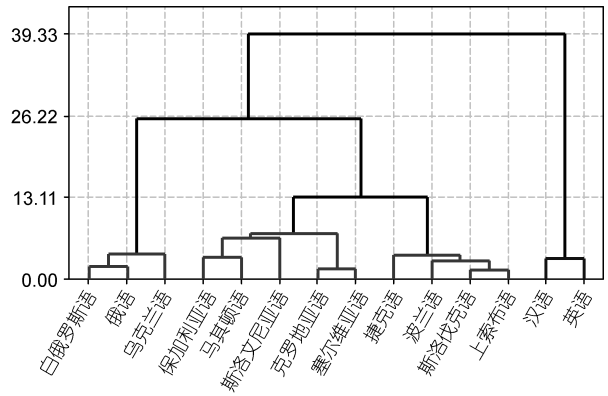

图 2 基于 8 个复杂网络参数对 14 个词同现网络的聚类结果

想象中的要小得多. 另外值得注意的是, 本研究对斯拉夫语言的分类效果要明显好于 Liu[22]采用依存方向等语序指标所得到的结果. 这是因为本研究所采用的方法依据的是语言作为一个系统的整体特征, 而非一系列难以反映语言整体性质的局部结构细节. 这也表明, 对于像斯拉夫语言这样具有较丰富的屈折形态变化的语言[23]来说, 语序可能不是其分类的最佳依据. 另外, 由于本研究的方法完全是从定量的角度去进行语言分类, 它反映出来的语言之间的异同是连续性的, 而非离散的.

3 结论

本研究考察在语言精细分类中使用复杂网络以及在基于复杂网络的语言分类中使用平行词同现网络替代句法依存网络的可行性. 我们在 12 种斯拉夫语言和 2 种非斯拉夫语言的平行文本的基础上分别构建了 14 个词同现网络, 并通过其主要复杂网络参数的不同组合对这些网络进行了聚类分析. 基于这些参数的恰当组合, 聚类分析能够将斯拉夫语言与非斯拉夫语言区分开来, 并能将斯拉夫语言正确地划分到各自的语支中去. 另外, 聚类也能反映某些斯拉夫语言在其语支内部的亲缘关系. 因此可以做出这样的结论: 平行词同现网络能够被用于语言的精细分类, 而且在基于复杂网络的语言分类中可被用作句法依存网络的一种更为便捷的替代品. 本研究所采用的方法也有助于建立一种注重整体特征的、定量的、能反映语言之间连续性差异的语言类型学研究路向. 本研究进一步证实了使用主要的复杂网络参数能够反映出真实世界中网络的多样性. 更为重要的是, 由于平行词同现网络被证明能够用于语言的精细分类, 复杂网络的应用因此能够被进一步拓展到人文与社会科学中更为具体的领域.

致谢 感谢 Emmerich Kelih 为本研究提供斯拉夫语平行文本.

参考文献

1 Costa L D F, Oliveira O N, Travieso G, et al. Analyzing and modeling real-world phenomena with complex networks: A survey of applications. Adv Phys, 2011, 60: 329–412
2 Choudhury M, Mukherjee A. The structure and dynamics of linguistic networks. In: Dynamics on and of Complex Networks, Modeling and Simulation in Science, Engineering and Technology. Boston: Birkhaeuser, 2009. 145–166
3 Kretzschmar W A. The Linguistics of Speech. New York: Cambridge University Press, 2009
4 Steyvers M, Tenenbaum J B. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. Cognit Sci, 2005, 29: 41–78
5 Ferrer i Cancho R, Solé R V, Köhler R. Patterns in syntactic dependency networks. Phys Rev E, 2004, 69: 051915
6 刘海涛. 汉语语义网络的统计特性. 科学通报, 2009, 54: 2060–2064
7 刘海涛. 语言复杂网络的聚类研究. 科学通报, 2010, 55: 2667–2674
8 Liu H T, Xu C S. Can syntactic networks indicate morphological complexity of a language? Europhys Lett, 2011, 93: 28005
9 Abramov O, Mehler A. Automatic language classification by means of syntactic dependency networks. J Quant Ling, 2011, 18: 291–336
10 Ruhlen M. A Guide to the World's Languages 1: Classification. Stanford: Stanford University Press, 1991
11 Shibatani M, Bynon T. Approaches to language typology: A conspectus. In: Approaches to Language Typology. New York: Oxford University Press, 1995. 1–26
12 Ferrer i Cancho R, Solé R V. The small world of human language. Proc R Soc Lond B, 2001, 268: 2261–2265
13 Liu H T. Dependency distance as a metric of language comprehension difficulty. J Cognit Sci, 2008, 9: 159–191

14 Solé R V, Corominas-Murtra B, Valverde S, et al. Language networks: Their structure, function and evolution. Complexity, 2010, 15: 20–26
15 陈芯莹, 刘海涛. 汉语句法网络的中心节点研究. 科学通报, 2011, 56: 735–740
16 Katzner K. The Languages of the World (New Edition). London and New York: Routledge, 1995
17 Kelih E. The type-token relationship in Slavic parallel texts. Glottometrics, 2010, 20: 1–11
18 Assenov Y, Ramirez F, Schelhorn S E, et al. Computing topological parameters of biological networks. Bioinformatics, 2008, 24: 282–284
19 Costa L D F, Rodrigues F A, Travieso G, et al. Characterization of complex networks: A survey of measurements. Adv Physics, 2007, 56: 167–242
20 Altmann G, Lehfeldt W. Allgemeine Sprachtypologie. Munich: Fink, 1973
21 Novotná P, Blažek V. Glottochronolgy and its application to the Balto-Slavic languages. Baltistica, 2007, XLII: 185–210
22 Liu H T. Dependency direction as a means of word-order typology: A method based on dependency treebanks. Lingua, 2010, 120: 1567–1578
23 Comrie B, Corbett G G. Introduction. In: The Slavonic Languages. London: Routledge, 2002. 1–19