DOI: 10.3785/j.issn.1008-942X.CN33-6000/C.2015.12.231

依存距离分布有规律吗?

陆 前 刘海涛
(浙江大学 外国语言文化与国际交流学院,浙江 杭州 310058)

[摘 要]探索语言的普遍特征一直是语言学研究的重要内容,当前依存距离最小化已经被证实是人类语言的一种普遍规律。为了发现这一规律背后的动因,对30种语言的依存距离分布情况进行研究,通过多种模型拟合对比,发现广延指数分布和指数截断的幂律分布分别适合拟合"短句"与"长句"的依存距离分布。研究结果还显示,人类语言的依存距离分布介于指数分布和幂律分布之间,可用指数和幂律混合的模型来描述。在此基础上,利用不同模型拟合对比来探讨依存距离分布的方法和路径,结果揭示出人类语言的依存距离可能遵循一种普遍性的分布模式,反映了省力原则和人类认知机制在语言结构运用与演化过程中发挥着重要的支配作用。

[关键词]依存距离;依存距离分布;幂律分布;依存树库;认知机制;语言结构

Does Dependency Distance Distribute Regularly?

Lu Qian Liu Haitao
(*School of International Studies, Zhejiang University, Hangzhou* 310058, *China*)

Abstract: Universal properties of languages have always been important in traditional linguistics study. In recent years, studies have increasingly presented a trend which integrates multiple disciplines and methods, e. g. cognitive science, network science, big data analysis and quantitative techniques. So far, results of the survey on large-scale cross-language materials have indicated that human languages have a tendency toward dependency distance minimization. This tendency suggests that, although human languages differ in pronunciation, vocabulary and grammar, etc., their syntax may be bound by universal mechanisms, and their evolution may also have a universal model.

Dependency distance, which is defined as the linear distance between two words which are syntactically related, can reflect the comprehension difficulty of syntactic structure. Therefore, the dependency distance minimization is considered as resulting from cognitive mechanism and the

[收稿日期]2015-12-23
[在线优先出版日期]2016-04-22
[本刊网址·在线杂志]http://www.journals.zju.edu.cn/soc
[网络连续型出版物号]CN33-6000/C
[基金项目]国家社会科学基金重大项目(11&ZD188);中国博士后科学基金资助项目(2015M571852)

[作者简介]1.陆前(http://orcid.org/0000-0003-2470-3598),男,浙江大学外国语言文化与国际交流学院博士后研究人员,文学博士,主要从事计量语言学、计算语言学等方面的研究;2.刘海涛(http://orcid.org/ 0000-0003-1724-4418)(通信作者),男,浙江大学外国语言文化与国际交流学院求是特聘教授,博士生导师,文学博士,主要从事计量语言学、语言复杂网络、配价理论与依存语法、语言规划与语言政策等方面的研究。

effect of "the principle of least effort" on syntactic structure. It also proves that humans prefer to avoid the use of long-distance dependencies to reduce cognitive cost. As a result, dependency distance distribution may present a certain pattern. Revealing this pattern will help us understand how human cognitive mechanism works on syntactic structure.

But the question is which of the probability distributions can fit the pattern of dependency distance distribution more properly — the power law distribution or the exponential distribution? To find out the answer, this paper uses the following methods and materials to analyze dependency distance distribution: 1) Complementary Cumulative Distribution Function (CCDF) is used to smooth data, to avoid statistical fluctuation, and to lower fitting error; 2) Maximum likelihood estimation and likelihood ratio test are used to fit and compare five kinds of "heavy tail" distribution, including exponential and power law; 3) HamleDT 2.0 dependence treebank is adopted, especially for language materials which are annotated with Prague Dependencies Scheme, because this annotation scheme is most similar to the traditional dependency grammar, and more helpful to find the rules of language structure.

With these methods and materials, this research analyzes dependency treebanks of 30 languages, and summarises the following findings: 1) Complementary Cumulative Distribution indicates that distribution of dependency distance in human languages proved has certain regularity; 2) for the majority of 30 languages, the distribution of dependency distance conforms to certain models, namely, Stretched Exponential Distribution (SED) for "short sentences" and Truncated Power Law Distribution (TPLD) for "long sentences;" 3) although dependency distance distribution patterns differ among languages, they all fit, in essence, to a mixed exponential and power law distribution; 4) the debate over exponential distribution and power law distribution might mainly be caused by different fitting methods, different languages, different sentences length and text size, etc. The findings of this research will help us better understand the nature of the dependency distance minimization. It may also reveal that the dependency distance in human languages may abide by a certain universal distribution pattern. At the same time, the findings may contribute significantly to constructing the syntactical synergetic subsystem in the framework of dependency grammar.

Key words: dependency distance; dependency distance distribution; power law; dependency treebank; cognitive mechanism; language structure

一、引 言

探索语言的普遍特征一直是传统语言学研究的重要内容[1-3],并日渐呈现出多学科交叉与多方法融合的研究态势[4-5]。针对大规模跨语言材料的实证研究已表明:人类语言存在依存距离最小化的倾向[6-9]。这种倾向性说明人类语言尽管在语音、词汇和语法等方面千差万别,但语法结构可能受相同机制的约束,语言系统的演化也具有统一的模式。事实上,依存距离最小化思想已被应

用于构建高效的自然语言处理系统中①，但依存距离为什么会倾向于最小化以及受什么机制影响等问题还有待探讨。因此，深入挖掘依存距离最小化现象背后所隐藏的动因，将有助于我们进一步阐释语言的普遍特征。

依存距离是依存语法研究领域的一个重要概念，指的是句中两个有句法关系的词之间的线性距离[10-11]。依存距离的大小反映了人类认知机制对句法结构关系的约束作用[6-8,12]。有关句子加工机制的语言复杂性研究表明，当进行句法分析时，句子中的词被不断地储存于工作记忆之中，只有当词的支配词出现时，它才能被删除[2,6]。因此，依存距离越大意味着词被存储的时间越长。由于人类的工作记忆容量有限，当存储的词的数量超过记忆容量时，就会造成理解困难。与之相反，较小的依存距离能在一定程度上降低句法复杂性。也就是说，依存距离倾向最小化与人类认知机制的运作方式密不可分，它可能是降低语言理解难度的重要手段之一[6,13]。

同时，依存距离倾向最小化还与人类行为所遵循的“省力原则”[14]有关。正是出于“省力”的考虑，在语言运用中，人类会尽量避免使用可能导致认知成本增加的长距离依存关系。已有研究认为，省力原则是语言单位的频率、长度等计量特征呈幂律分布的内在机制[15]。从复杂网络角度看，幂律分布可能是系统中某一对象或变量的增长和择优过程（preferential attachment process）的数理表达方式[16]。语言也是一种复杂网络[4]，而依存距离则是其句法关系在线性距离上的量化单位。因此，我们会很自然地想到幂律可能是导致依存距离最小化的动因。换言之，如果我们确定了真实语言的依存距离分布源于一种概率分布模型，那么这种分布模型就可能在人类认知机制约束句法关系的显现中发挥关键作用。

然而，当前关于依存距离分布模型的研究却有不同的意见。刘海涛认为，句子中依存距离的大小与其出现频率的关系遵从一个简单假设，即变量频率与变量的相对变化率之间成反比关系，并由此推导出依存距离的概率分布是一种右截尾的 Zeta 分布（Right-truncated Zeta Distribution）[17]。这表明依存距离分布符合齐夫定律，是一种离散型的幂律分布。但是，Ferrer-i-Cancho 将依存距离的先验分布和平均依存距离受限作为约束条件，通过最大熵原则推导出依存距离分布是一种指数分布[7]。针对这一争议，蒋景阳与刘海涛用 6 种幂律类分布和指数分布分别对英汉平行语料进行了拟合优度检验，发现这些拟合结果在统计学意义上均可被接受[18]。为了搞清楚依存距离究竟更符合哪一种分布，我们认为，值得对下列问题进行研究：依存距离分布有没有规律？依存距离分布的规律是否符合一定的分布模型？依存距离分布模型是否会因不同语言而异？导致上面两种分布模型分歧的主要原因是什么？

本文拟对 30 种语言的真实语料进行依存距离分布研究以回答上述问题。采用大规模多语种语料是出于以下两方面的考虑：首先，尽量避免依存距离受句长、文本类型和语法标注方式的影响[18-19]，这些因素可能会导致数据拟合偏差与错误结论；其次，依存距离最小化是人类语言的普遍规律，已在几十种语言中被证实[6-7,9]，那么确定依存距离分布是否有规律可能也需要更多的语言来验证，这样才能使结论更具客观性和普遍性。

---

① N. A. Smith & J. Eisner，"Annealing Structural Bias in Multilingual Weighted Grammar Induction，" in Proceedings of the 21st International Conference on Computational Linguistic and the 44th Annual Meeting of the Association for Computational Linguistics，Sydney：Association for Computational Linguistics，2006，pp. 569－576；D. Klein & C. D. Manning，"Corpus-based Induction of Syntactic Structure：Models of Dependency and Constituency，" in Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics，Barcelona：Association for Computational Linguistics，2004，pp. 478－485.

二、方法与材料

(一) 依存距离计算方法

依存语法认为,一种语言的句法结构由词之间的依存关系构成,并遵循如下约束规则:(1)依存关系是词与词的二元关系;(2)依存关系是一种非对称关系,两个句法相关词中一个为支配词,另一个为从属词;(3)依存关系是有标记的,其类型可以被显性地标示出来。因此,句子是多个由支配词和从属词构成的依存关系连接而成的线性序列[3,20]。

图1是句子"I like red apples"的依存树,其中"like"是句子的根,句子下方为词性标注。句中有三个依存关系分别为:"like → I"形成的主谓关系(SBV),"like → apples"形成的动宾关系(VOB),"apples → red"形成的定中关系(ATT)。上面每组关系中词与词或相互临近,或相隔几个词,这种词间的间隔距离可以用依存距离(Dependency Distance,DD)来计算[10-11,17]。例如,图1中"red"和"apples"的依存距离 DD("apples","red")=π("apples")−π("red")=4−3=1,其中 π("apples")表示词在句中的线性位置编号。

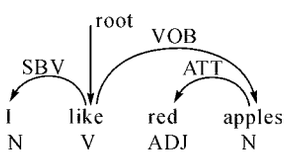


图1 句子"I like red apples"的依存树

(二) 概率分布模型拟合方法

在数学上,如果观测数据符合幂律分布(Power Law Distribution,PLD),那么它的概率分布为:

$$p(x)\propto x^{-\alpha} \tag{1}$$

其中 α 为幂指数。当观测数据为离散型时,其 PLD 密度函数为:

$$p(x)=\Pr(X=x)=Cx^{-\alpha} \tag{2}$$

公式中常数 C 一般用黎曼函数 ζ 进行归一化。因此,公式(2)可以改写成:

$$p(x)=\frac{x^{-\alpha}}{\zeta(\alpha)} \tag{3}$$

如果对公式(3)等式两边取对数,那么变换后的函数在双对数坐标上为一条直线。而观测数据若符合指数分布(Exponential Distribution,ED),那么它具有如下的概率密度函数:

$$p(x)=\Pr(X=x)=\lambda e^{-\lambda x} \tag{4}$$

在线性坐标下,ED 函数曲线与 PLD 类似,都是具有"长尾"特征的内凹曲线。在双对数坐标下,ED 呈外凸的曲线,而 PLD 则为一条直线,这个函数性质通常被用来检验数据是否服从幂律关系。然而,由于数据采样数量不足,往往会产生统计波动,表现为依存距离分布图形在尾部逐渐变宽。为了减少这种拟合误差,一般用互补累积分布函数(Complementary Cumulative Distribution Function,CCDF)来平滑数据[21-24],这是因为 PLD 的累积分布仍然服从幂律[23],只不过幂指数会

减 1。但是，在基于最小二乘法的拟合优度检验中，非 PLD 容易得到较高的决定系数（R^2），所以高 R^2 很难成为一种判断 PLD 假设的有力依据[24]。

事实上，在现实中很少有观测数据能够完全契合 PLD。Clauset 等人认为数据中往往是大于某个数值 $x_{\min}$ 的部分才符合 PLD[24]。因此，公式（3）中黎曼函数 $\zeta(\alpha)$ 可以用赫尔维茨 ζ 函数 $\zeta(\alpha,x_{\min})$ 来替代：

$$p(x)=\frac{x^{-\alpha}}{\zeta(\alpha,x_{\min})} \tag{5}$$

其中，$x_{\min}$ 可以通过计算理论拟合模型与原始数据之间的最小 Kolmogorov-Smirnov（KS）距离获得，该方法也能用于拟合优度检验；而模型参数 α 的计算则可用极大似然估计法，该法优于最小二乘法[25]。与此相应，ED 可以由下列公式进行定义：

$$p(x)=(1-e^{-\lambda x})e^{\lambda x_{\min}}\cdot e^{-\lambda x} \tag{6}$$

PLD 不仅与 ED 近似，还与其他众多概率分布模型相近，例如对数正态分布（Log-normal Distribution，LND）、指数截断的幂律分布（Truncated Power Law Distribution，TPLD）、广延指数分布（Stretched Exponential Distribution，SED）等等。在本文中，上面三种“长尾”分布的定义如表 1 所示。它们之间的拟合效果差异可以用似然比检验进行判断[22,24]。本文将用 R 表示两种模型之间的似然比，正值（$+R$）表示前一个模型更契合数据，而负值（$-R$）则相反，且当 $p<0.05$ 时表示对比结果具有统计学意义上的显著性。

表 1　其他三种用于模型对比的“长尾”分布定义

分布	概率密度函数
LND	$p(x)=\sqrt{\frac{2}{\pi\sigma^2}}\left[\operatorname{erfc}\left(\frac{\ln x_{\min}-\mu}{\sqrt{2}\sigma}\right)\right]^{-1}\cdot\frac{1}{x}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$
TPLD	$p(x)=\frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha,\lambda x_{\min})}\cdot x^{-\alpha}e^{-\lambda x}$
SED	$p(x)=\beta\lambda e^{\lambda x_{\min}^{\beta}}\cdot x^{\beta-1}e^{-\lambda x^{\beta}}$

（三）语料

本研究的语料为 HamleDT 2.0 依存树库[25]，该树库是专为研制与评测依存句法分析系统而构建的大规模多语种语料库，其中每种语言树库分别用布拉格标注法（PRG）[①]和通用斯坦福标注法（USD）[26]进行标注。本文选用了布拉格标注法所标注的语料，目的是为了避免不同标注法对数据拟合产生不利影响。除此之外，与斯坦福标注体系相比，布拉格标注方法也更接近于传统依存语法[3,20]。

HamleDT 2.0 依存树库的相关统计数据如表 2 所示。在表 2 中，“体裁”为依存树库中语料的语体类型，“结构”指树库在转换为依存树库（PRG 标注）之前的句法标注形式。

在分析依存距离分布规律前，我们对依存树库进行了预处理，主要操作有：（1）将树库中的训练集和测试集合并为一个语料集合。（2）标点符号的处理。如果标点只作为从属节点，那么删除该标点；如果标点是其他词的支配节点，那么保留该节点，并将依存距离归一化为 1。（3）空元素（Null

① J. Hajic, J. Panevová & E. Hajicová et al., "Prague Dependency Treebank 2.0," CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, ISBN 1-58563-370-4, Philadelphia, 2006.

elements)的处理。它们主要出现在孟加拉语、印地语和泰卢固语等几种语言树库中,处理方式如(2)。(4)将树库中全部句子中词的序号进行无间隔的编排,以减少统计误差。

表2 HamleDT 2.0依存树库中30种语言的相关数据统计

语言	词次	句子数	平均句长	体裁	结构
阿拉伯语(ar)	277 423	7 547	36.76	新闻	依存结构
保加利亚语(bg)	196 151	13 221	14.84	混合	短语结构
孟加拉语(bn)	7 252	1 129	6.42	—	依存结构
加泰罗尼亚语(ca)	443 317	14 924	29.70	新闻	短语结构
捷克语(cs)	1 503 738	87 913	17.10	新闻	依存结构
丹麦语(da)	100 238	5 512	18.19	会话	依存结构
德语(de)	680 710	38 020	17.90	新闻	短语结构
希腊语(el)	70 223	2 902	24.20	混合	依存结构
英语(en)	451 576	18 791	24.03	新闻	短语结构
西班牙语(es)	477 810	15 984	29.89	新闻	短语结构
爱沙尼亚语(et)	9 491	1 315	7.22	小说	短语结构
巴斯克语(eu)	151 604	11 226	13.50	混合	依存结构
波斯语(fa)	189 572	12 455	15.22	新闻	依存结构
芬兰语(fi)	58 576	4 307	13.60	混合	依存结构
古希腊语(grc)	308 882	21 160	14.60	混合	依存结构
印地语(hi)	294 509	13 274	22.19	—	依存结构
匈牙利语(hu)	139 143	6 424	21.66	混合	短语结构
意大利语(it)	76 295	3 359	22.71	混合	依存结构
日语(ja)	157 172	17 753	8.85	会话	依存结构
拉丁语(la)	53 143	3 473	15.30	混合	依存结构
荷兰语(nl)	200 654	13 735	14.61	新闻	短语结构
葡萄牙语(pt)	212 545	9 359	22.71	新闻	短语结构
罗马尼亚语(ro)	36 150	4 042	8.94	新闻	依存结构
俄语(ru)	497 465	34 895	14.26	混合	依存结构
斯洛伐克语(sk)	901 298	57 435	15.69	混合	依存结构
斯洛文尼亚语(sl)	35 140	1 936	18.15	小说	依存结构
瑞典语(sv)	197 123	11 431	17.24	混合	短语结构
泰米尔语(ta)	9 581	600	15.97	新闻	依存结构
泰卢固语(te)	5 722	1 450	3.95	—	依存结构
土耳其语(tr)	69 695	5 935	11.74	混合	依存结构

注:孟加拉语、印地语、泰卢固语未注明体裁。

三、结果与讨论

（一）依存距离分布规律

我们分别统计了 30 种语言中句长为 20 的句子的依存距离累积分布情况，结果如图 2 所示（其中每一条线表示一种语言，泰卢固语语料较少，只抽取了句长为 11 的句子）。

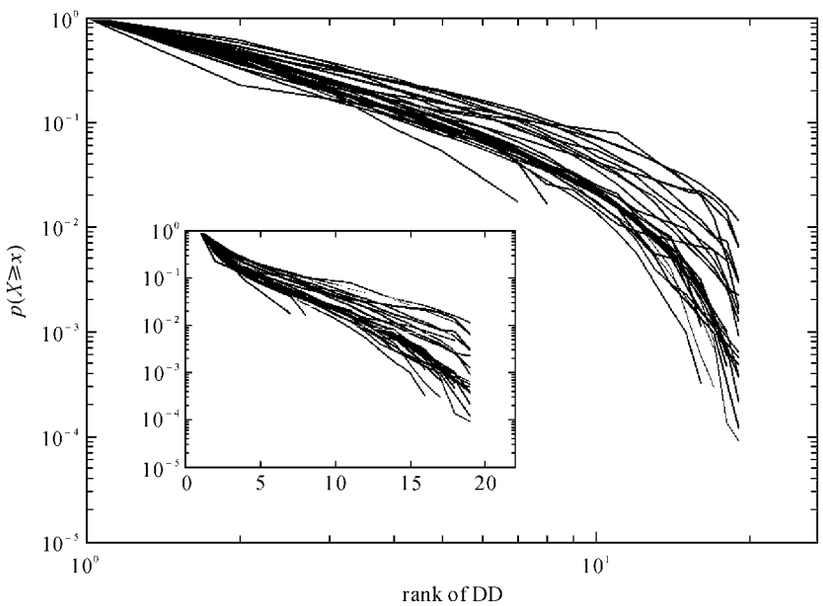


图 2　30 种语言句长为 20 的依存距离累积分布图

图 2 中主图是依存距离分布在双对数坐标下的情况，插入图展示的是在线性—对数坐标下的情况。在双对数坐标下，当依存距离的值小于 10 时，其分布近似直线；大于 10 后，则变成下弯的曲线。而在线性—对数坐标下，当依存距离为 1 至 10 时，其分布呈由上而下的曲线，然后逐渐近似于直线。图 2 中 30 种语言的依存距离分布趋势有迹可循，这说明依存距离分布具有一定的规律性。

（二）依存距离分布模型

图 2 揭示了依存距离分布的规律性，但并不能说明其符合哪种分布模型。另外，理论上在双对数坐标下 PLD 是一条直线，ED 是一条曲线，而在线性—对数坐标下两者图形正好相反。由图 2 来看，依存距离分布也不符合前文中提到的具有争议性的两类模型。我们认为，可以将极大似然估计和似然比检验引入到模型分析中来。Ferrer-i-Cancho 等人指出，不能用不同句长混杂的句子集来研究依存距离分布，因为拟合所得出的分布模型可能只是不同句长的分布模型的叠加[19]。另有研究表明，系统中个体模式（特定句长句子集）和整体模型（不同句长混合句子集）也可能存在差异[27]。因此，本文按句长对 HamleDT 2.0 依存树库中 30 种语言的语料进行归类，然后做升序排列，再对特定句长句子集的依存距离数据进行多模型拟合。为了保证分析结果的有效性，我们统计的语料范围是按句长划分的、数量超过 30 的句子集合。分析结果如图 3 所示。

图 3 展示了两种拟合方案的结果：被拟合数据起始值 $x_{\min}$ 为 KS 检验所得最优值；设 $x_{\min}$ 为 1，考察依存距离分布整体拟合状况。图中主要展示了 PLD 和 ED，TPLD 和 SED 的对比结果。可以

看出,从某个句长开始,在进行比较的两对模型中,PLD 和 TPLD 拟合数据的效果更好。这说明每种语言在不同句长的情况下,句集所适用的分布模型是有差别的。

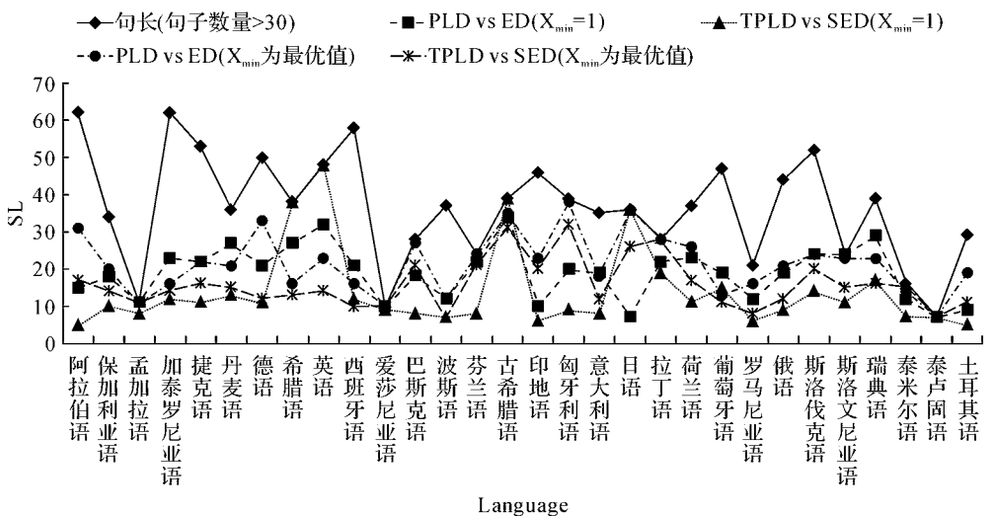


图 3 30 种语言的拟合模型与句长的关系

图 3 对依存距离数据进行两组分布模型拟合,其结果可粗略地划分为 4 类,如表 3 所示。

表 3 多模型拟合依存距离分布结果分类统计表

类别	语料	x_{min}为最优值 PLD vs. ED	语言(x_{min}为最优值)	PLD vs. TPLD 或 SED 或 LND	$x_{min}=1$ TPLD vs. SED	语言($x_{min}=1$)
C1	前端	$-R$	阿拉伯语,保加利亚语,加泰罗尼亚语,捷克语,丹麦语,德语,希腊语,英语,西班牙语,波斯语,印地语,意大利语,罗马尼亚语,俄语,斯洛伐克语,瑞典语,泰米尔语,土耳其语	$-R$,$p\approx0$	$-R$	阿拉伯语,保加利亚语,加泰罗尼亚语,捷克语,丹麦语,德语,西班牙语,巴斯克语,波斯语,芬兰语,印地语,意大利语,荷兰语,罗马尼亚语,俄语,斯洛伐克语,瑞典语,泰米尔语,土耳其语
	后端	$+R$		$-R$,$p\approx0$	$+R$	
C2	前端	$-R$	荷兰语,葡萄牙语	$-R$,$p\approx0$	$-R$	斯洛文尼亚语,瑞典语,拉丁语,匈牙利语,波斯语,丹麦语
	后端	$-R$ 或 $+R$		$-R$,$p\approx0$	$-R$ 或 $+R$	
C3	前端	$-R$	巴斯克语,芬兰语,古希腊语,匈牙利语,日语,拉丁语,斯洛文尼亚语	$-R$,$p\approx0$	$-R$	希腊语,英语,古希腊语,日语
	后端	$-R$		$-R$,$p\approx0$	$-R$	
C4	前端	$-R$	爱沙尼亚语,孟加拉语,泰卢固语	$-R$,$p\approx0$	$-R$	爱沙尼亚语,孟加拉语,泰卢固语
	后端	$-R$		$-R$,$p\approx0$	$-R$	

图 3 和表 3 中的统计数据向我们揭示了以下几点:

(1)对大多数语言而言,现有的单一模型无法很好地描述不同句长的依存距离分布,因为语料前端(较短句的句集)依存距离分布倾向于 ED,而随句长的增加,语料后端(较长句的句集)则倾向于 PLD。

(2)对大多数语言而言，在 PLD 与其他模型的对比中，x_{min} 为最优值和 x_{min} 为 1 两种情况下，似然比 R 绝大多数为负值，并且 p 近似为 0.00。这说明 TPLD、SED 和 LND 等模型都比 PLD 拟合效果好。

(3)大多数语言的拟合结果表明，TPLD 非常适合较长句的句集依存距离分布，而较短句的句集则更适合用 SED 进行拟合。

(4)当 x_{min} 为最优值时，与 PLD 相比，有 7 种语言完全倾向于 ED，但若以 TPLD 与 SED 拟合数据，则仅有芬兰语、拉丁语近似完全倾向于 SED；而当 x_{min} 为 1 时，有 4 种语言符合 SED。产生这种情况的原因是 x_{min} 取值不同，导致模型拟合依存距离的效果各不相同。

(5)孟加拉语、爱沙尼亚语和泰卢固语这三种语言的 5 个相关指标几乎重叠，这表明它们适合用 ED 或 SED 描述。但由于这几种语言语料本身较少，所以当我们截取每种句长句子数量大于 30 的句子集时，最大句长仅约为 10，因此无法判断较长句子适合哪种分布模型。

综上可以看出，依存距离分布模型介于 ED 和 PLD 之间，并且与 ED 和 PLD 相比，SED 和 TPLD 更适合拟合“短句”与“长句”的依存距离分布。尤其是当 x_{min} 为 1，大多数语言在句长约小于 10 时，SED 拟合较好；大于 10 后，TPLD 拟合较好。

(三) 不同语言依存距离分布模式异同

从上文的研究中，我们知道不同语言的依存距离分布规律存在差异。主要表现在，大多数语言中不同句长的句子并不适用单一模型，需要用 SED 和 TPLD 这两种分布模型来描述，且两者描述的句长区间也往往不同。此外，一些语言的拟合结果完全倾向于 SED。表面看来，上述情况似乎说明不同语言系统可能具有不同的依存距离分布模型，但事实并非如此。实际上，不管是 SED 还是 TPLD，都能够较好地拟合类似幂律行为的混合分布模型。

首先，从这两个模型的数学定义来看。TPLD 概率密度函数是在幂律分布函数后面简单地加上指数衰减因子，因此在分布前端的有限区域内近似于幂律行为，分布的尾部呈指数下降趋势；在 SED 函数公式中，参数 β 等于 1 时，函数变为 ED，而当参数 β 小于 1 时则变为指数幂律函数混合分布，尤其是参数 β 越小，整个分布函数就越接近于 PLD。由于这两种分布都具有两个参数，因此能够比 PLD 更好地对一些类似幂律行为的现象进行描述。下面，我们以表 3 中当 x_{min} 为 1 时依存距离分布倾向于 SED 的四种语言来说明 SED 与 ED 和 PLD 的关系。

我们统计了英语、日语、希腊语和古希腊语四种语言的依存距离分布，当以 SED 拟合时，其参数 β 随句长变化的趋势如图 4 所示。图 4 说明，当英语、希腊语和古希腊语句长为 6—20，日语为 4 时，参数 β 接近于 1，即此时 SED 表现了依存距离分布近似 ED 的情况；随着句长的增加，参数 β 逐渐减小，SED 表现为指数幂律混合的分布。这一点充分说明，尽管表面看来，不同语言满足不同分布模型，甚至有的满足混合分布模型，但其背后隐藏的规律是相同的，即依存距离分布随句长增加由类指数分布逐渐向类幂律分布过渡。从图 4 中，我们还发现日语与另外三种语言随着句长的增大，参数 β 的变化趋势差异较大。这种表现可能与语言的平均句长有关(日语平均句长约为 9，英语、希腊语、古希腊语分别为 24、24、15)，限于篇幅，相关问题我们将另文讨论。

其次，从语言系统的运作方式来看。我们的研究表明，大多数语言较短句句集的依存距离分布是一种近似 ED 的模型，这一点可以用最大熵原则进行验证[7]。但随着句长增加，长距离依存关系出现的概率不断增大，语言系统面临句法复杂性不断增大的压力[6,12-13,18]。这时，如果长句依然表现为 ED，那么根据 ED 收敛速度快于 PLD 这一函数性质，可以推断，长句中较小依存距离的占比相近，而较大依存距离的占比急剧下降。这可能会使语言系统依存距离趋向最小化，进而降低句法复杂性。但与此同时，较小依存距离的聚集规模并不明显，并且严重缺乏长依存关系。实际上，语

言使用的省力原则一方面要求依存距离为1的比例趋向于占绝对多数;另一方面,为了保证交流的顺利进行,必须依赖一定数量的长依存关系来增加句式的多样性。因此,作为一种自适应系统[4,13],语言在长句上采取的调节举措是:一方面使依存距离分布的前端表现为幂律,即不同依存距离的占比相差悬殊;另一方面,最大程度地缩小那些几乎等于句长的长依存关系的占比。这就导致了在长句中依存距离分布的前端呈一条逐渐下降的直线,而尾部表现为指数下降的趋势,这恰恰可以用TPLD进行较好地描述,也充分说明了长句的依存距离最小化可能更能体现语言作为一种自适应系统的特点。

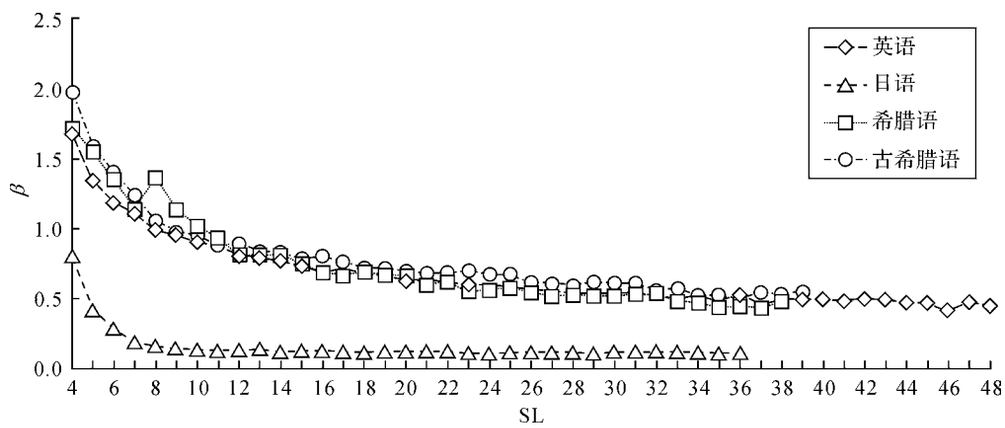


图4 英语、日语、希腊语和古希腊语SED中参数β随句长变化的趋势

综上所述,我们可以认为30种语言的依存距离分布都可以用幂律与指数混合的分布模型来描述。

(四)当前分布模型分歧的原因

根据本文研究,我们发现造成当前依存距离分布研究分歧的主要原因有三点。首先,由于幂律行为可能出现在分布的局部位置,故相关拟合手段都不太精确,很难区分PLD和其他相近分布之间的差异。例如,刘海涛用PLD拟合汉语的依存距离分布[17],蒋景阳与刘海涛使用PLD和ED拟合英汉两种语言($10\leqslant SL\leqslant 30$)[18],拟合结果都具有较高的决定系数,但利用最小二乘法对比PLD和其他模型之间的差异是比较困难的[24];而Ferrer-i-Cancho则是在线性—对数坐标下通过观察分布曲线的形状来判断模型类别的[7]。其次,不同语言、句长和文本量都对分布模型的确定有着至关重要的影响。Ferrer-i-Cancho使用的语料是罗马尼亚语($SL=6,9$)和捷克语($SL=12,18$)两种语言树库中特定句长的句子[7]。本研究表明,罗马尼亚语在3—11的句长区间,捷克语在3—21的句长区间内表现为类指数分布,而在其他句长上则表现为类幂律分布。这说明依存距离分布研究中,句长在一定程度上限制了幂律分布规律的隐现。最后,由于依存距离分布是一种指数和幂律混合的模型,因此,ED和PLD在一定程度上都仅对依存距离分布的部分特点进行了描述,这就需要我们从模型拟合效果和依存距离分布生成机制两个方面进行分析,才能最大限度地避免这种分歧。

此外,刘海涛以5个汉语依存树库为基础探讨了相邻与非相邻依存关系分布[28],发现在汉语依存树库中依存距离为1的依存关系比例较大,约为47.9%—56.6%。但他在文中并没有深入考察非相邻依存关系中不同依存距离的分布情况,因此,也就无法精确地得出依存距离的概率分布模型。本文在此基础上,将依存距离分布研究进一步精确化,有助于更科学地揭示依存距离最小化的动因。

还需要注意的是,本文并没有展开对PLD拟合优度(p值)的讨论。Clauset等提出可以采用

自举法(bootstrapping)和 KS 检验来计算 PLD 拟合优度[24]，但该法不适合衡量特定句长的依存距离分布模型的拟合优劣。因为依存距离的取值范围较小(本文研究的句子最大句长为 62)，在句子数非常庞大的情况下，依存距离仍然分布在固定的那些数值区域内，这相当于放大了理论值和观测值之间的距离，自然会扭曲实际的 p 值大小。因此，通过计算多模型之间的似然比，获取拟合效果较佳的模型，可能是当前在依存距离分布研究中判断幂律行为的一种较为合理的方法。

(五) 加泰罗尼亚语的依存距离分布——个案分析

为了更好地展现我们的研究结果，下面将以表 3 中 C1 类的语言加泰罗尼亚语作为个案，详细阐述我们关于依存距离分布的研究，统计 PLD 与 ED 及其他模型之间的差异性指标，如表 4 所示。

表 4　加泰罗尼亚语依存距离多模型拟合结果对比表(x_{min}为最优值)

句长	x_{min}	PLD vs. ED		PLD vs. SED		PLD vs. TPLD		PLD vs. LND		TPLD vs. SED	
		R	p	R	p	R	p	R	p	R	p
3	1	−10.74	0.00^{+}	−26.88	0.00^{+}	−8.92	0.00^{+}	−26.88	0.00^{+}	−2.37	0.02^{+}
4	2	−9.67	0.00^{+}	−12.39	0.00^{+}	−7.20	0.00^{+}	−12.39	0.00^{+}	−2.78	0.01^{+}
5	3	−5.95	0.00^{+}	−6.23	0.00^{+}	−4.72	0.01^{+}	−6.23	0.00^{+}	−2.23	0.03^{+}
6	2	−7.51	0.00^{+}	−6.56	0.00^{+}	−7.54	0.00^{+}	−6.12	0.00^{+}	−0.95	0.34
7	4	−4.20	0.00^{+}	−3.25	0.00^{+}	−3.70	0.04^{+}	−3.03	0.00^{+}	−0.89	0.37
8	5	−1.54	0.12	−1.69	0.09	−1.42	0.12	−1.47	0.14	−0.20	0.84
9	2	−8.41	0.00^{+}	−8.93	0.00^{+}	−12.53	0.00^{+}	−8.47	0.00^{+}	−0.81	0.42
10	8	−0.97	0.33	−5.50	0.00^{+}	—	—	−5.50	0.00^{+}	—	—
11	8	−1.42	0.16	−1.68	0.09	—	—	−1.56	0.12	—	—
12	9	−2.26	0.02^{+}	−2.55	0.01^{+}	—	—	−2.55	0.01^{+}	—	—
13	10	−0.95	0.34	−2.57	0.01^{+}	—	—	−2.57	0.01^{+}	—	—
14	2	−2.28	0.02^{+}	−11.14	0.00^{+}	−12.78	0.00^{+}	−10.69	0.00^{+}	8.60	0.00^{+}
15	10	−4.05	0.00^{+}	−1.51	0.13	−3.97	0.14	−1.52	0.13	−1.01	0.31
16−19	2	0.95	0.36	−11.58	0.00^{+}	−13.52	0.00^{+}	−11.09	0.00^{+}	10.04	0.00^{+}
20−28	2	6.57	0.00^{+}	−12.10	0.00^{+}	−12.60	0.00^{+}	13.84	0.00^{+}	13.83	0.00^{+}
29	15	−0.83	0.00^{+}	−0.95	0.34	−1.05	0.19	−0.92	0.36	0.28	0.78
30−62	2	10.63	0.00^{+}	−7.21	0.00^{+}	−9.91	0.00^{+}	−7.01	0.00^{+}	10.97	0.00^{+}

注：“—”表示由于数据太少，模型无法成功拟合；$^{+}$表示对比结果具有显著性，下同；表中倒数第 1、3 和 4 行中的相关数值都是平均值。

表 4 中被拟合数据起始值 x_{min} 为 KS 检验所得最优值。结果显示，当句长为 3—15 和 29 时，幂律与指数分布之间的似然比 R 为负值，而其他情况下 R 都为正值，尤其是当句长大于 19 时，p 值大多数等于 0.00。这表明较短句子的依存距离分布不服从 PLD，而长句则相反。但较短句子是否服从 ED 是值得商榷的。这是因为当句长为 3—15 和 29 时，x_{min} 的估计值较大，这导致这些集合中进行拟合的依存距离仅为尾部的几个数据，可能会造成拟合对比结果的扭曲。

PLD 与其他分布模型的对比结果表明，SED、TPLD 和 LND 都更好地拟合依存距离分布。从

表4中第5、9列可以看出,SED和LND比较相近,而第11列则表明TPLD更好地拟合了句长较长的句集(14≤SL≤62,除了15)。其他语言也表现出类似结果,这就是图3中只展示了TPLD与SED对比结果的原因。

表5是设置x_{min}为1的拟合情况。表中第3列显示,句长为3—23时,ED比PLD拟合的效果好(R是负值);当句长为24—62时,PLD则拟合得更好一些(R是正值)。然而,表中第5列显示其他分布模型比上面的两者更适合描述依存距离分布(R都是负值)。表中第8列是对比TPLD与SED的拟合结果,可发现,当句长为3—11时,后者较好地拟合了依存距离分布;而当句长为12—62时,前者则更加契合依存距离分布,其中具有统计显著性的占72%。

表5 加泰罗尼亚语依存距离多模型拟合结果对比表(x_{min}为1)

句长	x_{min}	PLD vs. ED		PLD vs. SED或TPLD或LND		句长	TPLD vs. SED	
		R	p	R	p		R	p
3—20	1	−13.01	0.00^{+}	$-R$	0.00^{+}	3—9	−2.79	0.00^{+}
21	1	−9.67	0.16	$-R$	0.00^{+}	10—11	−0.85	0.46
22	1	−5.95	0.55	$-R$	0.00^{+}	12—15	1.35	0.28
23	1	−7.51	0.86	$-R$	0.00^{+}	16	−0.12	0.90
24—62	1	9.32	0.00^{+}	$-R$	0.00^{+}	17—62	3.06	0.00^{+}(72%)

根据以上研究可以认为,加泰罗尼亚语中句长为3—11的句子中的依存距离分布适合用SED进行刻画,而其他句长的句子则更适合用TPLD进行描述。

四、结 论

本文利用大规模多语种依存树库对30种语言的依存距离分布进行了详细分析。我们提出了利用不同模型拟合对比来探讨依存距离分布的方法和路径,并较合理地回答了引言中所提出的几个问题:(1)从依存距离累积分布趋势上,人类语言依存距离分布模式具有一定的规律性;(2)多数语言符合一定的分布模型,SED和TPLD分别较适合拟合“短句”与“长句”的依存距离分布;(3)尽管不同语言依存距离分布规律表现出一些特性,但在本质上,其分布都是一种指数和幂律混合的分布模型;(4)当前研究中出现两种分布模型分歧的主要原因可能是拟合方法、不同语言、句长和文本量等方面的影响。

本文的研究有助于我们更深刻地理解依存距离最小化的本质,并进一步揭示人类语言的依存距离可能遵循一种普遍性的分布模式。语言中依存距离倾向于最小化,表明在实际的构句过程中,依存相关的支配词和从属词之间不能离得太远,它们更常相邻地在句子中依次出现[6,12,18]。在短句中,类幂律行为表现并不十分显著;但随着句长增大,依存关系的组配逐渐复杂,依存距离分布也渐渐表现为幂律行为,这正反映了Zipf提出的“省力原则”在依存句法关系层面所起的作用。这种原则约束长依存关系的出现,增加短依存的比例,在会话双方的博弈下,最终使依存距离的分布达到一种平衡。在数学上,这种平衡可以用指数和幂律混合的函数来表达。

作为依存句法计量研究的重要内容,依存距离分布研究也是构建基于依存语法的句法协同子系统不可或缺的一部分[5,29],本文在这方面做了一些有益的探索。但是,在HamleDT 2.0依存树库中,一些语言所包含的句子数量较少,导致我们难以判断这些语言依存距离分布拟合结果的准确性。此外,本文采用的幂律分布拟合方法是当前该领域的常用方法,与传统方法(基于最小二乘法)

相比，能够更加精确地拟合幂律分布模型[24]，但对幂律行为的判断和拟合还需要考察与借鉴更多的数据拟合与检验方法。因此，未来有必要基于包含更多语言的多维均衡的依存树库，进一步验证相关方法和结论的准确性与可靠性，更科学地探究依存距离分布规律与人类认知、语言类型之间的关系。

［参　考　文　献］

［1］N. Chomsky, *Lectures on Government and Binding: The Pisa Lectures*, Berlin/New York: Walter de Gruyter, 1993.

［2］R. Hudson, *Word Grammar*, Oxford: Blackwell, 1984.

［3］I. A. Mel'čuk, *Dependency Syntax: Theory and Practice*, Albany: State University Press of New York, 1988.

［4］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）》2011 年第 2 期，第 169－180 页。［Liu Haitao, "Linguistic Networks: Metaphor or Tool?" *Journal of Zhejiang University (Humanities and Social Sciences)*, No. 2(2011), pp. 169－180.］

［5］刘海涛、黄伟：《计量语言学的现状、理论与方法》，《浙江大学学报（人文社会科学版）》2012 年第 2 期，第 178－192页。［Liu Haitao & Huang Wei, "Quantitative Linguistic: State of the Art, Theories and Methods," *Journal of Zhejiang University (Humanities and Social Sciences)*, No. 2(2012), pp. 178－192.］

［6］H. Liu, "Dependency Distance as a Metric of Language Comprehension Difficulty," *Journal of Cognitive Science*, Vol. 9, No. 2(2008), pp. 159－191.

［7］R. Ferrer-i-Cancho, "Euclidean Distance between Syntactically Linked Words," *Physical Review E*, Vol. 70, No. 5(2004), 056135.

［8］梁君英、刘海涛：《语言学的交叉学科研究：语言普遍性、人类认知、大数据》，《浙江大学学报（人文社会科学版）》2016 年第 1 期，第 108－118 页。［Liang Junying & Liu Haitao, "Interdisciplinary Studies of Linguistics: Language Universals, Human Cognition and Big-data Analysis," *Journal of Zhejiang University (Humanities and Social Sciences)*, No. 1(2016), pp. 108－118.］

［9］R. Ferrer-i-Cancho, "Why Do Syntactic Links Not Cross?" *EPL (Europhysics Letters)*, Vol. 76, No. 6(2006), pp. 1228－1234.

［10］R. Hudson, "Measuring Syntactic Difficulty," 2008－07－06, http://www.phon.ucl.ac.uk/hone/dick/diffculty.htm, 2015－12－17.

［11］H. J. Heringer, B. Strecke & R. Wimmer, *Syntax: Fragen, Lösungen, Alternativen*, München: Wilhelm Fink Verlag, 1980.

［12］R. Ferrer-i-Cancho, "A Stronger Null Hypothesis for Crossing Dependencies," *EPL (Europhysics Letters)*, Vol. 108, No. 5(2014), pp. 58003－58008.

［13］Q. Lu, C. Xu & H. Liu, "Can Chunking Reduce Syntactic Complexity of Natural Languages?" *Complexity*, DOI: 10.1002/cplx.21779, 2016－03－27, https://www.researchgate.net/publication/299459663_Can_chunking_reduce_syntactic_complexity_of_natural_languages, 2016－04－15.

［14］G. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, New York: Hafner, 1949.

［15］R. Ferrer-i-Cancho & R. V. Solé, "Least Effort and the Origins of Scaling in Human Language," *Proceedings of the National Academy of Sciences*, Vol. 100, No. 3(2003), pp. 788－791.

［16］A. L. Barabasi & R. Albert, "Emergence of Scaling in Random Networks," *Science*, Vol. 286, No. 5439(1999), pp. 509－512.

［17］H. Liu, "Probability Distribution of Dependency Distance," *Glottometrics*, Vol. 15(2007), pp. 1－12.

[18] J. Jiang & H. Liu,"The Effects of Sentence Length on Dependency Distance, Dependency Direction and the Implications-based on a Parallel English-Chinese Dependency Treebank," *Language Sciences*, Vol. 50(2015), pp. 93 - 104.

[19] R. Ferrer-i-Cancho & H. Liu," The Risks of Mixing Dependency Lengths from Sequences of Different Length," *Glottotheory*, Vol. 5, No. 2(2014), pp. 143 - 155.

[20] 刘海涛:《依存语法的理论与实践》,北京:科学出版社,2009年。[Liu Haitao, *Dependency Grammar: From Theory to Practice*, Beijing: Science Press, 2009.]

[21] M. E. J. Newman,"Power Laws, Pareto Distributions and Zipf's Law," *Contemporary Physics*, Vol. 46, No. 5(2005), pp. 323 - 351.

[22] J. Alstott, E. Bullmore & D. Plenz," Powerlaw: A Python Package for Analysis of Heavy-tailed Distributions," *PLoS One*, Vol. 9, No. 1(2014), e85777.

[23] L. A. Adamic & B. A. Huberman," Zipf's Law and the Internet," *Glottometrics*, Vol. 3, No. 1(2002), pp. 143 - 150.

[24] A. Clauset, C. R. Shalizi & M. E. J. Newman,"Power-law Distributions in Empirical Data," *SIAM Review*, Vol. 51, No. 4(2009), pp. 661 - 703.

[25] D. Zeman, O. Dušek & D. Mareček, et al. ,"HamleDT: Harmonized Multi-language Dependency Treebank," *Language Resources and Evaluation*, Vol. 48, No. 4(2014), pp. 601 - 637.

[26] M.-C. de Marneffe, T. Dozat & N. Silveira et al. ,"Universal Stanford Dependencies: A Cross-linguistic Typology," in N. Calzolari, K. Choukri & T. Declerck et al. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Paris: European Language Resources Association, 2014, pp. 4585 - 4592.

[27] C. A. Hidalgo,"Conditions for the Emergence of Scaling in the Inter-event Time of Uncorrelated and Seasonal Systems," *Physica A: Statistical Mechanics and Its Applications*, Vol. 369, No. 2(2006), pp. 877 - 883.

[28] 刘海涛:《基于依存树库的汉语句法计量研究》,《长江学术》2008年第3期,第120 - 128页。[Liu Haitao,"A Quantitative Study of Chinese Structures Based on Dependency Treebanks," *Yangtze River Academic*, No. 3(2008), pp. 120 - 128.]

[29] R. Köhler, *Quantitative Syntax Analysis*, Berlin/New York: de Gruyter Mouton, 2012.

李军司长调研《浙江大学学报(人文社会科学版)》

2016年6月1日,国家新闻出版广电总局新闻报刊司司长在浙江期间,专程到《浙江大学学报(人文社会科学版)》编辑部进行了调研。李司长在调研中充分肯定了《浙江大学学报(人文社会科学版)》近年来的办刊业绩,尤其在期刊精品化、数字化和提升期刊国际影响力等方面所做的探索,希望浙大社科学报能保持目前的发展势头。

浙江省新闻出版广电局副局长华宣飞、报刊处施丽琴处长,浙江大学出版社总编袁亚春、《浙江大学学报(人文社会科学版)》执行主编徐枫等有关单位和部门的负责人参加了本次调研活动。