现代汉语历时变化特征研究:结构复杂度

秦洪武　周　霞　孔　蕾

【摘　　要】本文基于现代汉语历时语料库,通过历时比较描述现代汉语结构复杂度百年演化特征,观察复杂度变化与句段长度的关系,并从类联接和依存距离两个可计量维度考察结构复杂度变化的内部成因。研究发现:(1)汉语结构复杂度在过去百年总体呈上升趋势,快速变化多出现在1920-1940年代;(2)结构复杂度表现出文体差异:小说文本历时变化幅度微小,学术文本变化幅度最大;(3)句段长度与结构复杂度变化轨迹相近,前者能够较好地预测后者;(4)类联接历时分析显示,部分含有副词的类联接构成的句段复杂度低,多在文学作品中用作话语标记;而含有助词"的"的类联接多用作修饰成分用于构建复杂的结构,多出现在学术文本中;(5)以依存距离表达的结构复杂度呈现历时增长,增幅最大的依旧是学术文本。研究认为,依据口语、书面语二分法来比较现代汉语历时变化具有合理性,但还需系统反映书面语内部体裁间的差异,本研究将这种差异表达为历时变化幅度层级:学术文本>通用文本>新闻文本>小说文本。

【关 键 词】现代汉语;历时语料库;结构复杂度;句段长度

【作者简介】秦洪武,曲阜师范大学外国语学院教授,博士,研究方向:语料库语言学,语言对比与翻译,文化海外传播,E-mail:qinhongwu@163.com;周霞,曲阜师范大学文学院博士生,研究方向:对比语言学;孔蕾,曲阜师范大学外国语学院教授,博士,研究方向:语言对比与翻译,文化海外传播,外语教学(山东　曲阜　273165)。

【原文出处】《外语与外语教学》(大连),2022.4.87~98

【基金项目】本文系国家社科基金项目"基于事件语义的语用标记词汇化跨语言对比研究"(项目编号:17BYY038)的阶段性研究成果。

1. 引言

现代汉语的发展与成形始于五四时期"躲文言,亲口语"(张中行2007:278)的文学革命,在1930年前后达到成熟(张中行2007:280)。一般认为,现代汉语历时变化是语法变化,主要涉及词法和句法两个层面(朱德熙1982:25),包括词汇变化(新词、动词"要""是"、副词"其实""十分"、介词"给""到"等的语法化)、结构变化(动补结构、比较句以及汉语特殊句式等)和语法系统演变(汉语的否定系统、方位系统、助词及体标记等)(如崔山佳2013;方梅2005;沈家煊1998;石毓智、李讷2000;张谊生2000)。汉语学界也尝试建立反映汉语语言特征的语法化理论,探索汉语语法化机制及演变规律(如刘丹青2001;沈家煊1998;吴福祥2004)。另外,翻译语言对现代汉语的影响(欧化)(王力2011:335;张中行2007:285)已成学界共识,已有不少研究关注翻译对现当代汉语的影响(如王克非、秦洪武2017;夏云、秦洪武2017;朱一凡2018),但主要聚焦五四时期(1920年前后)至20世纪30年代的文本,现代汉语在20世纪30年代以后的变化较少进入研究视野。

长期以来,汉语演化研究多注重观察特定语言项使用的变化,极少关注语言演化的系统性,缺少对汉语演化特征的整体描述。此外,汉语历时演变研究一般考察记叙文本,体裁单一,范围受限,缺少文体和体裁平衡的历时语料支撑,无法多层次反映现代汉语变化的样貌。鉴于此,本研究使用历时平衡语料,尝试基于词性和

句法标注从多个侧面描述现代汉语结构复杂度的历时变化。

2. 结构复杂度和相关概念

本研究基于历时平衡语料库，采用描写性研究方法，探索现代汉语结构复杂度历时变化特征，选择影响汉语结构复杂度的词性和句法结构设置主要变量。除了采用秦洪武和孔蕾(2018)提出的汉语结构复杂度计算方法描述现代汉语的变化趋势，本研究还把与汉语结构复杂度密切相关的句段长度单独列为观察对象，观察句段长度对汉语结构复杂度的预测力。为进一步揭示结构复杂度变化的内在原因，研究还观察和分析了两种影响复杂度的内部构成因素：类联接型式和依存距离。此外，学界在观察现代汉语历时变化时，传统上只作口笔语区分，书面语内部可能存在的差异长期未得到重视。为弥补这一不足，本研究基于体裁分类观察结构复杂度变化，尝试对汉语历时变化作出更精细的刻画。

2.1 汉语结构复杂度

句法复杂程度(syntactic complexity)与心理学上的认知复杂性有关联(Givón 1998：10)。同等情况下，句子越长结构越复杂，读者付出的处理努力就可能越大，可读性因之越低。根据Kintsch(1998)的分析，词汇数量是反映概念复杂度的指数，句子越长，就越有可能嵌入多重离散的概念/命题。表现为字符串长度的句子长度在语言表层最直观，也最容易计数。同时，容量本身也是可读性的重要参数。一般说来，字符串越长，占用的工作记忆就越多，语言使用者所付出的处理努力也就越多。本研究认为，除了词汇数量，句法复杂度还受其他因素影响，如特定结构内部的构成。对结构封闭的汉语而言，若结构内部修饰成分多或内嵌式成分、小句性修饰语混合使用，结构也会变得复杂，处理难度增大，对可读性或影响。鉴于此，本研究在计量结构复杂度时兼顾词的数量和结构内部构成。由于影响结构复杂度的因素众多，本文只关注四个主要因素：结构封闭性、句段长度、类联接和依存距离(封闭结构的容量)。

2.2 相关概念

2.2.1 汉语结构的封闭性

汉语缺少修饰、限定成分后置的语法手段，基本结构不允许向右扩展，属于结构封闭的语言。“介词-方位词”“数量短语-名词”“动词-宾语”“指示代词-名词”“形容词-名词”等结构右边的词为结构终点，修饰成分只能置于被修饰成分之前，导致汉语结构对容量敏感，结构容纳的词汇数量有限(秦洪武2010)。已有研究发现，结构容量内部构成复杂，中心词之前修饰成分过多是导致汉语结构复杂的主要原因(秦洪武2010；秦洪武、孔蕾2018；王克非、秦洪武2017)。通常情况下，封闭结构内部的前置修饰成分短小且数量少；使用多个修饰成分时，一般会使用平行结构以降低处理难度；若修饰成分多，结构不平行，甚至混有小句式定语，则会破坏读者期待的阅读节奏，迫使读者采用基于记忆的理解策略付出更多认知努力(秦洪武、孔蕾2018)。汉语结构的封闭性使结构内部容量和构成的计量成为可能，基于词数、各类小句数量以及封闭结构从属小句数量、内嵌小句数量和修饰成分数量可综合计算出汉语的句法复杂度(秦洪武、孔蕾2018)。

2.2.2 句段长度

在汉语里，句段是一种独特的存在。它包含流水句式(连谓句)，也包含其他句子片段；它可以是单个句子，也可以是短语，更像是赵元任和沈家煊所说的“零句”(沈家煊2012)。汉语句段切分且用于计数的标志除句子

标记外还有逗号和冒号。句段长度不等于结构复杂度，但二者紧密关联。首先，句段之间的显性连接是结构复杂度的指标(逻辑连接词)；其次，封闭结构包含在句段之中，结构复杂度和句段长度必然关联：句段长度相同时，句段内含的封闭结构容量越大复杂度越高，封闭结构中内嵌小句或非平行表达形式越多复杂度越高。修饰成分、内嵌成分或小句性修饰语混合使用会产生大容量结构和内部复杂的长句段，导致认知处理困难。这说明，句段长度可以作为判断结构复杂度的重要关联变量。鉴于此，本研究在描述汉语结构复杂度历时变化时，将句段长度一并纳入考察范围。

2.2.3 类联接

结构复杂度成因复杂，需要综合考察多种变量方能对其进行有意义的分析。汉语封闭结构对容量敏感，仅凭词数难以真正反映结构的复杂度，需考察结构内部构成以达成合理判断。鉴于词性间的线性关系可产生有规律的句法组合，可以从词性组合方式上(类联接)分析汉语结构复杂度变化的原因。学界对类联接认识不同，可以指词与词在语法范畴之间的关系(Firth 1957：181)，即在词汇类别上的组合关系，也可指词形或句法上促使语言因素之间相互结合的条件(Bussmann 2000：81)。本研究中汉语语法层面的类联接仅指词类间的线性组合。类联接的历时变化，以及特定类联接与结构复杂度之间的关系，是本研究考察和分析的重点内容。

2.2.4 依存距离：封闭结构容量

汉语封闭结构的容量可以通过依存距离进行测量，这种句法关系的描述与结构复杂度直接相关(Ai & Lu 2013；Ortega 2003；Pallotti 2015)。依存句法分析表达句子成分间的依存关系，而基于依存关系可测量成分间的依存距离，并根据依存距离判断结构的复杂程度。一般来说，依存距离越长，结构越复杂(Lin 1996)。汉语封闭结构在依存语法中表达为结构起始词和结束词之间的依存关系，依存关系表达为以词计数的依存距离，依存距离越长容量越大，容量越大结构越复杂。鉴于此，本研究将依存距离作为观察结构复杂度内部构成的另一维度。限于篇幅，本文仅以动宾结构为例考察以依存距离表达的汉语结构容量的历时变化。

3. 研究方法和数据

本研究采用探索性方法，基于语料库标注信息，通过频率比较考察汉语的历时变化特征，发掘可能存在的变化规律。使用的数据和分析工具如下。

3.1 数据

一般来说，历时样本间的时间间隔越短越有助于揭示变化细节(秦洪武、王克非 2014)。鉴于此，本研究以五年为间隔收集语料样本。考虑到语言特征存在文体和体裁差异，本研究使用平衡语料，包括四个子库(见表 1)。研究建设了“1920-2020 百年汉语历时语料库”，该库基于 Brown/LOB 语料库抽样框架，参照 LCMC 的体裁分类[①]和 Biber(1993)对语料样本代表性的基本观点[②]收集 1920 至 2020 年间的汉语语料。每一时间点都按照 Brown/LOB 语料库抽样框架抽取语料，以保证语料的代表性和历时一致性。时间设置上，自 1920 年起每隔 5 年收集一次语料，共设置 21 个时间点(允许在节点年之前或之后一年收集语料)。由于历时语料时间点多，抽样难度大，研究按照 LCMC(100 万词)10%的比例收集语料。每一时点的库容如下(表 1)。

表1 “1920–2020”百年汉语历时语料库

序号	时点(年)	媒体子库(字)	小说子库(字)	通用子库(字)	学术子库(字)	库容(字)
1	1920	28,932	57,682	48,111	25,516	160,241
2	1925	28,872	57,669	48,094	25,547	160,182
3	1930	28,796	57,704	47,992	25,479	159,971
4	1935	28,808	57,744	48,073	25,599	160,224
5	1940	28,873	57,592	48,052	25,488	160,005
6	1945	28,794	57,582	48,107	25,521	160,004
7	1950	28,890	57,674	48,091	25,582	160,237
8	1955	28,895	57,737	48,246	25,541	160,419
9	1960	28,907	57,716	48,072	25,609	160,304
10	1965	28,809	57,615	47,913	25,556	159,893
11	1970	28,861	57,699	46,976	25,600	159,136
12	1975	28,883	57,738	47,958	25,613	160,192
13	1980	28,792	57,671	48,091	25,519	169,973
14	1985	28,912	57,680	48,139	25,590	160,321
15	1990	28,872	57,723	48,097	25,692	160,384
16	1995	28,888	57,714	47,989	25,604	160,195
17	2000	28,848	57,667	48,117	25,613	160,245
18	2005	28,763	57,681	48,066	25,558	160,068
19	2010	28,874	58,371	48,099	25,599	160,948
20	2015	28,739	57,628	47,996	25,596	159,959
21	2020	28,691	57,778	47,973	25,542	159,984
总计(字)		605,699	1,212,065	1,008,252	536,964	3,372,885

3.2 数据提取

语料标注方式多样,常见的是词性标注和句法标注。相比而言,词性标注准确率较高,其主要功能是允许按类检索和提取语言使用信息。根据Wynne(2008:716),检索标注符有助于发现语料库中的语法型式(grammatical patterns),也可用于发现与特定词有共现倾向的语法型式,而且多维度语域分析中的变量选取大都以词性为线索(Biber & Conrad 2019: 65)。考察词性间的线性关系可以捕捉有规律的句法组合,是语法(短语结构或句法结构)描写的基础。本文考察的结构复杂度将功能词尤其是结构助词和动态助词列入考察重点,对与之关联的词性和词性组合描述和分析。此外,句法标注可直接反映句子成分间的关系,本文基于依存句法标注描述、分析成分间的依存距离。词性标注采用汉语语言处理工具包HanLP1.3.2,依存句法标注采用Stanford-corenlp-4.1.0。

3.2.1 结构复杂度提取和统计

根据秦洪武和孔蕾(2018)以及夏云和秦洪武(2017)的研究,影响现代汉语结构复杂程度的因素可概括为:封闭结构内部的词数(Wr,计数词)、用于计算句段长度的标点符号数量(Cl,除顿号、书名号、引号外的所有标点符号)、从属小句数量(Su,计数从属连词)、内嵌小句数量(Eb,计数动词、“所”“着”“了”“过”)[③]以及修饰成分数量(Nmd,计数“的”“之”)。从属小句数量、内嵌小句数量直接影响可读性,而词数并不直接对可读性产生影响。鉴于此,需要对导致句法复杂的因素赋予较大权重,即词的权重小于修饰成分,修饰成分权重小于小句。本研

究使用秦洪武和孔蕾(2018)的复杂度计算公式,根据词性和小句权重计算复杂度,采用线性综合评价模型得到结构复杂度,使用的计算公式如下:

复杂度=((NWr*2)+(NEb*40)+(NSu*40)+(NCl*13)+(NMd*5))/10+NWr/NCl*0.4

该计算公式计量总的句段数、句段内含的内嵌成分或从句数量(NEb),以及结构内部词数(NSu)和小句总数(NCl)之比。采用Excel的计数和运算函数完成计数和加权统计,直接得到复杂度。

3.2.2 类联接提取和统计

研究从4类文本84个样本中提取类联接数据,也就是说每类文本有21个时点的标准化历时类联接频率数据。分析时,先考虑在所有样本中出现的类联接数据,然后将这些数据进行主成分分析,得到降维后的类联接信息。分析时使用R中的Stylo工具包获得频率数据,为获得较好的可视化效果,采用R自带的分析工具实施主成分分析。

3.2.3封闭结构距离提取

研究基于依存距离计量汉语封闭结构的容量,即以词数计量结构起止词之间的"距离"。在4类文本21个时点84个样本中,根据句法关系标记使用正则表达式提取容量数据,以此描述特定封闭结构依存距离的历时变化。

4. 主要发现

以下是基于"1920-2020百年汉语历时语料库"分子库描述和分析的结果。首先是结构复杂度宏观变化特征的描述,然后是句段长度变化分析。此外,为考察结构容量和复杂度,还进行了体现结构复杂度的类联接历时聚类特征分析,并以动宾结构为例,观察封闭结构容量(基于依存距离计量)在各子库中的历时变化以及子库间的历时变化差异。

4.1 结构复杂度历时变化

结构复杂度历时变化分四个子库(通用文本、学术文本、小说文本和媒体文本)进行描述。研究发现(见图1),变化最为明显的是学术子库,媒体和通用子库次之,而小说子库未表现出明显变化。文学文本与非文学文本在结构复杂度变化上形成鲜明对比,可能的原因是,文学文本更接近口语,而口语受翻译语言影响较小,更可能体现语言演化的惰性(Inertial Theory of language change, Longobardi 2001)[④],为检验这一推断,我们将在4.1.1节作出进一步分析。

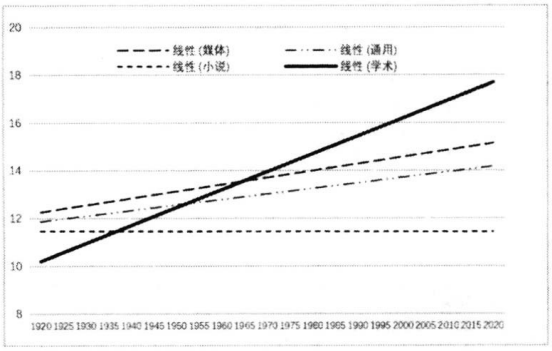

图1 各子库结构复杂度的历时变化

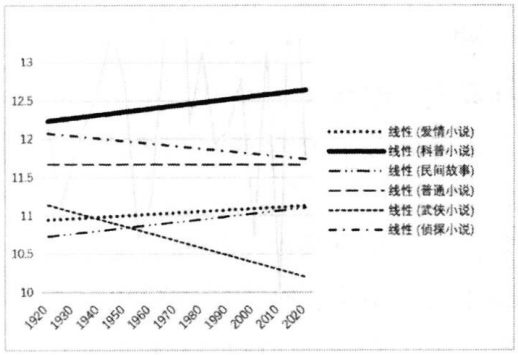

图2 小说文本结构复杂度历时变化

不同子库文本在结构复杂度上呈现不同历时变化特征,那么不同子库文本内部究竟发生了怎样的变化?

为回答这一问题,下文按体裁分类对汉语结构复杂度变化进行历时描述。

4.1.1 小说文本结构复杂度历时变化

根据上文描述,小说子库文本的结构复杂度变化微小。为进一步观察小说子库内部变化特征,我们分析各类小说文本结构复杂度的历时变化。需要指出的是,小说文本属于文学文本,相较于信息性文本,文学文本主体为叙述,风格更接近口语。这说明,口语在句法变化上的惰性可能在文学文本中反映出来。如果这一解释成立,文学子库中的各类体裁也应体现这一惰性特征。相关分析如下。

图2显示,普通小说、爱情小说和民间故事未显示出明显的变化趋势,侦探小说和武侠小说的结构复杂度甚至呈下降趋势,唯一例外的是科普小说,其结构复杂度出现小幅历时增长。可能的原因是,科普小说兼具文学语言和非文学语言的特征,其非文学语言的那部分特征将其与其他类别小说区分开来,体现为微弱的结构复杂度的增长。

总的来看,在小说子库文本内部,结构复杂度仅在个别体裁中有微幅增长,总体未表现出显著历时差异。小说子库分析显示,只有科普小说在结构复杂度上呈现增长趋势,而口语特征相对明显的其他文学体裁中结构复杂度变化幅度相对较小。这也从另一角度说明,不能单纯基于文学和非文学分类观察汉语变化,至少需要在体裁层面作出细粒度的描述。

4.1.2 信息性文本结构复杂度历时变化

信息性文本包括学术、新闻、通用子库,三个子库包含8种体裁,以下是结构复杂度在各体裁中的历时变化情况。

学术文本结构复杂度的历时变化。学术子库由科技文本构成,图3显示,科技文本结构复杂度明显增长,且表现出阶段性:1920–1930年急速增长,1960–1970年较快增长,1975年后复杂度变化趋于稳定(图3)。

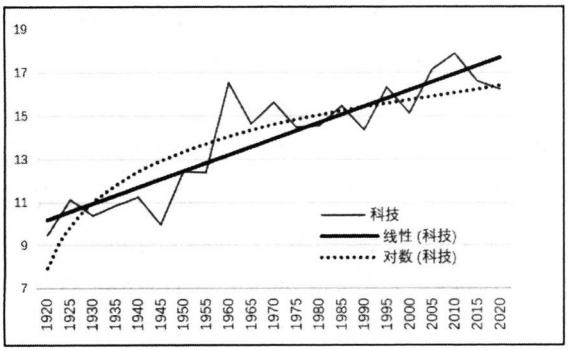

图3 学术文本结构复杂度历时变化

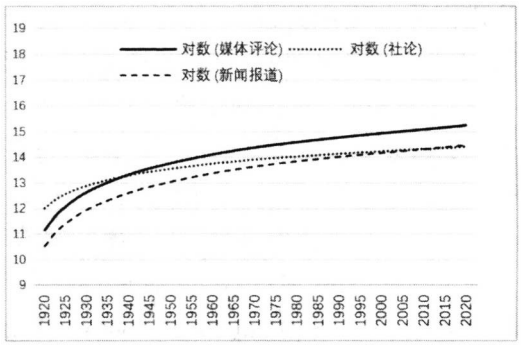

图4 媒体子库结构复杂度历时变化

新闻文本结构复杂度历时变化。媒体子库由社论、新闻报道和媒体评论三部分构成,对媒体子库各体裁的分析显示,三类媒体文本结构复杂度总体趋势是上升(图4)。对数趋势线显示,三类文本的结构复杂度基本都是在1945年前快速增加,1945年后增长趋于平稳,媒体评论类文本结构复杂度变化最明显。

通用文本结构复杂度历时变化。通用子库包含技能贸易风俗、书信演讲传记、杂类和宗教文本。图5显示,与科技、媒体子库一样,通用文本在结构复杂度上呈现历时增长趋势,其中杂类和技能贸易风俗类文本的变化最明显,显著变化均发生在1935年之前;相比而言,宗教文本变化幅度最小,变化趋势不明显。

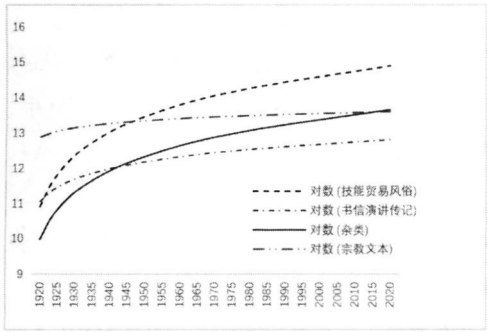

图5 通用文本结构复杂度历时变化

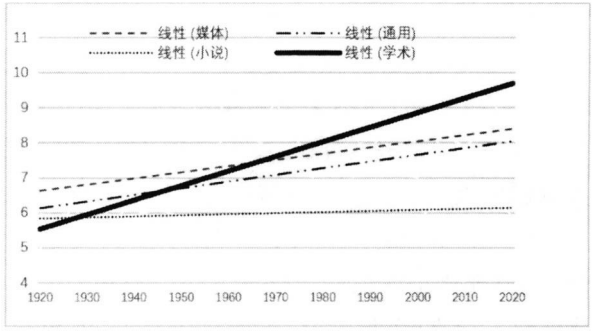

图6 各子库平均句段长度历时变化

综上分析,整体看来,1920–1940年代为现代汉语结构复杂度快速增长期,1940年代后增长趋于平稳,但不同子库、不同体裁的表现不同:小说语言历时变化幅度小,而信息性文本尤其是学术文本变化最大,其中科技语言历时变化幅度最大,新闻报道和技能贸易风俗类次之。

4.2 句段长度历时变化

句段容量与汉语结构复杂度直接相关。句段越长,可能容纳的句法成分就越多,句子就越复杂(秦洪武、孔蕾2018)。这提示我们,结构复杂度的历时变化或许可以借助句段长度这一更易测量的变量进行预测。鉴于此,本节从平均句段长度入手,观察句段长度对汉语结构复杂度的预测力。

图6显示,各子库中平均句段长增长最大的是学术文本,其次为媒体和通用文本,而小说文本几乎没有出现明显历时变化。学术文本句段长度的增长幅度远大于其他类文本。对比图1发现,除文学外,其余三个子库在平均句段长和结构复杂度上均表现出一致的历时变化倾向,这说明结构复杂度在很大程度上可以通过平均句段长反映出来。

对小说子库内部各体裁文本平均句段长变化的分析同样显示(图7),只有科普小说出现明显历时增大趋势,这种变化相对于其他体裁(如学术、媒体等)幅度微小,但也从另一角度证明科普小说兼具文学语言和非文学语言的特征。

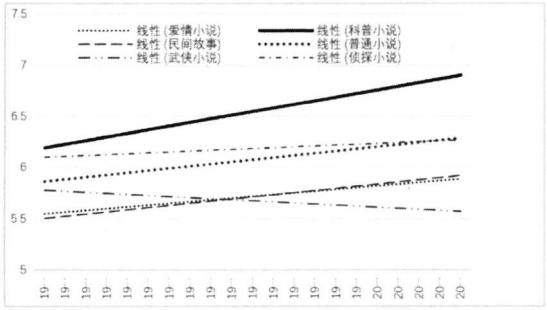

图7 小说文本句段长度历时变化

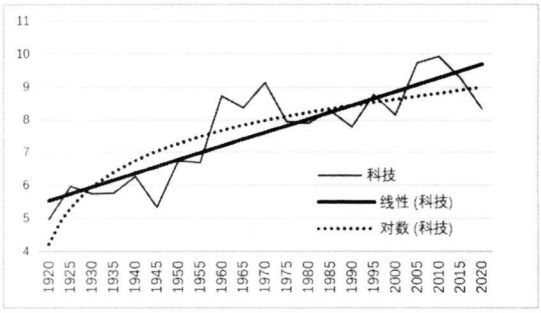

图8 学术文本句段长度历时变化

综上所述,在小说子库文本内部,结构复杂度和平均句段长两个变量仅在个别体裁中有微幅增长,总体未表现出显著历时差异。小说子库分类描述显示,只有科普小说在结构复杂度和句段长度上呈现增长趋势,而在口语特征相对明显的其他文学体裁中两个变量的变化幅度相对较小。这进一步说明,在体裁层面作出细粒度描述非常必要。

学术文本在平均句段长度上增长明显(图8),由1920年的4.96增加到2020年的8.34。图8对数线显示,学术文本句段长度历时增长经历了三个阶段:1920–1930年,1960–1970年,1975年后。这三个阶段与图3显示的学术文本结构复杂度三阶段历时变化趋势近乎一致,可以据此推测句段长度对结构复杂度有预测力。

图9显示,媒体子库文本句段长度的变化在1945年前增速最为明显,1945年后稳步提升,这与媒体子库结构复杂度变化趋势相似。媒体子库在1920–1945年间表现出的快速变化与这一时期借助"欧化"改造汉语有关,但却不像学术文本那样完全同步。1920年至文白论战(20世纪30年代)之前"欧化"之风盛行于知识界,长定语、多修饰语成为行文严谨、精确的主要实现方式,汉语结构容量因之在短时间内快速扩增,句段长度随之快速增大。而在20世纪30年代(1930–1940年)的文白论战和文白论战之后,媒体语言结构复杂度和句段长度的增幅明显减小,极端欧化式微,汉语媒体语言结构复杂度进入平稳发展时期。

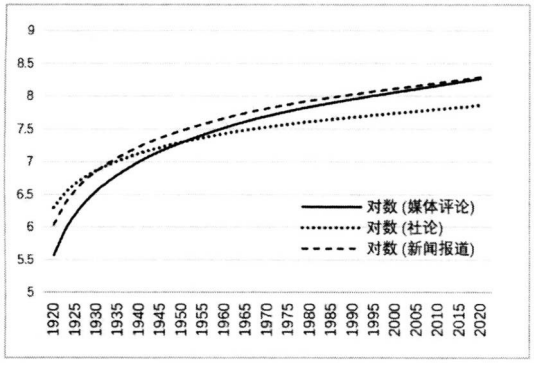

图9 媒体子库句段长度历时变化

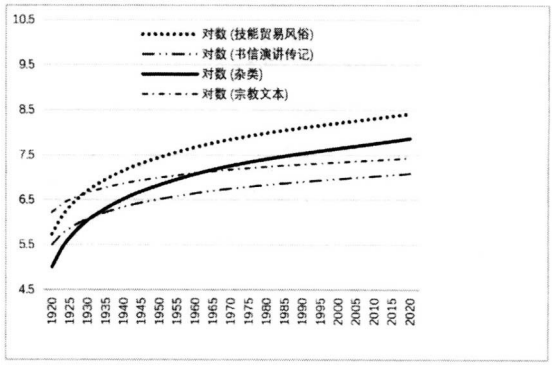

图10 通用文本句段长度历时变化

图10显示,与科技、媒体子库一样,通用文本在句段长度上呈现历时增长趋势,其中杂类和技能贸易风俗类文本的变化最明显,显著变化都发生在1935年之前;与之相比,宗教文本变化幅度最小,变化趋势不明显。通用文本句段长度与结构复杂度历时变化趋势图5一致。

总的来看,各子库文本在句段长度和结构复杂度历时变化上表现出近乎平行的变化特征,说明句段长度与结构复杂度的关联不仅在理论上成立,在统计上也成立。前者能够预测后者,但预测能力强弱还有待进一步研究。

4.3 类联接历时变化

词与词在语法层面的关系被称为类联接(colligation)(McEnery et al. 2006)。一般认为,类联接就是语法类别或者范畴的组合形式,也可以扩展到一个词和不同语法类别或范畴间的关系。历时语料中,类联接可能带有某些时代或体裁特征。我们采用主成分分析方法,探索类联接与各类文本可能存在的历时关联。为考察可能存在的句法关系,研究采用3项目序列(3-gram)获取词性符串。3-gram词性串分析显示(图11)[⑤],某些词性组合在当代(20世纪80年代以后)突出,主要出现在非文学文本之中,非文学文本使用的组合方式较文学类文本变化大。

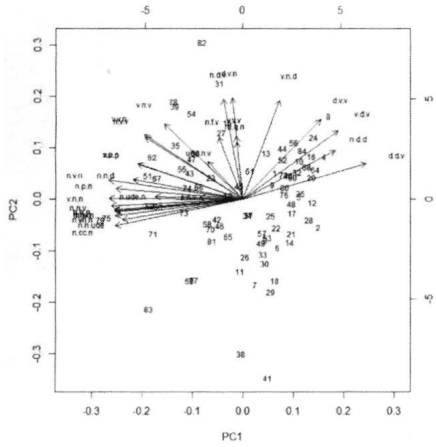

图11 类联接主成分分析表达的历时变化

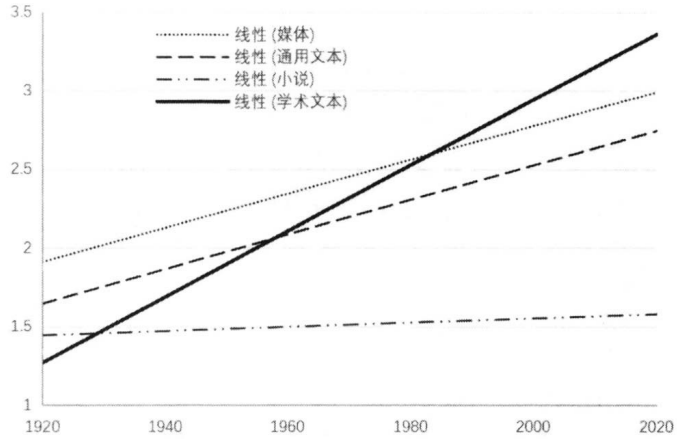

图12 dobj(动词–宾语)依存距离历时变化

图11显示，包含副词(d)的类联接变量(d v v、v d v、d d v)与主成分正相关且贡献大，如“日益/d发展/vn壮大/v、应/v立即/d振作/v、便/d大声/d喊/v”，这类表达形式多为独立使用的句段，无内嵌结构，复杂度低。从样本代码看，上述类联接多数出现在文学样本，常用于叙述和交际场景，与文学的叙事性和口语化有关，是文学语言接近口语的直接证据。从样本代码所处的位置看，文学样本处于主成分一(PC1)的相关性高值(如编号4、8、12、16的样本)和低值(如编号84、72、68、60、56、52的样本)没有表现出由小到大的历时变化分布特征。由此可以判断，至少从类联结上看，文学文本的结构复杂度没有表现出明显渐增或渐低的历时变化趋势。

观察图11中的主成分二(PC2)，贡献大的类联接变量为n ude n、n n ude、n vn n、n n n、n p n、v v n，靠近相关系数线的样本点(表达为数字)主要是“非文学学术文本”。提取到的词性串实例显示，这些类联接主要为修饰语或包含修饰成分的语言片段(如“劳动者/n的/ude1工资/n”“中产阶级/n色彩/n的/ude1”“结构性/n改革/vn力度/n”“先锋/n模范/n作用/n”)，以及意义不完整的表述(如“市场/n为/p导向/n”“继续/v发挥/v领导/n”)。以上两种类联接表达的语义或结构不完整，多用于构建更长、结构更复杂的句段，不像主成分一中的类联接那样构成语义或语用相对完整的句段。从样本代码所处的位置看，非文学样本尤其是学术文本的代码随箭头方向呈现有规律的增大趋势，说明非文学文本尤其是学术文本使用结构复杂类联接的频率呈历时增长趋势。

图11展示的类联接在文本类型和时间上的变化说明：类联接与文类和体裁关联，在一定程度上反映出口笔语之间的差异，说明体裁是考察句法特征的重要维度；从时间上看，非文学尤其是学术文本和文学文本表现出不同的变化。文学文本类联接变量未呈现有规律的历时变化趋势，学术文本则呈现出明显的历时变化趋势。此外，非文学通用和非文学新闻文本处于学术文本和文学文本之间没有表达出变化方向，因而没有明显的历时变化趋势。

4.4 封闭结构(“动词–宾语”结构)容量的历时变化

鉴于封闭结构直接关联结构复杂度和句段长度，本文使用依存距离观察汉语封闭结构内部容量的历时变化。研究从汉语句法标注库中提取动词–宾语(dobj)结构125671个，将提取到的结构按照子库分类(4类)，从21个时点获得4x21个平均距离值。图12显示的是动宾关系在依存距离(结构容量)上的历时变化。

根据图12，动词–宾语(dobj)间依存距离表达的结构容量在各子库均呈现历时增长，说明近百年来这个高

度活跃的汉语结构在容量上不断扩增。其中，增长幅度较大的是学术文本和通用子文本，幅度极小的依旧是小说文本。学术文本的结构容量变化最明显，1950年之前，汉语信息性文本动宾结构中动词和宾语的平均依存距离不到2词，1970年以后结构平均依存距离大于4词。如例(1)动词“呈现”和宾语“性质”之间的依存距离为6词，例(2)动词“具有”和宾语“功能”之间的依存距离则达到了35词。

(1)……很多材料呈现出奇特而有用的性质。[dobj(呈现-22，性质-28)](选自科技文本，《中国纳米科研现状分析与思考》)

(2)……具有结合并控制转录起始前复合物的装配、定位转录起始位点并控制转录的方向、对细胞内临近或远处的激活子或抑制子做出响应的功能，……[dobj(具有-22，功能-57)](选自科技文本，《维管组织定位表达启动子研究》)

需要说明，例(1)和例(2)的差异不仅表现为依存距离(6词和32词)，还体现于内部构成。从内部构成看，构成例(2)长依存距离的是以句子形式出现的修饰成分，且成分内部还包含较长的并列成分。依存距离长意味着处理动词后不能及时处理宾语，结构复杂意味着需要花费更多努力处理宾语前的语义关系，这增大了工作记忆负荷和处理难度，是封闭结构内部构成复杂导致结构复杂的典型例证。

对动词-宾语(dobj)结构的分析说明，依存距离能扩展封闭结构的观察范围，便于从内部构成角度揭示汉语结构复杂度变化的成因。

5. 讨论和结语

在过去的百年里，现代汉语结构复杂度总体呈增长趋势，增长幅度存在体裁差异，层级关系为：学术文本 > 通用文本 > 媒体文本 > 小说文本。具体说来，学术子库变化幅度最大，小说子库变化幅度极微。较传统的语言演化研究，基于取样间隔时间短的历时语料和多体裁、多变量分析有助于细致描述语言变化的趋势和体裁内部差异，弥补既往研究缺少平衡语料数据支持的不足，丰富汉语历时描写的研究方法和研究思路。

研究发现，汉语结构复杂度呈现阶段性变化特征，1920-1930年间经历了快速变化，但这种变化主要发生在非文学文本中。一般认为，汉语句法复杂度增大与白话文运动借重西方语言改造汉语的尝试有关。本文的描述和分析说明，由于模糊了文体和体裁的界限，现有的汉语结构复杂化趋势描述掩盖了这一变化趋势在体裁上的差异。

体现汉语句子组合特征的句段在长度上也经历了历时变化，变化趋势和幅度与结构复杂度基本一致，说明句段长度对结构复杂度有较好的预测能力，未来研究可进一步探讨句段长度对可读性和体裁的预测能力，使之成为具有较强解释力的变量。

与结构复杂度密切相关、反映结构内部构成的类联接和依存距离分析也表现出体裁差异。比如，“的”字修饰成分在非文学文本中的使用频率远高于文学文本，这是非文学文本结构复杂度增加的主要原因。这说明，考察现代汉语百年变化，体裁差异维度不可或缺。早在20世纪40年代，王力就指出欧化只是文法上的欧化，“还不大看见它在口语里出现”(王力2011：334)。如果小说文体接近口语这一说法成立，就可以说汉语变化多体现于书面文本，但未在口语中表现出来。

王力(2011：337)曾指出，现代汉语语法欧化在1940年之前已经“完成了十分之九的路程”。从本文所描述的变化趋势看，汉语结构复杂度在学术语言中的增大趋势一直在延续，但这不能用欧化来解释。事实上，是否

存在欧化也是学界争议的问题。本文对汉语句法复杂度的历时描述说明，汉语结构的变化主要体现在非文学语言结构内部容量增大和内部成分复杂度增大，句法规则并未发生根本变化。

另外，研究认为，除了时间顺序和主题突出这两个广受认可的特征，汉语描述还应重视另一个重要的类型学特征：结构封闭性。一般认为，现代汉语发展是自“五四”之后，表现为句子延长、复杂(王力2011：348-349；向熹2010：806-819)。相关研究聚焦于局部语法特征，如次品谓语形式(王力2011：349)和“DV”句(崔山佳2013：796-877)，但都没有考虑汉语结构的封闭特性，而不在结构内部观察局部语法特征就难以揭示汉语句子复杂的内在原因。本研究所做的类联接和依存距离描述一定程度上揭示了结构容量和复杂度变化的内部成因，表明类联接和依存距离是描述结构变化的有效切入点。

就语言演化而言，基于短时间间隔取样的语料、依照体裁分类实施的句段长度和多变量复杂度分析明显提升了描述精细度，更易于挖掘传统方法难以发现的演化趋势和文类差异。有自然语言处理技术的支持，语言演化研究就可能扩展变量选择和观察范围，对大规模经验性语言数据进行多角度、多层面分析，实现对语言的充分描述。有了充分的描述才有望探索新问题，开辟新疆域，完成对语言的充分解释。

注释：

①本库参照兰卡斯特汉语语料库(简称LCMC)的体裁分类，由媒体(Press)、小说(Fiction)、通用(General Prose)和学术(Learned)四个子库构成，收录包括新闻报道、社论、新闻评论、宗教等(幽默文本除外)14类体裁的文本。

②Biber(1993)认为，评估语料库设计语料抽样的代表性要考虑特定语言文本类型的涵盖范围和特定语言特征的分布范围。

③对容量敏感的小句结构若带有时范畴“着”“了”“过”，则会增加句子的信息负载，也会相应增加汉语句子的处理难度(龙果夫1958；雅洪托夫1959)。同理，“所”“之”“的”等结构助词的使用也正是处理难度增加的词汇标志。

④Longobardi(2001)认为句法在历时演化上存在惰性，除非受音系、词汇或语义演化或通过界面和语法外部压力，驱使其发生改变，否则句法本身不会发生变化。

⑤图11中的数字为语料库中的样本编号，按体裁类别历时编列，样本编号N的计算方式：N=r+ON*4。其中，r为文本类型起始数字：r=1～4(1为非文学通用；2为非文学_新闻；3为非文学_学术；4为文学)；ON为序数，即第n-1个同类文本：ON=1～21。其中，1、5、9、13、17、21……53……65、69、73、77、81为非文学通用文本；2、6、10、14、18、22……54……66、70、74、78、82为非文学新闻文本；3、7、11、15、19、23……55……71、75、79、83为非文学学术文本；4、8、12、16、20、24……56……68、72、76、80、84为文学文本。

参考文献：

[1]Ai, H. & X. Lu. 2013. A corpus-based comparison of syntactic complexity in NNS and NS university students´ writing[A]. In A. Diaz-Negrillo, N. Ballier & P. Thompson(eds.). Automatic Treatment and Analysis of Learner Corpus Data[C].Amsterdam: John Benjamins.

[2]Biber, D. 1993. Representativeness in corpus design[J]. Literary and Linguistic Computing, (4): 243-257.

[3]Biber, D. & S. Conrad. 2019. Register, Genre, and Style[M]. Cambridge: Cambridge University Press.

[4]Bussmann, H. 2000. Routledge Dictionary of Language and Linguistics[M]. Shanghai: Foreign Language Teaching and Researching Press.

[5]Firth, J. 1957. Modes of meaning[A]. In J. Firth(ed.). Papers in Linguistics 1934-1951[C]. London: Oxford University Press.

[6]Givón, T. 1998. The functional approach to grammar[A]. In M. Tomasello(ed.). The New Psychology of Language[C]. Norwood: Lawrence Erlbaum Associates.

[7]Kintsch, W. 1998. Comprehension: A Paradigm for Cognition[M]. Cambridge: Cambridge University Press.

[8]Lin, D. 1996. On the structural complexity of natural language sentences[R].The 16th International Conference on Computational Linguistics. Copenhagen.

[9]Longobardi, G. 2001. Formal syntax, diachronic minimalism, and etymology: The history of French, Chez[J]. Linguistic Inquiry, (2): 275-302.

[10]McEnery, T., R. Xiao & Y. Tono. 2006. Corpus-Based Language Studies: An Advanced Resource Book[M]. New York: Routledge.

[11]Ortega, L. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing[J]. Applied Linguistics, (4): 492-518.

[12]Pallotti, G. 2015. A simple view of linguistic complexity[J]. Second Language Research, (1): 117-134.

[13]Wynne, M. 2008. Search and concordanceing[A]. In A. Lüdeling & M. Kyt{H1AH1199.jpg} (eds.). Corpus Linguistics: An International Handbook[C]. Berlin & New York: Walter de Gruyter.

[14]崔山佳.2013.汉语欧化语法现象专题研究[M].成都:巴蜀书社.

[15]方梅.2005.认证义谓宾动词的虚化——从谓宾的动词到语用标准[J].中国语文,(6):495-507.

[16]刘丹青.2001.语法化中的更新、强化与叠加[J].语言研究,(2):71-81.

[17]龙果夫.1958.现代汉语语法研究[M].北京:科学出版社.

[18]秦洪武.2010.英译汉翻译语言的结构容量:基于多译本语料库的研究[J].外国语,(4):73-80.

[19]秦洪武 孔蕾.2018.翻译语言影响原创语言的途径和方式——基于汉语结构复杂度的分析[J].外国语,(5):15-26.

[20]秦洪武 王克非.2014.历史语料库:类型、研制与应用[J].外语与外语教学,(4):1-7.

[21]沈家煊.1998.实词虚化的机制——《演化而来的语法》评介[J].当代语言学,(3):41-46.

[22]沈家煊.2012."零句"和流水句[J].中国语文,(5):403-415

[23]石毓智 李讷.2000.十五世纪前后的句法变化与现代汉语否定标记系统的形成——否定标记"没(有)"产生的句法背景及其语法化过程[J].语言研究,(2):40-54.

[24]王克非 秦洪武.2017.基于历时复合语料库的翻译与现代汉语变化考察[J].外语教学与研究,(1):37-50.

[25]王力.2011.中国现代语法[M].北京:商务印书馆.

[26]吴福祥.2004.近年来语法化研究的进展[J].外语教学与研究,(1):18-24.

[27]夏云 秦洪武.2017.翻译与现代汉语结构容量的变化——以"介词……方位词"结构为例[J].外国语,(6):77-85.

[28]向熹.2010.简明汉语史[M].北京:商务印书馆.

[29]雅洪托夫.1959.汉语的动词范畴[M].北京:中华书局.

[30]张谊生.2000.论与汉语副词相关的虚化机制——兼论现代汉语副词的性质、分类与范围[J].中国语文,(1):3-15.

[31]张中行.2007.文言和白话[M].北京:中华书局.

[32]朱德熙.1982.语法讲义[M].北京:商务印书馆.

[33]朱一凡.2018.基于语料库的英汉翻译对当代汉语影响的研究[M].上海:上海交通大学出版社.

