基于标注语料库的《新闻联播》语言特征统计分析

王彬 王依然 文采菊 周鑫

（中国传媒大学 应用语言学系 北京 100024）

摘 要： 本文利用传媒语言语料库及各种相关软件工具，对《新闻联播》的部分语料进行了词性、结构、句法和语义关系标注，并从词汇、句法、语义三方面进行数据提取、统计，在此基础上分析其语言特征，得出的结论是《新闻联播》所使用的词语书面语色彩较浓，句子比较复杂，语义关系比较固定。这种定量分析与定性解释相结合的方法，对于作为中国官方正式的新闻发布管道的《新闻联播》的语言特征分析，显得比较客观全面，同时可以为传媒语言的研究提供一些新的思路。

关键词： 语料库；标注；新闻联播；统计分析

A Statistic Analysis on the Linguistic Features of News Broadcasting on Tagged Corpus

Wang Bin Wang Yiran Wen Caiju Zhou Xin

(Applied Linguistics Department, Communication University of China, Beijing 100024)

Abstract: This paper conducted some statistical research and analysis on the CCTV news program *News Broadcasting* based on the tagged media language corpus and kinds of relative software and tools to describe some linguistic features on the lexical, syntactic and semantic dimensions. It is concluded that the words it uses are rather formal, the sentences are somehow complex, and the semantic relations are relatively fixed. This research method which combined quantitative analysis with qualitative explanation is useful for the objective and comprehensive analyses of the linguistic features of *News Broadcasting*, which is regarded as the Chinese official news issuance channel. Meanwhile it provides some new thoughts for media language study.

Keywords: corpus; annotation; News Broadcasting; statistic analysis

1 前言

随着新闻传播事业的飞速发展，有很多学者对新闻语言进行了不同角度的研究，但是绝大多数研究都缺乏较大规模真实语料的支持。20 世纪 60 年代以来，由于计算机的发展，基于带标语料库统计的语言分析方法使得人们的研究语料处理速度更快，分析精确度更高，得出的结论更可靠。对语言的定量分析与定性解释相结合，使得我们对语言的描述更客观、更全面。《新闻联播》作为中国官方正式的新闻发布管道，应该对其语言特征予以重视。正是出于上述原因，我们在中国传媒大学传媒语言语料库中随机选取了 10 篇《新闻联播》文本语料，进行预处理后，对它进行了较全面的标注，从词汇、句法、语义三方面进行数据提取、统计、分析，试图描述《新闻联播》的一些语言特征。

作者简介：王彬（1983-），王依然（1984-），文采菊（1984-），周鑫（1983-），女，中国传媒大学应用语言系 2005 级硕士研究生，研究方向为应用语言学，E-mail:wdxj0308@cuc.edu.cn

2 研究思路与方法

由于新闻联播中的字幕标题及同期声部分属于非播报内容，有其自身的语言特点，与正式的播报内容有较大差异，为了求得研究对象的同质性，我们去除了该部分，只保留纯播报内容。

我们首先对语料进行分词及词性标注，采用的是北京大学计算语言所开发、中国传媒大学应用语言学系完善的自动分词标注系统[1]。

接下来进行了结构、句法、语义关系的标注，在完成了所有标注工作后，我们使用了中国传媒大学的 TCRS 语料库检索系统对加工过的熟语料进行数据的提取及统计。

3 词汇特征分析

3.1 词语使用

我们使用中国传媒大学 TCRS 语料库检索系统对标注的 10 篇《新闻联播》熟语料进行了词频统计，为求具有可比性，我们又随机选取了与《新闻联播》字数大致相等的北京电视台民生新闻节目《第七日》的 14 篇语料进行词频统计。各类词具体使用的数据对比见下表：

词性	数量		词性	数量		范畴	数量	
	新闻联播	第七日		新闻联播	第七日		新闻联播	第七日
名词	2646	2004	代词	93	90	习用语	63	6
动词	2456	2134	方位词	81	74	成语	37	0
形容词	402	500	处所词	62	50	专业术语词	35	14
副词	304	377	介词	57	62	新词	29	19
时间词	213	144	连词	61	87	口语词	25	137
数词	143	131	助词	17	21	方言词	0	23
量词	127	156	语气词	6	20	惯用语	0	9
区别词	115	72						

在统计过程中，笔者发现人名、地名、机构名、简称略语在数量上的差别是由于节目题材不同引起的，故在此不予以讨论。以下是二者在词汇使用上的一些差别：

（1）《新闻联播》中使用了 63 个习用语，其中四字格的习用语例如：“对外开放”、“固定资产”、“爱岗敬业”等有 55 个，占 87.31%，而《第七日》中只使用了 6 个习用语，全部为三字格，分别是“差不多”、“有意思”、“傻了眼”、“一连串”、“交朋友”、“添麻烦”；

（2）《新闻联播》中使用了 37 个成语，例如：“来之不易”、“千方百计”、“脍炙人口”等，而《第七日》则没有使用成语；

（3）《新闻联播》中使用口语词 25 个，例如：“定心丸”、“秋老虎”、“盼头”，而《第七日》中则使用了 137 个；

（4）《新闻联播》中方言词、惯用语的使用次数均为 0，而在《第七日》当中这两类词都有使用。

从这些差别中可以看出，《新闻联播》极少使用口语色彩浓的方言词、惯用语、口语词，而这些词在《第七日》中经常使用；同时，《新闻联播》中使用的书面语色彩浓的成语在《第七日》中则没有使用；在习用语方面，《新闻联播》倾向于使用整齐的四字格形式。虽然同是新闻类节目，《新闻联播》作为中国官方正式的新闻发布管道，是关乎国计民生的硬新闻，所以书面色彩更浓，语体更加正式。

[1] 该分词标注系统正确率达不到 100%，但错误率基本控制在 2%左右，对统计结果不会有太大的影响。

3.2 词汇密度

词汇密度(lexical density)这一概念是 Halliday 在 1985 年时提出的。Halliday 认为句子由词汇项(lexical item)以及语法项(grammatical item)组成。前者主要由实词等组成，而后者主要指功能词。词汇密度指词汇项在整个句子中所占的比例。通常而言，词汇密度的大小是口语与书面语的重要区别，也是句子复杂度的重要指标。在本研究中，我们用词汇密度来指称实词(lexical words)与话语中词汇总数的比率，我们可以用以下的公式来表示：

T = 话语的词汇总数

L = 话语中的实词数

词汇密度 = L / T × 100%

由于目前对于什么是词仍然没有一个公认的定义，对于实词和虚词的定义和理解我们采用的是黄伯荣，廖序东在《现代汉语》增订三版中对于词类的划分。其中实词包括：名词(n)，动词(v)，形容词(a)，区别词(b)，数词(m)，量词(q)，副词(d)，代词(r)，拟声词(o)，叹词(e)；虚词包括：介词(p)，连词(c)，助词(u)，语气词(y)。

统计结果是：《新闻联播》播报内容的词汇密度=L/T×100%= 44518/51075×100% = 87.16%，

《第七日》播报内容的词汇密度= L/T×100%= 50043/59155×100% = 84.60%。

这两个数据远远超过了口语语体词汇密度 40%的标准[2]，从这个角度来说，两者都具有明显的书面语色彩。但是在统计虚词过程中，我们发现在《新闻联播》中，语气词使用了 87 次，占其总词次的 0.17%，而《第七日》中语气词使用了 1873 次，占其总词次的 3.17%，由此可见后者作为民生新闻节目，所使用的语气词数量大大超过了前者，其语体不如《新闻联播》正式，而语体越正式，其词汇密度越高。所以尽管同是新闻类节目，《第七日》的词汇密度比《新闻联播》要低。

4 句法特征分析

在这一部分中，我们主要从引发句、句子的复杂情况及整句与零句三个方面来描述和解析新闻联播语言在句法方面的特点。

4.1 关于引发句

在语料的实际标注过程中我们发现，新闻联播中时常出现诸如“XX 说/指出/表示/强调，……”之类的句子，且动词后面所关涉的内容往往较为复杂，以复句居多。对于此类句子传统语法通常分析为“主语+谓语+宾语”，即不管后面的部分有多复杂，都统统视作宾语，但这从逻辑上和人们的语感上看都是不太符合实际的，因而我们在这里把这种特殊的句子单独提出来，作为一类，称之为引发句，标注时记作 YF，其后面的部分作为独立的句子来分析标注。例如：

(**YF**_ZJ (O1_SU 李/nr 长春/nr)(I 强调/v) ,/w) (DJ_ZJ(O1_SU_NP(A1_FW 各级/r)(A2_XD 党政/b/j)(! 领导/n)(NV 要/v)(E_NR_PP 把/p 加强/v 和/c 改进/v 未/d 成年人/n 的/u 思想/n 道德/n 建设/v)(I 作为/v)(O2_NR_NP(A1_XD_VP 贯彻/v 落实/v "/w 三个代表/j "/w 重要/a 思想/n 的/u)(A2_XS 重要/a)(! 任务/n)) 。/w)

通过 TCRS 语料库检索系统，我们统计出所选取新闻联播语料中引发句共 242 个，占句子总数的 11.15%。引发句中的谓语多由以下二价动词来充当，如：说、表示、指出、强调、认为、宣布、证实、透露、表明、预测、提出、公布、分析、显示、估计、推测等，其中“说、表示、指出、强调”位居前四位，分别占 27.69%、17.77%、12.81%和 8.26%。引发句多以人名、机构名、文件通知等做主语，其中国家领导人和权威机构为施事主语的占绝大多数，这也从一个侧面体现了新闻联播作为中国官方信息发布的管道，以上情下达为宗旨。

4.2 关于句子复杂情况

对句子复杂度的计算是一项相当复杂的工作，涉及哪些因素，加权系数多少，都需要严密的计算和论证，目前我们的能力还达不到，所以这里我们仅从平均句长、平均分句数和复句的复杂度三方面来稍加分析，力图反映

[2] 该标准是 Ure 采用计算篇章词汇密度的方法对她所采集的篇章进行分析得出的结论。

新闻联播中所使用句子的相对复杂状况。

平均句长是衡量句子复杂情况的一个重要的量化指标，我们用总字数与小句总数（引发句、单句和所有复句的分句之和）相除，得出平均句长为 17.92 字；分句数量的多少也可以反映句子的长短，对于平均分句数的统计，我们是用小句总数除以单句与复句的数量总和，得到的结果为 2.12 句，即平均每个句子包含 2.12 个分句，与平均句长结合起来看，则平均每个句子包含 37.99 个字，这说明《新闻联播》中的句子相对较长；此外复句的复杂状况可以通过复句中所含的分句个数和复句的层次数两个指标来考查：从检索结果可以看出，九万多字的新闻联播语料中单句共有 880 个，占句子总数的 40.55%，而复句共有 1048 个，占总数的 48.30%，略多于单句，其余的都是引发句；就其复杂度而言，分句个数方面，复句中分句最多可达 13 个，层次方面，最高为 6 层，包含 11 个分句，是所有复句中包含层次关系最多的。具体如下表所示：

复句中的分句个数	2	3	4	5	6	7	8	10	9	11	12	13
相应的复句数	665	175	103	54	29	10	4	3	2	1	1	1
占复句总数的百分比	63.45%	16.70%	9.83%	5.15%	2.77%	0.95%	0.38%	0.29%	0.19%	0.10%	0.10%	0.10%
复句的层次数	1		2		3		4		5		6	
相应的复句数	724		236		57		24		5		2	
占复句总数的百分比	69.08%		22.52%		5.44%		2.29%		0.48%		0.19%	

（注：上表是按包含不同分句个数和层次数的复句数从大到小排列）

由此可见，总体来说，复句的使用略多于单句；复句中分句的个数与其出现频率大体上成反比，即分句个数越少、越简单的使用频率越高，两个分句组成的复句占了全部复句的 63.45%，而超过 5 个分句的复句出现频率较低；就层次来看，复句的层次数与其出现频率也成反比，1 层复句最多，占了 69.08%， 3 层以上的复句只偶尔出现。如此看来，新闻联播中的复句严格来说并不算十分复杂，与平均句长、平均分句数相对照，或许可以反映出新闻联播作为新闻语体，一方面力图在有限的篇幅内使用最为严整的语言来涵盖最大的信息量，而另一方面作为消息播报，又必须充分考虑受众的接受能力，因而需在两者之间做出协调，以找到最佳平衡点，表现在句子的复杂情况上，则一方面平均每句用字较多，而另一方面所用复句的分句和层次又不太复杂，由此来寻找到那个经过调适的兼具严密性和适听性的度。

4.3 关于整句与零句

通常来说，整句（ZJ）是指主谓俱全的句子，零句（LJ）是指不具备“主语—谓语”形式的句子。在实际的标注中，我们发现，《新闻联播》中的零句绝大部分都是缺省了主语的小句。例如： (FJ(FJ1_**ZJ**(E_SJ_NP 今年/t 一/m 季度/n)(O1_DS_NP (A_XD_NP 我国/r 经济/n 增长/v)(! 速度/n))(I 达到/v)(O2_NR 9.7%/m) ，/w)LGF(FJ2_**LJ**(E_FS 明显/a)(I 高于/v)(O2_DS_NP (A1_XD_VP 年初/t 我国/r 确定/v 的/u)(A2_XD 7%/n 的/u)(A3_XD 预定/v)(! 目标/n)) 。/w))

检索《新闻联播》语料得到的整句共 2677 个，占总基数的 65.40%；而零句只有 1416 个，仅占 34.60%，可见整句的数量远远多于零句。其中单句中整句有 816 个，零句仅有 64 个，整句的大量使用体现了新闻联播的严密性、规整性；复句中整句有 1619 个，零句有 1352 个，并且零句中省略的基本上都是主语，即主要通过省略的手段来实现零形回指，因此，虽然复句中仍以整句居多，但整句与零句相差并不太大，从而体现了新闻联播语言的连贯性。

5 语义关系特征分析

这里的语义关系，主要指的是句子中各项论元与谓语动词之间以及定语与中心语之间的逻辑关系。在该部分，我们从多项定语、多项状语角度做了统计分析，试图说明新闻联播语言在语义搭配上的一些特点。

5.1 多项定语的语义关系统计

我们首先对定语总和进行了统计，单项定语为 A，多项定语从第一项开始依次记作 A1、A2、A3…An。经统计，定语共有 4004 个，n 的最大值为 5，其中单项定语 A 为 1490 个，多项定语中 A1 为 1083 个，A2 为 1067 个，A3 为 295 个，A4 为 62 个，A5 为 7 个。

接下来分别统计了限定、修饰、领属、范围等 21 种语义关系的定语的总和，统计后排名前九位的定语如下：

所含语义关系	总数	所占比例	所含语义关系	总数	所占比例
限定（XD）	1864	46.55%	同指 (TZ)	113	2.82%
修饰（XS）	434	10.83%	内容 (NR)	96	2.40%
数量（NU）	350	8.74%	时间 (SJ)	96	2.40%
领属 (LS)	307	7.68%	场所 (CS)	86	2.15%
范围 (FW)	230	5.74%			

注： 由于标注存在误差，某些定语未标注语义关系，故含有语义关系的各项定语总和比实际定语总和略少。

这里需要说明的是：我们在对定语语义关系进行归类时，发现有一类语义关系明显不是表示修饰语义关系，但也无法归入数量、时间、同指等语义关系类别中的任何一类，例如“内资企业”中的“内资”显然不是修饰“企业”，但也不是其他的语义关系，所以针对这种情况我们单列出了一类——限定语义关系。不过，表示限定语义关系的定语和表示数量、领属、范围、同指等语义关系的定语一样都采用了限定性语言，而表示修饰语义关系的定语则采用了描写性语言。

阐述明白这些，我们再来看数据：表示限定语义关系的定语最多，几乎占一半，再加上表示数量、领属、范围、同指、内容、时间、场所等语义关系的定语，共占所有定语总和的 78.48%，而表示修饰语义关系的定语占了 10.83%。由此看出限定性定语在新闻联播中有绝对优势，而之所以会出现这种情况和《新闻联播》在传播学上的特性以及电视新闻的特性有关。首先，《新闻联播》是消息类电视新闻节目的典型代表，所以必须内容简要，信息量大，而限定性的定语传达了有关数字、时间、地点、人与物的关系、人与人的关系等大量的信息，但又不会增加句子个数，所以在《新闻联播》中，限定性定语会出现得较多。其次，《新闻联播》是电视新闻，其一大特征就是声画互为映衬，没有冗余信息。画面可以表现的人或事物无需用语言再进行描述，所以在《新闻联播》中出现大量的描写性定语是不大可能的。

5.2 多项状语的语义关系统计

我们还对状语进行了统计，单项状语为 E，多项状语从第一项开始依次记作 E1、E2、E3…En。检索出的状语一共是 3796 个，n 的最大值为 5。其中单项状语 E 为 1506 个，占到全部状语数量的 39.67%，E1 有 977 项、E2 有 960 项、E3 是 300 项、E4 是 44 项、E5 是 5 项。

通过语义标注以及统计我们还发现，在所有状语表示的语义关系上，表示时间的共 976 项，这里的“时间”指的是具体的时间，例如“今年”、“4 月 11 号”，占所有状语的 25.71%；表示情态的共有 603 项，这里所说的“情态”既包括“已、已经、将要、还是、永远、一向”这样的表示抽象时间、频率的词，还包括“就、可”这样的语气词，占所有状语的 15.88%，而表示方式和处所的也比较多，分别是 9.45%和 6.27%。这四种语义关系的使用占所有 27 项语义关系使用的一半还多。其中：

（1）单项状语 E 的 1506 项中，语义关系位于前五位的分别是时间（383 个）、情态（211 个）、方式（157 个）、范围（100 个），状态（85 个）；

（2）多项状语的 977 个第一项状语 E1 中，语义关系位于前五位的分别是时间（437 个）、情态（162 个）、方式（61 个）、范围（40 个）、状态（39 个）；960 个第二项状语 E2 中，语义关系位于前五位的分别是情态（179 个）、时间（131 个）、处所（112 个）、方式（94 个）、修饰（44 个）； 300 个第三项状语 E3 中，语义关系位于前五位的分别是情态（38 个）、方式（36 个）、处所（28 个）、时间（24 个）、目标（19 个）；44 个第四项状语 E4 中，语义关系位于前五位的分别是情态（12 个）、方式（8 个）、状态（5 个）、范围（3 个）、与事（3 个）、修饰（2 个）；第五项状语 E5 只有 9 个，表示情态、方式的分别有 2 个，表示原因、修饰、限定、状态以及程度的均为 1 个。

所以无论是单项状语还是多项状语，在语义关系上明显偏重于表示时间、情态、方式以及处所。从新闻传播的角度来讲，《新闻联播》作为消息类电视新闻节目的典型，必须要让观众清晰地得知何时（When）、何地(Where)、何人(Who)、何故(Why)、采用何法(How)做了何事(What)，而状语在新闻当中恰恰就承担了表述何时、何地、该事件如何的任务。我们的统计也在一定程度上体现新闻的时效性，新闻联播一天一播报，所以时间词中频率最高的就是“今天”一词，一共使用了 185 次。副词使用频率最高的 10 个中，表示情态的占了 7 个，分别是“将、也、就、都、还、已经、已”，它们的使用频次一共有 734 次。表示方式的“进一步、共同、先后、相互、一起、逐步、始终”使用频次也很高。

6 结语

从以上三部分的统计分析可以发现《新闻联播》在词汇、句法、语义上的若干特征：

（1） 词汇上，《新闻联播》倾向于使用书面语色彩浓的词语，与其他电视新闻类节目相比语体更加正式。

（2） 句法上，《新闻联播》所用的句子相对比较复杂，同时也体现了严整性、庄重性与连贯性的协调统一。

（3） 语义上，定语多表示限定性语义关系，状语多表示时间、情态、方式、处所的语义关系。这些都体现了《新闻联播》言简意赅、信息含量大的特点。

总之，基于一定规模的标注语料库的统计方法，对于研究传媒语言不失为一种科学客观的现代语言学研究方法。

参考文献

[1] 黄匡宇， 李岩， 张联编著. 广播电视新闻学[M]. 北京：高等教育出版社，2002.

[2] 孙道功， 亢世勇， 孙茂松. 基于标注语料库的现代汉语单句句型句模的对应关系研究. [C] 自然语言理解与大规模内容计算. 北京：清华大学出版社. 2005.

[3] 沈紫薇. 英语新闻的一些语言特点：语料库分析法.[J] 经济与社会发展. 2003 年 3 月.

[4] 赵元任著，吕叔湘译. 汉语口语语法[M]. 北京：商务印书馆, 1979.

[5] 陈会君. 词汇密度与难易度感知——科学论文及其摘要的对比研究. [J] 外语与外语教学. 2003 年第 4 期.