Modern Linguistics 现代语言学, 2021, 9(5), 1294-1299
Published Online October 2021 in Hans. http://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2021.95176

语料库视角下十八大报告与十九大报告词汇对比研究

陈　劼

华中科技大学人文学院，湖北　武汉

收稿日期：2021年8月17日；录用日期：2021年10月15日；发布日期：2021年10月27日

摘　要

本文借助CorpusWordParser、Excel、AntConc等工具对党的“十八大报告”和“十九大报告”做字词频、主题词等相关统计和比较。透过词汇的“不变”，透析党中央坚持系统谋划、统筹推进党和国家各项事业的一脉相承；透过词汇的“变”，展现党中央不断根据新的实践需要，探索形成新思想、新目标、新部署、新要求的与时俱进，在社会发展中揭示语言的动态本质。

关键词

词频，主题词，语料库，统计分析

A Contrastive Study on the Vocabulary of the 18th CPC National Congress Report and the 19th CPC National Congress Report from the Perspective of Corpus

Jie Chen

School of Humanities, Huazhong University of Science and Technology, Wuhan Hubei

Received: Aug. 17th, 2021; accepted: Oct. 15th, 2021; published: Oct. 27th, 2021

Abstract

With the help of CorpusWordParser, Excel, AntConc and other tools, this paper makes statistics

文章引用: 陈劼. 语料库视角下十八大报告与十九大报告词汇对比研究[J]. 现代语言学, 2021, 9(5): 1294-1299.
DOI: 10.12677/ml.2021.95176

and comparison on the word frequency and Keywords of the "18th National Congress report" and "19th national congress report" of the Communist Party of China. Through the invariant words, this paper analyzes that the CPC Central Committee adheres to systematic planning and overall promotion of various undertakings of the party and the state; through the variant words, it shows that the Party Central Committee is constantly exploring the formation of new ideas, new goals, new deployments, and new requirements in accordance with new practical needs. It reveals the dynamic nature of language in social development.

Keywords

Word Frequency, Keywords, Corpus, Statistical Analysis

1. 引言

语言是一种特殊的社会现象，是表征社会的一种形式。社会性作为语言的本质属性，使得其形成、发展、消亡都取决于社会意志和社会需要。在一定的社会群体中，其代表性语言载体能较为充分表征社会的现状、发展及变化。中国共产党是中国特色社会主义事业的领导核心，每五年一次的党的全国代表大会是党和政府工作的航标。大会所作报告作为典型的政治语篇，具有权威性、指导性和纲领性。语言系统中变化最快、最明显的是词汇。本研究基于语料库工具技术，从词汇入手，对中国共产党十八大报告《坚定不移沿着中国特色社会主义道路前进 为全面建成小康社会而奋斗》和十九大报告《决胜全面建成小康社会 夺取新时代中国特色社会主义伟大胜利》做比较研究，并尝试结合语言使用与社会发展的共变关系及规律对其进行简单说明。

2. 比较“十八大报告”和“十九大报告”词表的异同

本文以从中国共产党历次全国代表大会数据库(http://cpc.people.com.cn/)获得的两次报告内容为语料，在对其进行整理和筛选的基础上，综合 CorpusWordParser 和 AntConc 两个软件的分词成果，人工梳理得出最终分词结果[1]。接着使用 Python 语言，对原始数据按空格进行切割(str.split)，并去除标点符号类的词，然后将词作为键，出现次数作为值，构建一个字典数据结构(dict)，遍历所有的词，对每个词的键所对应的值进行自增操作，得到一个完整的词频统计字典。由于字典的无序属性，我们需要重新对字典按值进行排序，得到有序元组列表数据结构(tuplelist)。最后遍历有序元组列表，逐行以英文逗号分隔分别输出“词”“频数”“频率”，得到一个 csv 格式的文件并导入 excel，形成两份报告各自的专属词表。

经统计，十八大报告全文词语数为 13529，词种数为 2530；十九大报告全文词语数为 15404，词种数为 3023。借助筛选，我们得到二者相同的词有 1771 个，例如“的”“和”“发展”“建设”“党”“中国”“社会主义”“人民”“坚持”“是”“社会”“特色”“在”“国家”“要”“制度”“文化”“新”“全面”“推进”等。十八大独有词 759 个，例如“有利于”“单位”“党建”“十年”“反腐倡廉”“热爱”“借鉴”“调解”“关乎”“台阶”等。十九大独有词 1252 个，例如“中国梦”“梦想”“本领”“脱贫”“文艺”“决胜”“激励”“能够”“初心”“一流”“一带一路”“共建”“抓”

[1]不同软件、不同方法所得的分词结果存在差异，但大体趋势和相对数据无明显区别。

“底线”“攻坚战”等。

3. 两份报告高频词比较及其差异显著性调查

我们利用 excel，对两份报告的高频词进行比较，发现了一些共性和差异。

(一) 高频词比较与透析

如图 1 所示，从内容上看，十九大报告和十八大报告高频前十位的词是完全一致的，可以在一定程度上反映出，党始终把“发展”摆在重要位置，强调“建设”的行动力，从国家的大局出发，始终将“人民”作为首要关注，“坚持”“社会主义”发展道路不动摇。此外，具有主体性义项的词“党”“中国”“人民”等词频在十九大报告中要普遍高于十八大报告，也在一定程度上反映和体现了党的执政理念，表明了党执政的宗旨、目的和价值追求。

	十八大报告	频数	频率
1	的	514	3.80%
2	和	347	2.56%
3	发展	279	2.06%
4	建设	185	1.37%
5	党	131	0.97%
6	中国	127	0.94%
7	社会主义	154	1.14%
8	人民	127	0.94%
9	坚持	107	0.79%
10	是	104	0.77%

	十九大报告	频数	频率
1	的	696	4.52%
2	和	375	2.43%
3	党	219	1.42%
4	发展	216	1.40%
5	中国	165	1.07%
6	人民	164	1.06%
7	建设	161	1.05%
8	社会主义	134	0.87%
9	是	133	0.86%
10	坚持	132	0.86%

Figure 1. Comparison of high-frequency words in the reports of the 18th and 19th National Congress of the Communist Party of China
图 1. 十八大报告与十九大报告高频词比较

从数据上看，数据的变化能够体现出党执政能力的不断提高和结构的不断优化。比如“党”一词在十八大报告中排第 6 位，在十九大报告中排第 3 位，频数也增加了近一倍，可以在一定程度上反映出党领导地位和领导能力的强化；再如“伟大”一词在十八大报告中排第 88 位，在十九大报告中排第 22 位，得益于“新时代中国共产党的历史使命”一章中“伟大斗争、伟大工程、伟大事业、伟大梦想”的提出，用使命凝心聚力中国梦；还有“国家”一词在十八大报告中排第 26 位，在十九大报告中排第 11 位；“中华民族”一词在十八大报告中排第 130 位，在十九大报告中排第 42 位等。

(二) 高频共现词 spss 差异显著性调查

我们以两份报告的共用词在各自文件中的频率为基础，取两者平均值从高到低排序，取前 50 词。接着计算出这些词在各自报告中每万字的“频率”作为“共现词频率”的数据进行“独立样本 T 检验”[1]。组别 1 是十八大报告的数据，组别 2 是十九大报告的数据，统计结果如图 2 所示。

从图中我们可以看到 T 检验分组统计的结果：十八大报告和十九大报告这 50 个词的平均词频分别为每万字约为 64.60 和 63.37，标准偏差约为 63.07 和 68.55。独立样本 T 检验结果：方差齐性检验的 F = 0.010，显著性为 0.920，大于 0.05 的显著性水平，取“共现词频率”一栏第一行的方差齐性检验结果进行推断，t = 0.093，双尾检验相伴概率 Sig = 0.926，大于 0.05 的显著性水平，说明两个样本之间不存在显著性差异，即两个样本所代表的“十八大报告”与“十九大报告”共现词的词频没有显著差异，其差异不具有统计学上的显著意义，这也在一定程度上反映出二者是一脉相承的。

T-检验

[数据集0]

组统计

	组别	个案数	平均值	标准偏差	标准误差平均值
共现词频率	1	50	64.601966	63.0730478	8.9198760
	2	50	63.373150	68.5549485	9.6951338

独立样本检验

		莱文方差等同性检验		平均值等同性t检验						
									差值95%置信区间	
		F	显著性	t	自由度	Sig. (双尾)	平均值差值	标准误差差值	下限	上限
共现词频率	假定等方差	0.010	0.920	0.093	98	0.926	1.2288163	13.1742099	-24.9149745	27.3726071
	不假定等方差			0.093	97.327	0.926	1.2288163	13.1742099	-24.9172338	27.3748665

Figure 2. Independent sample T test of high-frequency co-occurrence words in the reports of the 18th and 19th National Congress of the Communist Party of China
图 2. 十八大报告与十九大报告高频共现词独立样本 T 检验

4. 两份报告主题词的比较与透析

主题词的提取原理是通过对比一个连续的整篇文本和一个更大的参照语料库，把文本中词频具有显著差异的词语提取出来，生成一个主题词表。因此，统计主题词需要建立两个语料库[2]，一是观察语料库，二是参照语料库。在本研究中，我们分别进行两份报告的主题词提取。一是以十八大报告作为观察语料库(总字数 26,164)，十二大到十九大报告(除去十八大报告)作为参照语料库(总字数 170,215)；二是以十九大报告作为观察语料库(总字数 29,267)，十二大到十八大报告作为参照语料库(总字数 167,112)。

通过 AntConc 的 Keywordlist 功能，我们得到十八大报告主题词共 20 个，十九大报告主题词 49 个。部分如图 3 所示。

十八大报告+L1:U28					十九大报告				
Rank	Freq	Keyness(LL4	Effect(DICE)	Keyword	Rank	Freq	Keyness(LL4	Effect(DICE)	Keyword
1	94	+44.19	0.0136	特色	1	66	+90.84	0.0085	时代
2	74	+43.26	0.0108	体系	2	39	+83.28	0.0051	治理
3	52	+40.21	0.0076	推动	3	33	+54.46	0.0043	法治
4	36	+36.37	0.0053	生态	4	15	+49.97	0.0019	绿色
5	31	+30.44	0.0046	和谐	5	13	+49.51	0.0017	梦
6	279	+28.78	0.0371	发展	6	13	+49.51	0.0017	梦想
7	48	+27.64	0.007	服务	7	54	+48.07	0.007	安全
8	22	+27.12	0.0032	公共	8	76	+47.34	0.0097	伟大
9	57	+26.33	0.0083	创新	9	32	+46.82	0.0041	复兴
10	20	+23.28	0.003	持续	10	21	+44.22	0.0027	力
11	40	+22.96	0.0059	机制	11	40	+40.02	0.0052	生态

Figure 3. Comparison of keywords in the reports of the 18th and 19th National Congress of the Communist Party of China
图 3. 十八大报告与十九大报告主题词比较

借助筛选，我们得到二者共有的主题词有 5 个，其主题性(Keyness (LL4))如图 4 所示：

	十八大报告	十九大报告
体系	+43.26	+35.04
推动	+40.21	+21.71
生态	+36.37	+40.02
创新	+26.33	+21.12
全面	+18.23	+38.27

Figure 4. Keyness comparison of keywords in the reports of the 18th and 19th National Congress of the Communist Party of China
图 4. 十八大报告与十九大报告共有主题词的主题性比较

以此数据为基点，通过索引并查看其前后语境，我们发现，共有的主题词可以反映出十八大以来，党对各类体系建设、生态建设以及创新方面重视的延续性；“推动”主题性的减少主要表现推动格局的提高和结构的优化；“全面”主题性的增加主要表现在随着综合国力的提升，国家把各项建设推向纵深发展的布局更加完整、周密、具体，尤其表现在“四个全面”战略布局的提出。

当然，主题词的差异也体现得非常明显，如图 5 所示：

十八大报告独有主题词	十九大报告独有主题词
特色、和谐、发展、服务、公共、持续、机制、空间、城乡、观、推进、文化、增强、加快、化	时代、治理、法治、绿色、中国梦、梦想、安全、伟大、复兴、力、构建、中国、中华民族、脱贫、本领、美好、理念、自信、强国、决胜、建成、文明、者、人民、美丽、文艺、政治、人类、性、斗争、强军、人民军队、意识、统筹、引领、坚持、推进、一带一路、亲、共建、底线、强化

Figure 5. The unique keywords of the report of the 18th CPC national congress and the unique keywords of the report of the19th CPC national congress
图 5. 十八大报告与十九大报告独有主题词

从这些差异中我们可以看到，十八大报告相对于改革开放以来召开的历次党的全国代表大会报告，有更多的继承性，主题词数量相对较少，而十九大报告则具有更多的开创性，其主题词中的许多词都是新提出的，例如“中国梦”是习近平总书记 2012 年 11 月 29 日在国家博物馆参观“复兴之路”展览时首次阐释的；“本领”一词在十八大报告并未述及，而党的十九大报告则把“执政本领”单列提出，阐明了“全面增强执政本领”的必然性并立足八个方面阐发了“本领”的具体意涵；“一带一路”是习近平总书记于 2013 年 9 月、10 月分别提出的“丝绸之路经济带”和“21 世纪海上丝绸之路”国家级顶层合作倡议的简称。

5. 结语

语言与社会必然联系在一起，词汇作为语言的一个重要组成部分，对经济、社会的发展变化反应最敏感、最直接，具有灵活性和动态性[3]。党的全国代表大会报告是党中央精神传达、决策部署的集中体现，借助语料库分析工具，我们可以从词频、主题词等角度，用数据的相似相异感知党的性质、宗旨、

执政理念，感受党中央对党和国家各项事业既一脉相承又与时俱进的匠心布局，感悟变局中不断破局所需要的眼力、脑力、定力、动力、魄力。从词汇管窥语言呈现的规律性特点，为我们更科学地了解语言、社会及其共变，提供了重要的依据。

参考文献

[1] 钱颖. 十八大报告和政府工作报告字词频统计比较研究[J]. 华中人文论丛, 2013, 4(1): 77-80.

[2] 桂诗春, 宁春岩, 著. 语言学方法论[M]. 北京: 外语教学与研究出版社, 1997.

[3] 狄艳华, 杨忠. 基于语料库的中国政府工作报告核心主题词研究[J]. 外语学刊, 2010(6): 69-72.