1.2.4 词嵌入的回归：谁塑造了革命的内涵

在本项研究中，我们运用了一种结合了词嵌入模型与多元回归分析的词嵌入回归（ALC）技术，目的是为了量化分析历史事件对特定术语含义的影响。首先，我们将目标词汇及其上下文转化为向量形式，以此来捕捉每个词汇的语义属性。接下来，我们计算这些向量的平均值，并应用转换矩阵对常见词汇的影响力进行降权，这一步骤有助于我们更精确地把握特定上下文中的语义特征。随后，这些经过处理的词向量矩阵在回归模型中被用来估计参数，并进行因果推断。简而言之，我们的回归分析旨在确定，在满足特定条件的上下文中推断出的词向量与不满足这些条件的上下文中推断出的词向量之间是否存在显著差异，通过这种差异来推断因果关系，并解释研究结果。构建回归模型时，我们引入了一个由 ALC 向量组成的矩阵 Y，并通过回归方程

$$Y = \beta_0 + \beta_1 Post\_Revolution + \beta_3{}' Revolution' \times Post\_Revolution + \varepsilon$$

来分析。

通过这样的回归策略，我们分析了在革命中被多次讨论的重要概念，包括“平等”“自由”“民主“”民族”“公民”“权利”和“清”。估计范数β_3代表了在辛亥革命发生以后，Y 和“革命”一词之间的相似度。这个范数的显著性则代表了在辛亥革命前后，在与革命相关的语境中，Y 的词义是否有显著的变化。如果我们观察到β_3显著向为正，说明在辛亥革命以后，Y 的词义显著变化，并与“革命”一词的词义相关。

为了评估β值代表的影响是否在统计上显著，我们采用排列测试来计算经验p值。通过随机打乱 Y 矩阵中的信息，并基于这些打乱后的数据重新计算回归多次（如 999 次），我们能够记录每次回归后系数的欧几里得范数，并比较这些范数与真实数据下计算出的β系数的大小。计算得到的经验p值是通过比较随机数据下得到的规范化系数与真实数据下得到的规范化系数的比例来估计的。如果大量随机数据下得到的规范化系数大于真实数据下的规范化系数，得到的高p值将表明历史事件对术语含义变化的影响并不显著。通过这种方法，我们能够在统计上评估历史事件是否对特定术语的含义产生了显著影响，其中 β 系数揭示历史事件对我们关注的目标词汇的影响程度，而通过排列测试计算出的 p 值帮助我们判断这种影响是否具有统计学意义。

另外，为了明确辛亥革命对我们关注的目标概念造成的影响，我们设置了对照组进行回

归。对照组的回归方程如下：

$$Y = \beta_0 + \beta_1'Revolution' + \beta_2 Post\_Control + \beta_3'Revolution' \times Post\_Control + \varepsilon$$

第一个对照组我们选择辛亥革命前后的具体时间点进行对比分析。这样的分析框架旨在排除辛亥革命之前的事件或其他潜在混淆因素对词义变化观察结果的影响。

第二个对照组我们将袁世凯复辟定义为独立的对照组，由于袁世凯复辟发生在辛亥革命之后，为了排除辛亥革命的持续影响，我们将语料库进行了划分，并对两个语料库进行了回归。这一方法使我们能够基于时间节点明确区分并分析研究中的关键历史事件。

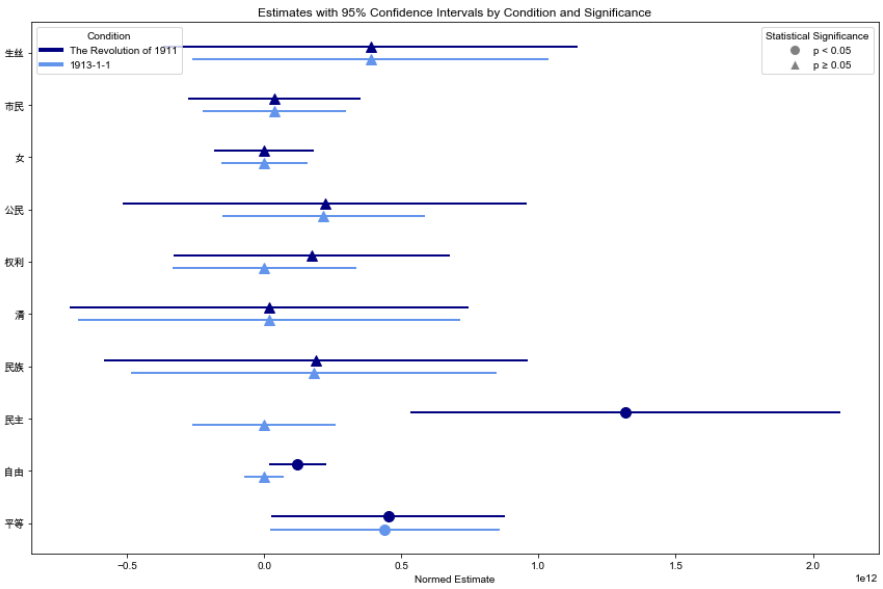
Table：The Estimated Shifts in Discourse following 1911 Revolution and the Control Event, across concepts.

	自由	**平等**	**民主**	**权利**	**公民**
1911 Revolution	1.23***	4.52***	1.32***	1.73	2.22
	(0.53)	(2.17)	(0.399)	(2.56)	(3.75)
Control1(1908/1/1)	1.21***	4.67*	1.47***	1.74	2.85
	(0.294)	(2.67)	(0.428)	3.25	(7.28)
Control2(1915/12/31)	0.005	4.21*	0.002	0.002	2.24
	(0.006)	(2.28)	(0.008)	772	(1.527)
Documents	5250	1413	495	978	3180

*p<0.1, **p<0.05, ***p<0.01

表格中的结果反映了交乘项的回归系数。我们可以看到，在辛亥革命发生之前，“自由”“平等”和“民主”等概念就开始和“革命”显著相关。但在袁世凯称帝之后，这些关键概念与“革命”之间开始减弱了。

下图更为直观地反映了这一结果，我们同样验证了在这一时期我们关注的其他概念。从结果中我们可以看出，“民主”一词在辛亥革命之后明显产生了正向的迁移，这代表“民主”一词在辛亥革命之后开始接近了革命的内涵，且这一变化是显著的。但在袁世凯上台之后，“民主”和“革命”之间的相似度又回归到 0 的周围，并且“民主”一词词义的变化不显著。这说明“革命”一词的民主内涵很快消失。“革命”中包含的“自由”的内涵也有相似的变化。但值得注意的是，辛亥革命带来的对“平等”的强调保留了下来。



2.1.2 公民身份的构建

为了讨论公民含义的变迁，我们同样构造了一个词嵌入回归模型。

$$Y = \beta_0 + \beta_1 'Ciziten' + \beta_2 Post\_Treatment + \beta_3 'Ciziten' \times Post\_Treatment + \varepsilon$$

其中 Y 是一系列我们关心的和公民身份相关的目标词汇，包括“选举”“权利”“教育”和“道德”。我们选择了一系列与历史事件有关的时间节点来构造二值变量$Post_Treatment$。$'Ciziten'$同样是一个二值变量，当目标词的上下文出现了“公民”的时候，我们将这个之赋值为 1。我们关注交互项$'Ciziten' \times Post_Treatment$的回归结果，当$\beta_3$显著时，代表着在特定历史时期以后，在上下文出现了“公民”的语段中，我们关注的目标词词义出现了显著地变化。如果我们观察到β_3显著地正向移动，则代表 Y 与“公民”在语义上更为接近。

下表给出了回归的结果，结果表明，在袁世凯称帝之后，社会舆论对“公民”的讨论更多地与“选举”这一政治权利联系在一起。并且随着时间的流逝，“公民”和“选举”在词义上更为接近。在蒋介石掌权之后，“公民”和“教育”在语义空间上更为接近。

Table：The Estimated Shifts in Discourse following Events, across concepts.

	选举	权利	教育	道德
Baseline（1912/1/1）	0.0008	1.07	0.00237	1.703
	(3.382)	(2.173)	2.445	4.291
Treat1（1913/7/1）	0.002	0.05	0.00107	1.724
	(4.303)	(1.688)	(3.34)	(3.632)
Treat2（1915/12/31）	0.002***	1.051	0.0006	1.675
	(0.0003)	(2.056)	(0.698)	2.959
Treat3（1927/2/1）	0.003***	0.006	0.003***	-
	(0.0001)	(0.619)	(0.0003)	-
Documents	15749	23680	978	962