朝鲜语口语与书面语实词间相关关系的一元线性回归分析*

卢星华　金　静

[提要] 一元线性回归分析主要研究两个变量之间的线性相关关系，是根据自变量 *x* 和因变量 *y* 的相关关系建立 *x* 与 *y* 的线性回归方程进行预测的方法。本文基于大型朝鲜语口语和书面语语料库，利用一元线性回归分析方法，计算各类实词之间的相关系数，并对其进行排序，再对中度以上的相关关系建立一元线性回归方程，并以此对朝鲜语口语和书面语进行比较研究。

[关键词] 朝鲜语　口语　书面语　一元线性回归　相关性分析

一　引　言

回归（regression）是指研究一组随机变量和另一组变量之间相关关系的统计分析方法。回归分析，是一种数学模型，是研究某一变量与另一个或多个变量之间的依存关系，用解释变量的已知值或固定值来估计或预测因变量的总体平均值（夏伯忠 1993:11）。回归分析就是要找出一个数学模型 *y=f (x)*，使得从 *x* 估计 *y* 可以用一个函数式去计算。

回归分析如果只涉及两个变量，就称为一元线性回归分析。一元线性回归分析的主要任务是从两个相关变量中选择一个变量去估计另一个变量，选择的变量称为自变量 *x*，而被估计的变量就称为因变量 *y*。当 *y=f (x)* 的形式是一个直线方程时，称为一元线性回归。这个方程一般可表示为：*y=a+bx*。其中，*a* 为截距项，*b* 为斜率项。

简单地说，一元线性回归是涉及一个自变量的回归分析，主要功能是处理两个变量（因变量与自变量）之间的线性关系，建立线性数学模型并进行评价预测。一元线性回归分析的应用领域非常广泛，如市场调查、人口分布监测、地质勘探、舆情监测等（박영호、김영화 2015；邓成竹 2017；林志伦 2018；김혜림、김동재 2018；姜道旭、吴文琴 2019；麦莉莉 2019）。

纵观国内外一元线性回归分析的应用研究，尚未发现其在语言学领域的应用研究成果。本文借鉴其他领域的应用研究成果，将一元线性回归分析作为一种新的比较研究方法引入语言学领域，基于大规模朝鲜语口语和书面语语料库，研究语料库中出现的各类实词之间的相关关系及线性趋势，把传统的单个实词出现频次的一维考察，提升至两类实词出现频次相关关系的二维考察，用数学的研究方法来检验传统语言学所观察到的实词之间的相关关系。

* 本文是国家社科基金一般项目“面向智能信息处理的韩国语口语词汇研究（16BYY176）”的阶段性成果。审稿专家提出相关修改意见和建议。谨此致谢。

二　语料来源与关注的问题

本文使用的语料库是朝鲜语口语和书面语的词素分析语料库。口语词素分析语料库包含“21 世纪世宗计划语料库”[①]中的 200 个纯口语[②]语音转写文本（访谈、师生之间的课堂交流、电话交谈等）以及 2009 年至 2022 年间笔者自建的 369 个准口语语音转写文本（电影、电视剧），共计 2,011,577 个语节[③]；书面语词素分析语料库由“21 世纪世宗计划语料库”中的 279 个书面语文本（报刊、小说、新闻等）组成，共计 10,156,140 个语节。两种语料库的词素标注均参照“21 世纪世宗计划语料库”中的词素标注体系，其中体词包括普通名词、固有名词、依存名词、代词、数词，谓词包括动词、形容词、补助谓词、肯定指定词、否定指定词，修饰词包括冠形词、普通副词、接续副词，独立词包括感叹词；关系词包括主格助词、补格助词、冠形格助词、宾格助词、副词格助词、呼格助词、引用格助词、补助词、接续助词，依存形态包括先语末词尾、终结词尾、连接词尾、名词型转成词尾、冠形词型转成词尾、名词形前缀、名词派生后缀、动词派生后缀、形容词派生后缀、词根等。朝鲜语实词包括体词、谓词、修饰词以及独立词；虚词包括关系词和依存形态。

研究中，本文将关注以下三个问题：

第一，把某两类实词的出现频次作为因变量和自变量画散点图，能否发现它们之间存在某种明显的趋势？比如，这种趋势近似直线还是曲线，或者看不出任何规律？如果这种趋势近似一条直线，那么是上升趋势还是下降趋势？在这方面，朝鲜语口语和书面语有何不同？

第二，计算某两类实词之间的相关系数后，首先根据相关系数大小进行排序，再对高度相关、中度相关、弱相关或者不相关的相关关系进行分类，其结构如何？朝鲜语口语和书面语在相关系数的排序和归类上有何不同？

第三，如果针对口语和书面语中具有高度相关和中度相关关系的实词画出直线图，代表口语和书面语实词的这两条直线是交叉还是重叠？如果是交叉，哪条直线在上面，哪条直线在下面？这都意味着什么？

三　一元线性回归分析的步骤

一元线性回归分析包括以下三个步骤：

第一步：通过画散点图来查看因变量 y 和自变量 x 之间是否存在线性相关关系。即把成对的数据用直角坐标系表示出来，如果观察到这些点的分布大致散布在一条直线的周围，则这两个变量是线性相关；如果这些点不在一条直线的周围，则这两个变量不是线性相关。

第二步：若存在线性相关关系，通过计算相关系数可查看相关程度。其计算公式如下：

$$\rho = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

① 本文使用的“21 世纪世宗计划语料库”为韩国国立国语院 2010 年 12 月出版的 DVD 修订版，下同。

② 纯口语通常指没有脚本的自然对话，准口语指有脚本的、人为编写的对话。

③ 朝鲜语中实词通常与虚词结合在一起组成一个语节，语节与语节之间用空格隔开。

其中，X_i 为第 i 个位置的自变量的值，$\bar{X}$ 为自变量的平均值；Y_i 为第 i 个位置的因变量的值，$\bar{Y}$ 为因变量的平均值。

相关系数 ρ 介于区间 [−1,1] 内。相关系数 ρ 为 −1 时，表示完全负相关。相关系数 ρ 为 +1 时，表示完全正相关。相关系数 ρ 为 0 时，表示不相关。ρ 的绝对值 $|\rho|$ 越接近 1，表示 x 与 y 两个变量之间的相关程度越强；反之，ρ 的绝对值 $|\rho|$ 越接近 0，x 与 y 两个变量之间的相关程度就越弱。通常来说，“$1\geq|\rho|\geq0.9$”为高度相关，“$0.9>|\rho|\geq0.8$”为中度相关，“$0.8>|\rho|\geq0.6$”为弱相关，“$0.6>|\rho|$”为不相关。

第三步：对高度相关、中度相关的自变量 x 和因变量 y，可以求出线性数学模型，即求出一元线性回归方程 $y=a+bx$ 的截距项 a 和斜率项 b，计算公式如下：

$$a=\bar{y}-b\bar{x}$$

$$b=\frac{\sum_{i=1}^{n}x_iy_i-\frac{1}{n}\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sum_{i=1}^{n}x_i^{\,2}-\frac{1}{n}(\sum_{i=1}^{n}x_i)^2}$$

本文使用的软件为笔者自主研发的“现代朝汉语言信息处理系统（2021 版）”。首先，利用其“统计\各语言单位的频次\朝鲜语”功能，分别对朝鲜语口语和书面语词素分析语料库中出现的普通名词、固有名词、依存名词、代词、数词、动词、形容词、补助谓词、肯定指定词、否定指定词、普通副词、接续副词、感叹词等实词计算出现频次。

口语语料库中的文本来自简短的生活对话、电影、电视剧等，因此，与以小说、报刊为主的书面语语料相比，语节数量有较大差异。因为多数书面语样本比口语样本大 10 倍左右，画出的散点图和直线图不能直观地观察到朝鲜语口语和书面语的差异，所以为便于可视化对比，本文采取整体扩放的方法，对口语语料库中出现的各类实词频次都乘以 10，整体扩大了 10 倍。

其次，利用软件的“数据分析与机器学习\一元线性回归模型”功能，按照一元线性回归分析步骤，对 569 个口语文本和 279 个书面语文本进行分析：第一步，把一类实词当作因变量 y，再把其他实词当作自变量 x，画出散点图，观察每两类实词间是否存在线性相关关系；第二步，计算变量的相关系数并查看其相关程度；第三步，对相关系数超过 0.8 的两类实词计算出一元线性回归方程并在散点图中画出其直线，查看代表口语和书面语的直线有何差异。

四　结果与讨论

由于篇幅所限，本文仅对体词中的普通名词、谓词中的动词、修饰词中的普通副词、独立词中的感叹词与除冠形词[①]外其他 12 类实词的出现频次及相关关系进行研究分析。

（一）普通名词

朝鲜语词汇中绝大多数都是名词。朝鲜语的名词分为普通名词、固有名词、依存名词。普通名词通常表示人、事物、处所、事态等的名称，在句子中充当主语、宾语、补语，经常受定语的修饰（구본관외 2015:166）。

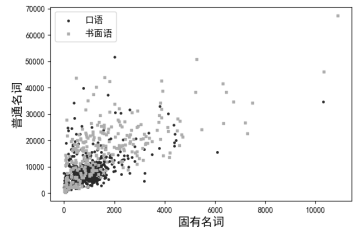
① 冠形词只有修饰体词的功能，因此本文未对该类词与其他实词的相关关系进行分析。

普通名词和其他实词的相关系数、相关程度及一元线性回归方程（以下简称“直线方程”）如表 1 所示：

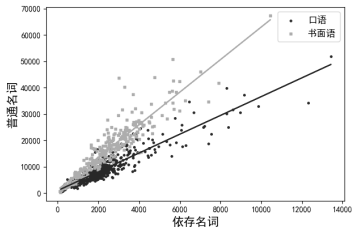
表 1　朝鲜语口语和书面语中普通名词与其他实词的相关系数、相关程度及直线方程

因变量与自变量	口语			书面语		
	相关系数	相关程度	直线方程	相关系数	相关程度	直线方程
普通名词与固有名词	0.470	不相关		0.647	弱相关	
普通名词与依存名词	0.921	高度相关	$y=953.09+3.56x$	0.922	高度相关	$y=1268.40+6.17x$
普通名词与代词	0.735	弱相关		0.346	不相关	
普通名词与数词	0.651	弱相关		0.568	不相关	
普通名词与动词	0.920	高度相关	$y=-808.35+1.29x$	0.842	中度相关	$y=1873.99+1.80x$
普通名词与形容词	0.874	中度相关	$y=123.00+3.43x$	0.693	弱相关	
普通名词与补助谓词	0.889	中度相关	$y=107.78+6.37x$	0.708	弱相关	
普通名词与肯定指定词	0.894	中度相关	$y=133.72+5.34x$	0.734	弱相关	
普通名词与否定指定词	0.662	弱相关		0.540	不相关	
普通名词与普通副词	0.861	中度相关	$y=1441.94+1.84x$	0.638	弱相关	
普通名词与接续副词	0.764	弱相关		0.559	不相关	
a 普通名词与感叹词	0.678	弱相关		−0.006	不相关	

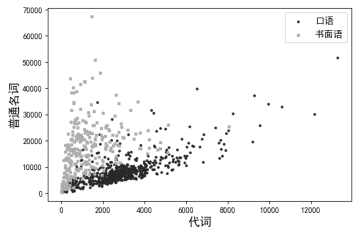
为便于观察各类实词间的相关程度及口语和书面语的差异，在同一张图中画出口语和书面语中某类实词与其他实词出现频次的散点图及直线图（图中分别用黑色和灰色代表口语和书面语，下同）。普通名词和其他实词出现频次的散点图与直线方程的直线图，如图 1 所示：



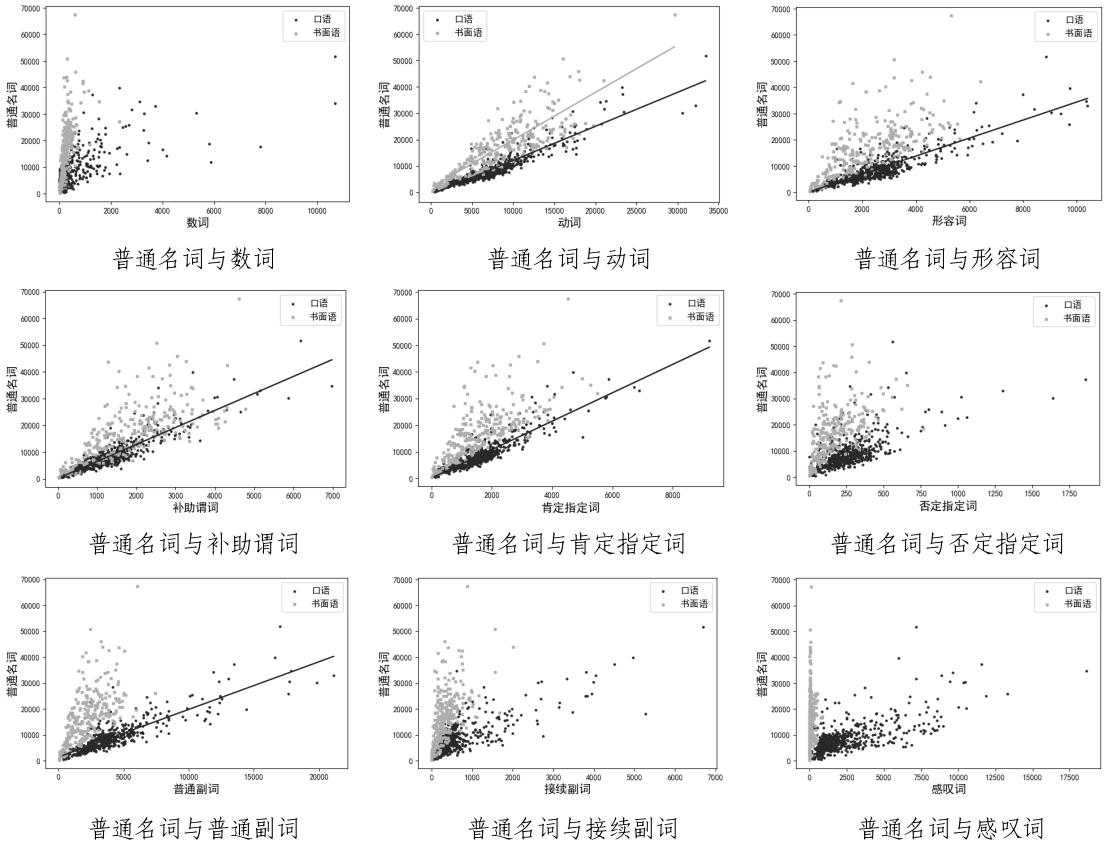
普通名词与固有名词



普通名词与依存名词



普通名词与代词



普通名词与数词　普通名词与动词　普通名词与形容词

普通名词与补助谓词　普通名词与肯定指定词　普通名词与否定指定词

普通名词与普通副词　普通名词与接续副词　普通名词与感叹词

图 1　朝鲜语普通名词与其他实词的散点图及直线图

查看表 1 和图 1，可以发现如下特点：

第一，表 1 显示，其他 12 类实词与普通名词的相关程度不高。与普通名词有高度相关关系的实词不多，口语中只有依存名词和动词 2 类（17%），书面语中只有依存名词 1 类（8%）。与普通名词有中度相关关系的实词，口语中有肯定指定词等 4 类（33%），书面语中只有动词 1 类（8%）；而与普通名词有弱相关关系或者不相关关系的实词，口语中有 6 类（50%），书面语中有 10 类（83%）。

第二，表 1 中的相关系数可按照降序排列如下：

口语：依存名词>动词>肯定指定词>补助谓词>形容词>普通副词>接续副词>代词>感叹词>否定指定词>数词>固有名词

书面语：依存名词>动词>肯定指定词>补助谓词>形容词>固有名词>普通副词>数词>接续副词>否定指定词>代词>感叹词

相关系数排序显示，朝鲜语口语和书面语中前 5 类实词的排序高度一致，后半部分体现出各自的特点。口语中，固有名词和数词排在倒数第一和第二的位置，而在书面语中排在第六和第八的位置；书面语中，感叹词排在倒数第一的位置，而在口语中排在第九的位置。

第三，观察图 1 中的直线，首先可以发现各条直线都是从下到上伸展，这说明普通名词与这些实词之间呈正相关关系，即随着其他实词出现频次的增长，普通名词的出现频次也跟

着线性增长。其次，在 12 类实词中，只有依存名词、动词与普通名词子图可以同时画出两条直线，这说明无论在口语还是在书面语中，这两类实词与普通名词都具有高度相关或者中度相关关系；口语中普通名词与形容词、补助谓词、肯定指定词、普通副词的相关关系均能画出直线图，表现出高度相关或者中度相关关系，而在书面语中画不出直线图，这说明普通名词与上述 4 类实词的相关关系在两个语料库中是不同的。再次，口语和书面语中其余实词与普通名词的相关关系均画不出直线图。最后，尽管口语和书面语中依存名词、动词与普通名词的相关关系均可画出直线，但由于各条直线斜率不相同，导致每张子图中两条直线都相互交叉，这表明依存名词、动词与普通名词的相关关系在口语和书面语中具有明显的差异；从这两张子图可以看到代表书面语的直线的斜率都大于代表口语的直线的斜率，因此，可以判定随着依存名词或动词使用量的增加，书面语中普通名词的使用量会明显大于口语中的。

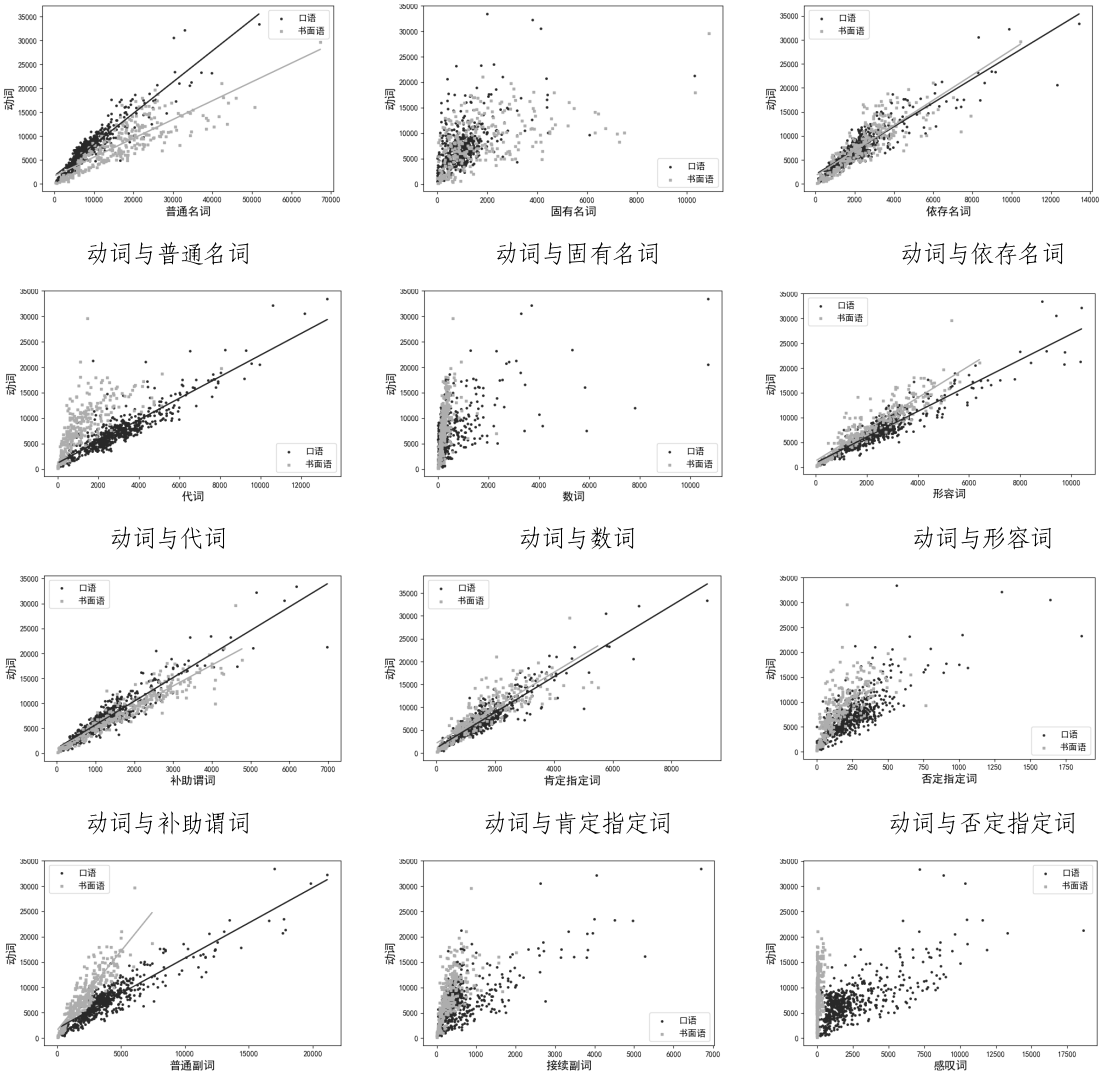
（二）动词

动词通常表示人或者事物的动作，后面加词尾后在句子中充当谓词，经常受普通副词的修饰（구본관외 2015:176）。动词和其他实词的相关系数、相关程度及直线方程如表 2 所示：

表 2　朝鲜语口语和书面语中动词与其他实词的相关系数、相关程度及直线方程

因变量与自变量	口语			书面语		
	相关系数	相关程度	直线方程	相关系数	相关程度	直线方程
动词与普通名词	0.920	高度相关	$y=1677.62+0.65x$	0.842	中度相关	$y=1657.45+0.39x$
动词与固有名词	0.540	不相关		0.521	不相关	
动词与依存名词	0.903	高度相关	$y=1957.90+2.49x$	0.839	中度相关	$y=1667.13+2.63x$
动词与代词	0.892	中度相关	$y=1171.85+2.12x$	0.707	弱相关	
动词与数词	0.558	不相关		0.485	不相关	
动词与形容词	0.929	高度相关	$y=868.50+2.60x$	0.917	高度相关	$y=1269.74+3.19x$
动词与补助谓词	0.927	高度相关	$y=990.10+4.72x$	0.938	高度相关	$y=916.46+4.19x$
动词与肯定指定词	0.915	高度相关	$y=1120.96+3.89x$	0.856	中度相关	$y=2158.40+3.87x$
动词与否定指定词	0.776	弱相关		0.721	弱相关	
动词与普通副词	0.914	高度相关	$y=1877.55+1.39x$	0.903	高度相关	$y=1562.04+3.12x$
动词与接续副词	0.694	弱相关		0.637	弱相关	
动词与感叹词	0.719	弱相关		0.380	不相关	

动词和其他实词出现频次的散点图与直线方程的直线图，如图 2 所示：



动词与普通名词　　动词与固有名词　　动词与依存名词

动词与代词　　动词与数词　　动词与形容词

动词与补助谓词　　动词与肯定指定词　　动词与否定指定词

动词与普通副词　　动词与接续副词　　动词与感叹词

图 2　朝鲜语动词与其他实词的散点图及直线图

根据表 2 中的相关系数、相关程度以及图 2 的散点图和直线图，可以发现如下特点：

第一，表 2 显示，动词与其他 12 类实词的相关程度明显高于普通名词。口语中，与动词有高度相关关系的有普通名词、依存名词、形容词、补助谓词、肯定指定词、普通副词等 6 类（50%），有中度相关关系的有代词 1 类（8%）；在书面语中，形容词、补助谓词、普通副词等 3 类（25%）与动词有高度相关关系，普通名词、依存名词、肯定指定词等 3 类（25%）与动词有中度相关关系。有弱相关关系和不相关关系的，口语中有固有名词、数词、否定指定词、接续副词、感叹词等 5 类（42%），书面语中有固有名词、代词、数词、否定指定词、接续副词、感叹词等 6 类（50%）。

第二，表 2 中的相关系数按照降序可排列如下：

口语：形容词>补助谓词>普通名词>肯定指定词>普通副词>依存名词>代词>否定指定词>感叹词>接续副词>数词>固有名词

书面语：补助谓词>形容词>普通副词>肯定指定词>普通名词>依存名词>否定指定词>代词>接续副词>固有名词>数词>感叹词

与普通名词相比，口语和书面语中动词与其他实词相关系数的排序有较大的波动，除了第四、六、十一位的排序相同外，其他排序均不相同。比如，口语对话带有较多感情色彩，因此口语中形容词排第一位，而在书面语中形容词排第二位。再比如，普通名词和普通副词的顺序也刚好颠倒，口语中普通名词与动词的相关系数大于普通副词与动词的相关系数，而在书面语中普通副词与动词的相关系数大于普通名词与动词的相关系数。动词与其他实词的相关系数中，我们还可以观察到书面语中动词与感叹词无相关关系，而在口语中动词与感叹词呈现弱相关关系。

第三，图 2 中，各条直线也都是从下到上伸展，这说明动词与这些实词之间呈正相关关系，即随着其他实词出现频次的增长，动词的出现频次也跟着线性增长。图 2 中，可以同时画出两条相关关系直线有动词与普通名词、依存名词、形容词、补助谓词、肯定指定词、普通副词等 6 类实词的子图，只能画出一条相关关系直线的仅有动词与代词 1 类实词的子图。固有名词、数词、否定指定词、接续副词、感叹词等其他 5 类实词与动词的子图都画不出直线图。观察画有两条直线的 6 张子图，也可以发现一些特点：从两条直线的位置来看，代表普通名词、补助谓词与动词相关关系的黑色直线在灰色直线的上方，而代表依存名词、形容词、肯定指定词、普通副词与动词相关关系的灰色直线在黑色直线的上方；从两条直线的夹角看，代表普通名词、普通副词与动词相关关系的两条直线的夹角较大，代表形容词和补助谓词与动词相关关系的两条直线的夹角较小，而代表依存名词、肯定指定词与动词相关关系的两条直线几乎重叠。位置信息说明变化大小，直线在上方表明其变化较大，大于在其下方的直线；夹角的大小说明变化幅度，即差异：夹角越大表示差异越大，夹角越小表示差异越小，如果重叠则表示口语和书面语没有明显差异。

（三）普通副词

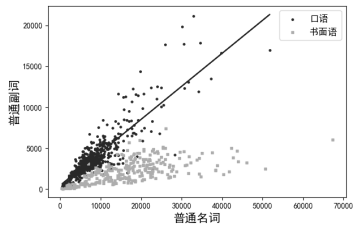
副词没有形态变化，可分为普通副词和接续副词（구본관외 2015:187）。普通副词和其他实词的相关系数、相关程度及直线方程如表 3 所示：

表 3　朝鲜语口语和书面语中普通副词与其他实词的相关系数、相关程度及直线方程

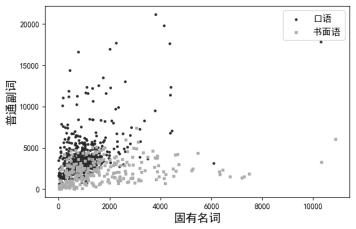
因变量与自变量	口语			书面语		
	相关系数	相关程度	直线方程	相关系数	相关程度	直线方程
普通副词与普通名词	0.861	中度相关	$y=454.07+0.40x$	0.638	弱相关	
普通副词与固有名词	0.486	不相关		0.306	不相关	
普通副词与依存名词	0.886	中度相关	$y=465.60+1.61x$	0.648	弱相关	
普通副词与代词	0.787	弱相关		0.807	中度相关	$y=946.24+0.98x$

普通副词与数词	0.608	弱相关		0.358	不相关	
普通副词与动词	0.914	高度相关	$y=-470.13+0.60x$	0.903	高度相关	$y=-15.73+0.26x$
普通副词与形容词	0.938	高度相关	$y=-360.53+1.72$x	0.931	高度相关	$y=90.55+0.94x$
普通副词与补助谓词	0.881	中度相关	$y=-30.73+2.95x$	0.917	高度相关	$y=67.09+1.18x$
普通副词与肯定指定词	0.814	中度相关	$y=307.09+2.28x$	0.836	中度相关	$y=418.62+1.09x$
普通副词与否定指定词	0.725	弱相关		0.784	弱相关	
普通副词与接续副词	0.815	中度相关	$y=2150.76+2.97x$	0.616	弱相关	
普通副词与感叹词	0.753	弱相关		0.551	不相关	

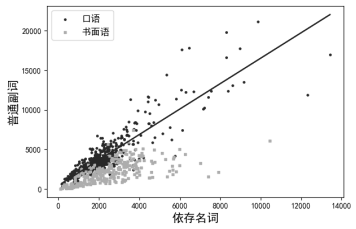
普通副词与其他实词出现频次的散点图与直线方程的直线图，如图 3 所示：



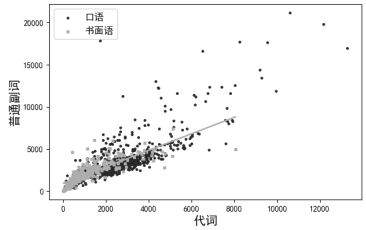
普通副词与普通名词



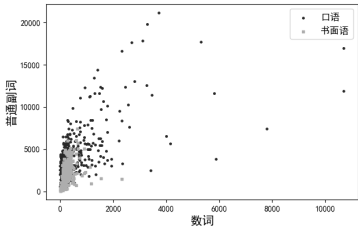
普通副词与固有名词



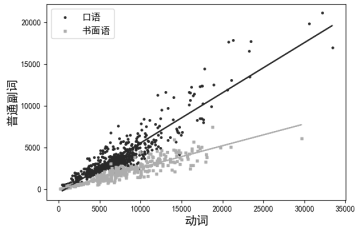
普通副词与依存名词



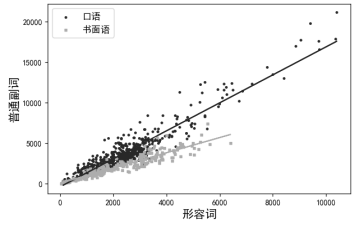
普通副词与代词



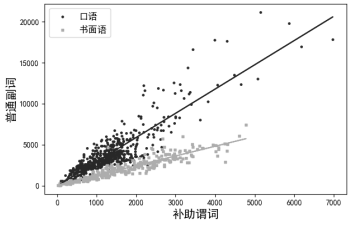
普通副词与数词



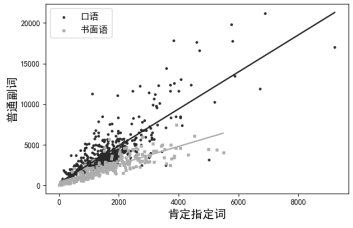
普通副词与动词



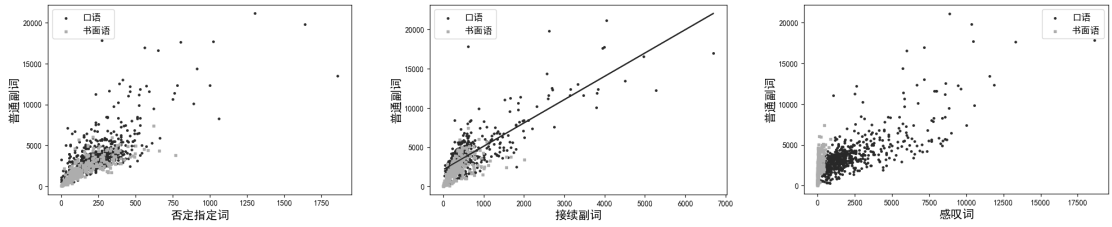
普通副词与形容词



普通副词与补助谓词



普通副词与肯定指定词



普通副词与否定指定词　　普通副词与接续副词　　普通副词与感叹词

图 3　朝鲜语普通副词与其他实词的散点图及直线图

根据表 3 中的相关系数、相关程度以及图 3 中的散点图和直线图，可以发现如下特点：

第一，口语中与普通副词呈现高度相关关系的实词有形容词和动词 2 类（17%），呈现中度相关关系的实词有依存名词、补助谓词、普通名词、接续副词、肯定指定词等 5 类（42%），呈现弱相关或不相关关系的有代词、感叹词、否定指定词、数词、固有名词等 5 类（42%）；书面语中与普通副词呈现高度相关关系的实词有形容词、补助谓词、动词等 3 类（25%），呈现中度相关关系的有肯定指定词、代词等 2 类（17%），呈现弱相关或不相关关系的有否定指定词、依存名词、普通名词、接续副词、感叹词、数词、固有名词等 7 类（59%）。比较普通副词与其他实词的相关程度，从圆括号里的百分数可以看出，口语中相关程度高的实词多于书面语中的。

第二，表 3 中的相关系数按照降序可排列如下：

口语：形容词>动词>依存名词>补助谓词>普通名词>接续副词>肯定指定词>代词>感叹词>否定指定词>数词>固有名词

书面语：形容词>补助谓词>动词>肯定指定词>代词>否定指定词>依存名词>普通名词>接续副词>感叹词>数词>固有名词

普通副词的主要功能是修饰谓词，因此，不管是口语还是书面语中，与普通副词高度相关的实词都是形容词排第一位；动词在口语中排第二位，而在书面语中则排第三位。书面语中补助谓词与普通副词具有高度相关关系，而在口语中二者仅具有中度相关关系。口语中体词中的依存名词、普通名词的排序靠前，书面语中这些体词的排序靠后。

第三，通过观察图 3 中的直线可以发现，各条直线也都是从下到上伸展，这说明普通副词与这些实词之间呈正相关关系，即随着其他实词出现频次的增长，普通副词的出现频次也跟着线性增长。查看该图还可以发现，同时能画出两条直线的有动词、形容词、补助谓词、肯定指定词等与普通副词 4 张子图；仅能画出一条直线的，有口语中普通名词、依存名词、接续副词与普通副词 3 张子图，以及书面语中代词与普通副词 1 张子图。再观察同时能画出两条直线的 4 张子图，首先各张子图中代表口语的黑色直线都在代表书面语的灰色直线的上方，而且夹角的开口度都比较大。这说明虽然口语和书面语中普通副词和动词、形容词、补助谓词、肯定指定词的相关关系上都呈现高度相关或者中度相关关系，但是在各自的使用上有较大的差异。

（四）感叹词

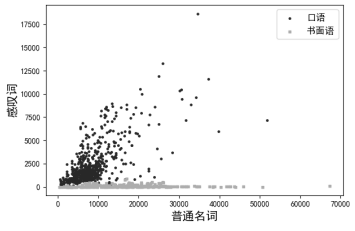
感叹词通常表示说话者的感觉或者意志，与冠形词和副词相同，没有形态变化，但是不修饰其他词，在句子中独立使用（구본관외 2015:191）。

由于感叹词与其他实词的相关程度较低，因此未能得出直线方程，所以表 4 中只列出感叹词与其他实词的相关系数和相关程度。

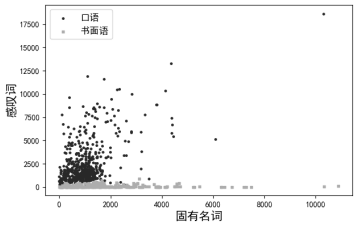
表 4　朝鲜语口语和书面语中感叹词与其他实词的相关系数和相关程度

因变量与自变量	口语		书面语	
	相关系数	相关程度	相关系数	相关程度
感叹词与普通名词	0.678	弱相关	−0.006	不相关
感叹词与固有名词	0.532	不相关	0.012	不相关
感叹词与依存名词	0.654	弱相关	0.034	不相关
感叹词与代词	0.686	弱相关	0.553	不相关
感叹词与数词	0.453	不相关	0.123	不相关
感叹词与动词	0.719	弱相关	0.380	不相关
感叹词与形容词	0.722	弱相关	0.373	不相关
感叹词与补助谓词	0.700	弱相关	0.385	不相关
感叹词与肯定指定词	0.661	弱相关	0.236	不相关
感叹词与否定指定词	0.589	不相关	0.276	不相关
感叹词与普通副词	0.753	弱相关	0.551	不相关
感叹词与接续副词	0.546	不相关	0.180	不相关

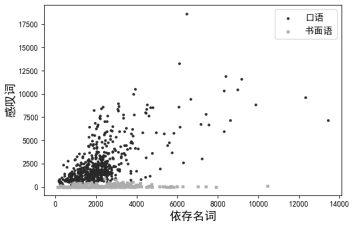
根据感叹词和其他实词的出现频次同样画出了散点图，如图 4 所示：



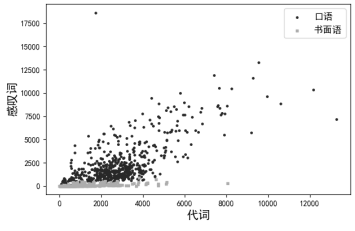
感叹词与普通名词



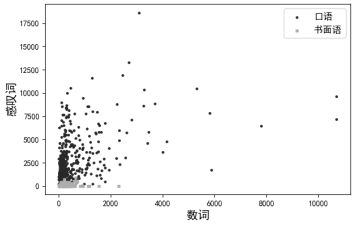
感叹词与固有名词



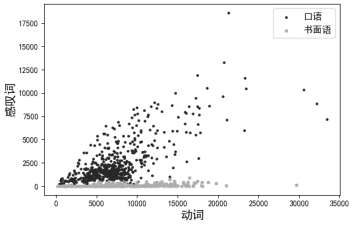
感叹词与依存名词



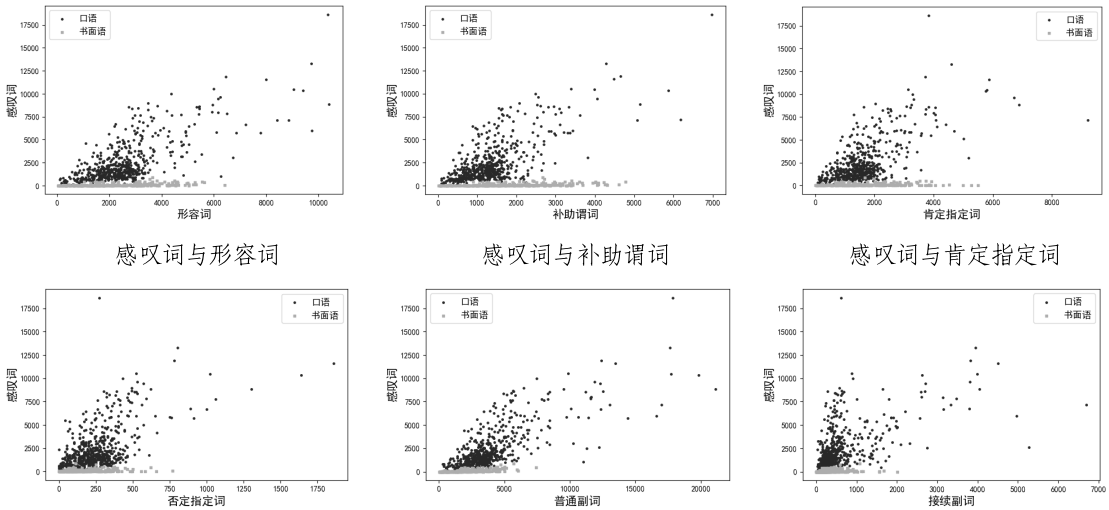
感叹词与代词



感叹词与数词



感叹词与动词



感叹词与形容词　　感叹词与补助谓词　　感叹词与肯定指定词

感叹词与否定指定词　　感叹词与普通副词　　感叹词与接续副词

图 4　朝鲜语感叹词与其他实词的散点图

查看表 4 和图 4，可以看出感叹词与其他实词没有高度相关或中度相关关系。虽然在口语中能看到随着各类实词出现频次的增长，感叹词的出现频次也在增长，但是其规律不明显。与口语相比，能明显看出书面语的特点。随着各类实词出现频次的增长，感叹词的出现频次在 0 的位置上，与 x 轴几乎平行，这说明在每个书面语样本中感叹词的出现频次都趋于 0。

五　结　语

本文的研究表明，一元线性回归分析可以用于语言学研究。针对某一语体的语料库，把一个语言特征当作自变量，把另一个语言特征当作因变量，并通过画散点图观察两个变量之间是否存在线性相关关系；如果存在线性关系，那么是正相关关系还是负相关关系。通过计算相关系数可以判定两个变量之间的相关关系强度，即高度相关、中度相关、弱相关还是不相关；对于高度相关和中度相关的两个变量可以求得直线方程，通过观察该直线的斜率还可以判断因变量和自变量之间的变化幅度。

本文的研究也表明，朝鲜语口语和书面语中出现的实词间的相关关系，与实词的组合能力和句法功能密切相关。实词的组合能力、句法功能越强，实词间的相关程度越高；反之，则越低，如普通名词与动词的组合能力、句法功能强于普通名词、动词与感叹词的组合能力、句法功能，普通名词与动词呈现高度相关关系，而普通名词、动词与感叹词呈现弱相关关系。本文实词间相关关系的研究结论与普通语言学获取的结论一致，说明一元线性回归分析用于词类间的相关关系研究是行之有效的。

本文的研究结果可用于文本自动归类。本文针对朝鲜语口语和书面语中出现的实词进行了相关关系分析，计算出各类实词间的相关系数和直线方程。对朝鲜语口语和书面语文本进行自动归类时，可将实词间相关系数和直线方程作为分类标准或分类参考，让计算机系统对未知的文本进行自动归类。

参考文献

[1] 邓成竹. 2017.《基于一元线性回归模型的航班业载与油量关系研究》,《民航学报》第 1 期.

[2] 姜道旭、吴文琴. 2019.《基于一元线性回归分析扬州市城镇和农村人均可支配收入与消费支出》,《现代营销》(经营版)第 12 期.

[3] 林志伦. 2018.《基于 R 语言的一元线性回归模型在经济变量间的应用》,《济源职业技术学院学报》第 2 期.

[4] 麦莉莉. 2019.《预测数据 精准资助——一元线性回归模型的研究》,《广西质量监督导报》第 3 期.

[5] 夏伯忠主编. 1993.《新会计大辞典》,北京:中国商业出版社.

[6] 刘 颖. 2014.《计算语言学》(修订版),北京:清华大学出版社.

[7] 김혜림、김동재. 2018.《여러 개의 단순선형회귀모형에서 순차기울기를 이용한 평행성검정》,《한국데이 터정보 과학회지》29(4): 873-884(金慧琳、金东才. 2018.《利用多个一元线性回归模型序列斜率的平行线测定》,《韩国数据信息科学会杂志》第 29 卷第 4 期第 873-884 页).

[8] 박영호、김영화. 2015.《단순선형회귀분석과에 지검출기에 근거한 영상잡음의 분산추정》,*Journal of the Korean Data Analysis Society* 17(1): 219-228.(朴永浩、金英花. 2015.《一元线性回归分析和边缘检测下的影像杂音分散推测》,《韩国数据分析协会联盟》第 17 卷第 1 期第 219-228 页).

[9] 구본관외. 2015.《한국어문법총론(1)》,서울:집문당(具本宽等. 2015.《韩国语语法总论(1)》,首尔:集文堂).

Correlation between Content Words in Spoken and Written Korean: A Univariate Linear Regression Analysis

LU Xinghua and JIN Jing

[Abstract] Univariate linear regression is commonly employed to study the linear relationship between two variables, the independent variable *x* and the dependent variable *y*. It is an approach for prediction by way of establishing the linear regression equation of *x* and *y* according to their correlation. Based on large-scale spoken and written Korean corpora, this paper uses a univariate linear regression model to calculate the correlation coefficients between each of the four categories of content words and the other twelve categories, sorts the correlation coefficients, and then establishes a univariate linear regression model of the correlation above the moderate, which is used to compare spoken and written Korean.

[Keywords] Korean language spoken language written language univariate linear regression correlation analysis

(通信地址:卢星华 133002 延吉 延边大学朝汉文学院/融合学院
金 静 133002 延吉 延边大学朝汉文学院)

【本文责编 胡鸿雁】