

中图分类号:

UDC:

学校代码: 10055

密级: 公开

南开大学

硕士学位论文

基于依存句法的高频动名兼类词研究题

A Study on High-frequency Verb-noun Heterosemous
Words based on Dependency Syntax

论文作者 杨牧

指导教师 蔡言胜 副教授

申请学位 文学硕士

培养单位 汉语言文化学院

学科专业 语言学及应用语言学

研究方向 对外汉语教学与研究

答辩委员会主席 郭继懋

评阅人 郭继懋 张万禾

南开大学研究生院

二〇二三年五月

南开大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下进行研究工作所取得的研究成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名： 杨牧

2023年5月10日

非公开学位论文标注说明

(本页表中填写内容须打印)

根据南开大学有关规定，非公开学位论文须经指导教师同意、作者本人申请和相关部门批准方能标注。未经批准的均为公开学位论文，公开学位论文本说明为空白。

论文题目			
申请密级	<input type="checkbox"/> 限制(≤2年)	<input type="checkbox"/> 秘密(≤10年)	<input type="checkbox"/> 机密(≤20年)
保密期限	20 年 月 日至 20 年 月 日		
审批表编号		批准日期	20 年 月 日

南开大学学位评定委员会办公室盖章(有效)

注：限制★2年(可少于2年)；秘密★10年(可少于10年)；机密★20年(可少于20年)

南开大学学位论文使用授权书

本人完全了解《南开大学关于研究生学位论文收藏和利用管理办法》关于南开大学(简称“学校”)研究生学位论文收藏和利用的管理规定,同意向南开大学提交本人的学位论文电子版及相应的纸质本。

本人了解南开大学拥有在《中华人民共和国著作权法》规定范围内的学位论文使用权,同意在以下几方面向学校授权。即:

1. 学校将学位论文编入《南开大学博硕士学位论文全文数据库》,并作为资料在学校图书馆等场所提供阅览,在校园网上提供论文目录检索、文摘及前 16 页的浏览等信息服务;

2. 学校可以采用影印、缩印或其他复制手段保存学位论文;学校根据规定向教育部指定的收藏和存档单位提交学位论文;

3. 非公开学位论文在解密后的使用权同公开论文。

本人承诺:本人的学位论文是在南开大学学习期间创作完成的作品,并已通过论文答辩;提交的学位论文电子版与纸质本论文的内容一致,如因不同造成不良后果由本人自负。

本人签署本授权书一份(此授权书为论文中一页),交图书馆留存。

学位论文作者暨授权人(亲笔)签字: 杨牧

2023年5月10日

南开大学研究生学位论文作者信息

论文题目	基于依存句法的高频高频动名兼类词研究				
姓名	杨牧	学号	2120201419	答辩日期	2023年5月20日
论文类别	博士 <input type="checkbox"/> 学历硕士 <input checked="" type="checkbox"/> 专业学位硕士 <input type="checkbox"/> 同等学力硕士 <input type="checkbox"/> 划 <input checked="" type="checkbox"/> 选择				
学院(单位)	汉语言文化学院		学科/专业(专业学位)名称		语言学及应用语言学
联系电话	18722120059		电子邮箱	hyym@mail.nankai.edu.cn	
通信地址(邮编):	南开大学津南校区文科(300381)				
非公开论文编号		备注			

注:本授权书适用我校授予的所有博士、硕士的学位论文。如已批准为非公开学位论文,须向图书馆提供批准通过的《南开大学研究生申请非公开学位论文审批表》复印件和“非公开学位论文标注说明”页原件。

摘要

作为汉语研究的基本问题之一，汉语的词类体系建设一直存在争议，争议的一个关键点就在动名兼类词。所谓动名兼类词，是指兼有动词性功能和名词性功能的词。信息技术的进步给动名兼类问题提供了更广阔的视野，定量研究、基于标注语料库的研究、跨学科研究等范式发挥的作用越来越显著。语言是个“人驱复杂适应系统”，结合更为先进的方法，计量动名兼类词在语言系统中的特征，有助于进一步探索汉语的结构模式和演化规律。

基于上述背景，本研究以动名兼类词为对象，基于依存树库，讨论了三个问题：（1）如何计量动名兼类词的语法功能？（2）如何计量动名兼类词与动词、名词的相似度？（3）所得指标如何应用于第二语言习得研究？研究成果如下：

（1）基于概率配价模式理论，从词和依存关系探讨了如何获取动名兼类词语法特征。前者关注词能与哪些词组合，后者关注词能在句中充当哪些句法成分。经统计，动名兼类词的主要功能是支配名词、小品词、动词、介词和从属动词、小品词、名词、助动词，以及支配修饰语、宾语、主语、补语和做宾语、修饰语。

（2）通过余弦相似度，计算出了动名兼类词的动名相似度。结果表明，73.4%典型动名兼类词的语法功能与动词和名词都有0.5以上的相似度，并且其动词相似度和名词相似度成负相关，这证实了“动名连续统”的假设。

（3）探讨了相似度指标在第二语言习得研究中的应用价值。本研究计算了中介语与目的语中动词、名词的相似度。结果表明，两词类配价整体上习得难度类似，做支配词时动词配价习得难，做从属词时名词配价习得难。配价习得呈现出非线性、动态特征，动词在80分阶段前后变化最显著，名词在60分阶段前后变化最显著。配价习得并非从零开始，且最高阶段也无法到达母语者水平。

本文探讨、验证并应用了一种计量动名兼类词语法功能的方法，计量了动名兼类词的语法功能，计算了其于名词和名词的相似度，还将相似度指标应用于第二语言习得研究。研究结果对于探讨动名兼类词在词类体系中的地位，探讨语言系统的整体与局部特性，促进词类研究的定量转向具有一定价值。同时，本文丰富了词（词类）计量的方法，能够为未来的语言计量研究提供参考。以上结果表明，基于依存树库的计量方法能在动名兼类词研究中发挥良好作用。

关键词：动名兼类词；相似度；词类；概率；计量语言学

Abstract

As a basic issue in Chinese linguistics, the construction of the Chinese word class system has always been controversial, and a key point is the verb-noun heterosemy, which refers to words that have both the grammatical function of verbs and nouns. The advancement of information technology has provided a broader perspective on this issue, and paradigms such as quantitative research, research based on annotated corpora, and interdisciplinary research have played an increasingly significant role. Language is a "human-driven complex adaptive system". Combining with more advanced methods, measuring the characteristics of verb-noun heterosemous words will help to further explore the structural patterns and evolution laws of Chinese.

Based on the background, three questions are discussed in this study: (1) How to measure the grammatical function of words and word classes? (2) How to measure the similarities between words or word classes? (3) How to apply similarity indicators to second language acquisition? The specific results are as follows:

(1) Based on the Probabilistic Valency Patterns Theory, this study discusses the method of obtaining the grammatical features from the perspectives of words and dependency relations. The former focuses on which words a word can be combined with, and the latter focuses on which syntactic components a word can serve as in a sentence. According to our statistics, the main functions of verb-noun heterosemous words are to dominate nouns, particles, verbs, prepositions and subordinate verbs, particles, nouns, auxiliary verbs, and to dominate modifiers, objects, subjects, complements, and to be objects and modifiers.

(2) Through the Cosine Similarity, similarities with verbs and nouns of verb-noun heterosemous words are calculated. The results show that the grammatical function of 73.4% of typical verbs and class words has a similarity of more than 0.5 with both verbs and nouns, and the similarities with verbs and nouns are negatively correlated, which

confirms the hypothesis of "verb-noun continuum".

(3) Discuss the application of the similarity index in second language acquisition research. This study calculates the similarity of verbs and nouns in the interlanguage and the target language. The results show that the acquisition difficulty of valency is similar on the whole. It is more difficult to acquire verb valency when it acts as a dominant word, and more difficult to acquire noun valency when it acts as a subordinate word. The acquisition of valency presents nonlinear and dynamic characteristics. Verb valency changes most significantly before and after 80 stages, and noun valency changes most significantly before and after 60 stages. Valency acquisition does not start from scratch, and the highest stage cannot reach the level of native speakers.

This paper discusses, verifies and applies a method of measuring the grammatical function of verb-noun heterosemous words, calculates their similarities with verbs and nouns, and discusses the application of the similarity index to second language acquisition. The research results are of certain value for clarifying the status of verb-noun heterosemous words in the word class system, exploring the overall and partial characteristics of the language system, and promoting the quantitative turn of word class research. What's more, this paper enriches the method of word (word class) measurement, highlights the probabilistic nature of language, and can provide theoretical and methodological reference for future linguistic quantitative research. The above results show that the econometric method based on the tagged corpus under the framework of dependency grammar can play a good impressive role in the study of verb-noun heterosemous words.

Key words: Verb-noun Hetersemy; Similarity; Word Class; Probability; Quantitative Linguistic

目录

第一章 绪论	1
第一节 研究现状.....	1
1.1.1 词类划分和动名兼类词的界定	1
1.1.2 动名兼类的理论研究	4
1.1.3 动名兼类的应用研究	6
1.1.4 动名兼类研究评述	8
第二节 研究问题和研究意义.....	10
第二章 资源和方法	12
第一节 资源.....	12
第二节 研究方法.....	15
2.2.1 分析工具	15
2.2.2 计量指标与计算方法	15
第三章 动名兼类词的语法功能计量	19
第一节 动名兼类词的选择.....	19
第二节 基于词的语法功能计量.....	20
3.2.1 以类为单位的语法功能统计.....	21
3.2.2 以词为单位的语法功能统计.....	22
3.2.3 本节小结	23
第三节 基于依存关系的语法功能计量.....	24
3.3.1 以类为单位的语法功能统计.....	27
3.3.2 以词为单位的语法功能统计.....	28
3.3.3 本节小结	29
第四节 本章小结.....	30

第四章 动名兼类词的动名相似度计量	31
第一节 动词和名词的选择.....	31
第二节 基于词的语法功能计量.....	31
第三节 基于依存关系的语法功能计量.....	34
第四节 动名兼类词的动名相似度计量.....	37
第五节 本章小结.....	41
第五章 基于相似度的汉语中介语动词、名词配价习得情况计量 .	42
第一节 资源和方法.....	42
第二节 动词配价和名词配价的习得情况计量.....	44
第三节 讨论.....	46
第四节 本章小结.....	51
第六章 结论	52
第一节 主要内容总结.....	52
第二节 不足与展望.....	53
参考文献.....	55
附录 A 统计信息附表	67
附录 B 动名兼类词词表及相似度.....	76
附录 C 动名兼类词语料例证	84
致谢.....	95
个人简历.....	96

第一章 绪论

就人类语言的普遍特征而言，动词和名词是最基本的两个词类，绝大多数语言都有动词和名词的区分。但它们并非泾渭分明。即使在有明确形态标记的印欧语中也存在动名兼类现象，在词类范畴非显性的汉语中动名兼类现象则更为常见。动名兼类现象在语言学理论研究和语言教学、语言处理等应用领域中都催生了不少问题。尽管动名兼类的研究历史并不短暂，但这些问题仍然存在。在数智时代，以定量手段研究动名兼类词具有重要意义。本章先回顾以往的研究，重点评述其中的定量研究，最后提出研究问题和阐明研究意义。

第一节 研究现状

本节先分述各家界定兼类词的标准，然后回顾以动名兼类为对象的理论研究和应用研究，最后就以上研究在资源和方法等方面的发展趋势进行评述。

1.1.1 词类划分和动名兼类词的界定

词类这个概念是印欧语言学界引入的。印欧语多为形态丰富的屈折语，大多数时候单凭词的形态就能够区别出词类。但现代汉语是孤立语，几乎没有形态，从词本身上看不出什么问题来，因此依据什么区分词类就成了问题。就此，学者们主要提出了三种依据：意义、形态、语法功能。

根据意义划分词类可以分为两种情况。一种是纯粹依靠意义分类，王力（1943）、吕叔湘（1947）都曾辨析意义得出词类。另一种是考虑意义的同时也参照其他标准，马建忠（1898）、黎锦熙（1924）用词义划分出词类系统，在处理一些“假借”或“活用”情况时又参照了词语的句法功能；高名凯（1948）根据词语的语法意义把词分为实虚两大类。其他学者在划分词类时也或多或少参考了意义。

根据意义区分虚实是有必要的，但在进行实词和虚词内部的词类划分时就不妥当了。根据意义分类的最大弊病在于没有看到词语意义和句法表现不是完全等同的，词义是模糊的，难以完全明晰。

根据形态来划分词类。高名凯（1948）认为形态是词类的本质划分依据，而

汉语没有形态，因此得出汉语无词类的结论。邢福义（2002）也认为形态是划分词类的依据，但他说的形态与印欧语中的形态不同，指的是构词和构形的语法形式。汉语中构词手段有前缀和后缀，如“老”作为前缀时能够视作名词的标志，“化”作为后缀能够视作动词的标志；构形手段有重叠和黏附，如“观看”能重叠和加“着”、“了”、“过”表示语法意义，而“观众”则不能。

现代汉语的形态成分已经几乎消失，显然形态并不适合用来区分词类。邢福义给的形态是有价值的，但区别能力仍然不够显著。

根据语法功能划分词类。语法功能指的是词能够出现的句法环境，实质是词的组合能力。陈承泽（1922）、金兆梓（1922）、方光焘（1939）、陈望道（1978）、胡明扬（1999）等都认为词的语法功能才是词类划分的主要依据；赵元任（1955、1979、2022）根据词的语法功能系统划分了词类；朱德熙（1961）也强调词类的本质是分布。

在有形态的语言中，形态无疑是最有效的词类分类手段，但形态反映的其实就是词的语法功能。因此语法功能的解释能力是最为令人信服的，这一结论也被广泛认可。词类划分多是先根据有无实际意义区分出实词和虚词，然后再根据语法功能的不同区别出具体的词类。尽管划分词类的大原则确定了下来，但实际的词类划分里仍出现了很多问题。兼类，特别是动名兼类，就是最严重的老大难问题之一（胡明扬，1996；俞士汶等，2020）。

兼类词，依安华林（2005）总结的主流现代汉语教材（黄伯荣、廖序东，2002；胡裕树，1995；邢福义，1991；张斌，2002），有以下定义：（1）经常具备两种或两种以上词类的主要语法功能；（2）在具体的语法环境中不存在兼类；（3）兼类词同音同形；（4）词义有联系；（5）兼类词数量上只占据少数；（6）兼类的主要类型为动、名、形之间的兼类。简单来说就是那些兼有动词和名词功能且在意义上有联系的词。但各家界定兼类词的标准有诸多细节上的差异。从总体看，动名兼类词的界定大致有两条主要原则：（1）语法功能兼属动名两类且语义有联系；（2）兼类用法有一定使用频率且兼类词规模不能太大。但各家对这两条原则的具体看法各有不同。

先说语法功能标准。钟棫（1980）给出了三条标准：（1）可受名量词限定；（2）可受一般形容词修饰；（3）可受名词修饰。胡明扬（1995a，1995b）认为判断一个词是否是动名兼类词，要从功能转移、使用范围和使用频率上来看。功能转移上，用了三条标准鉴定动词中兼有名词词性的词：（1）能否带名量词（除了“种”和“个”）；（2）能否直接做“有”的宾语并受前置动量词修饰；（3）能否直接受

名词修饰且受前置动量词修饰。杨同用（2008）肯定了胡明扬（1995a, 1995b）提出的标准。李立成（1999）认为由动词变为名词，而没有任何标记，在指称上可以有自指和转指两类的词是动名兼类词。动名兼类词可以进入“的”字自指结构，如“我们的合作很愉快”与“我们合作得很愉快”中的“合作”。郭锐（2002）还提出了两条标准：（1）可以受名词直接修饰；（2）可以作准谓宾动词的宾语。

史有为（1991、1993）提出“柔性处理”汉语的词类问题，先找出词类的典型成员，再计算边缘成员与其在分布特征上的联系值，以其联系值大小断定该词的词类归属。卢英顺（1998）给出计算动词“近似度”的五条标准及计算公式，即满足的标准的数量与标准总量的比值。袁毓林（1995、2005a、2005b）介绍了原型理论视角下的词类问题，探索了词类系统的各个词类的属性。袁毓林（2000、2005a、2005b）把典型成员的分布特征作为词类范畴的属性，给予每个属性分值，构建出一个词类评分系统，能够计算出某个词对各个词类的隶属度。隶属度越大，就越符合某个词类。如“当年”，时间词隶属度为 0.9，状态词隶属度为 0.7，动词隶属度为 0.4，副词隶属度为 1。以上研究以语法功能为标准，计算出了词对词类的隶属度，实际上也能被视为用语法功能区分兼类。

再说语义联系。对于兼类词内部的区别，郭锐（1999）提出了同型兼类词和异型兼类词，前者指词义相同的兼类词，如“小时”（量词、名词）、“区别”（动词、副词）；后者指词义不同的词，如“领导”（动词、名词）、“锁”（动词、名词）。郭文认为异型兼类词应当处理为兼类词，但同型兼类词则不一定要处理为兼类词。安华林（2005）也认为语义没有明显转变的，不处理为兼类词。聂志平、杨文全（2016）则提出相反意见，认为词的同—性来看，意义相近但语法功能不同的词不应该被视作兼类，而应该视作不同的词，实际上否定了对于兼类词的传统看法。

这些标准中“能否直接受名词修饰”和“能否受名量词修饰”取得了广泛共识，并且也经受住了实证检验（徐艳华、朱诗瑶，2020）。

最后再看频率和规模标准。胡明扬（1995a, 1995b）虽然提出了使用频率是兼类词判定的重要标准，但并未做出详细分析。郭文认为异型兼类词应当处理为兼类词，但同型兼类词则不一定要处理为兼类词，处理的原则既要使词类简单，兼类词的数量不能太多（这点对应于胡文中的使用频率），也要尽可能让词类内不同词的语法功能单一。就此原则，“研究”、“生产”、“管理”等词都应被划归为动名兼类词。翟淑英（2000）也肯定了使用频率量化在兼类问题中的地位。

高航（2009）认为传统上“词有定类”（吕叔湘，1954、1984）的观点存在问题。兼类问题的处理应当尊重语言事实，而不是依靠逻辑推论。此观点直指兼类

词划分中的使用频率原则——兼类词应当数量较少（郭锐，1999；安华林，2005；陆俭明，1994）。出于这样的考虑，“研究”这类数量庞大（郭文和陆文中的统计分别为 44%、31%）的兼有动词名词功能的词应当被处理为动词子类的“名动词”，而不应该处理为兼类。高航（2009）批评了这种“太理想化”的做法，认为应该以语言事实为准，放弃数量考虑也不会导致“词无定类”，因为词可能出现的结构终究是有限的。

大多数学者都认同将兼类词数量控制在小范围内的做法，经实证统计（胡明扬，1995b；安华林，2005；方清明，2010），兼类词也确实规模有限。

总而言之，动名兼类词是音形相同，兼有动词和名词的语法功能，语义有联系的词，并且不同用法均有一定规模的使用，总体数量较少。

1.1.2 动名兼类的理论研究

动名兼类问题的“老大难”主要集中在两点。第一，出现主宾语位置上的动词应当看作名词还是原词类？第二，出现在“N 的 X”中 X 位置上的动词应当看作名词还是原词类？为了回答以上两个问题，除却前文已经介绍的动名兼类词界定的相关研究，学者们从三个子主题进行了探讨：名物化和名词化、动词化、动名词类关系。

首先是名物化和名词化的探讨。

对于动词做主语和宾语的问题，各家看法如下。马建忠（1898）认为是“词类假借”，陈承泽（1922）、王力（1943）认为是“词类活用”。黎锦熙（1924）提出“散动词”，认为动词在特定句法环境时可以当其他词类看待，比如在做主宾时可以看作名词。这种看法明显受到印欧语（尤其是英语）中动名词、不定式等结构的影响。高名凯（1948）虽认为汉语无词类，但实际上也对汉语的实词分了类，不过是分为“具有动词功能的词”、“具有名词功能的词”等，且一个词可以同时具有几类词的功能。这种观点下两个问题中的动词展现的正是名词的功能。“暂拟汉语教学语法系统”（1956）提出了这些动词出现了名物化，类似的说法还有名词化（高航，2007），即动词或形容词失去一些原本的功能，获得一些名词的功能，如“这本书的出版”中的“出版”。朱德熙、卢甲文、马真（1961）批评了名物化，认为在语法意义上对“事物范畴”的区分不准确，名物化只是句子成分定类论的组成部分。在语法功能上名物化并没有获得名词的语法功能，而是名词和动词本来就都具有的共同功能。程工（1999）、项梦冰（1991）从向心结构出发，认

为“这本书的出版”整体表现出名词性，其中心词“出版”也应是名词。胡裕树、范晓（1994）认为名物化是发生在语义平面上的；张伯江（1994）承认此类动词谓词性质减弱，名词性质增强；杨成凯（1991）也有类似张伯江的观点，但他明确对名物化的态度，认为此类词部分已经成为名词，部分在向名词转化；李宇明（1986）、郭锐（2002）从表述功能角度对名物化进行解释。刘顺（2003）探讨了动词的零形式名词化，高航（2019）从认知角度探讨了回指构式中单音节动词的名词化。朱永生（2006）认为从语义功能上看，名词化指的就是把某个过程或特征看作事物，而词性转换只是这种现象得以实现的一种方式。王冬梅（2003）认为动词失去独立性和名物化间存在密切联系，认为动词的名物化是个程度问题。何元建、王玲玲（2007）认为它名物化结构是从相应的谓词结构转换而来的。金立鑫（2019）认为“这本书的出版”结构中的“出版”是名词的后置谓词性定语。

名物化和名词化仍然是研究的热点，并且仍未讨论出一个让众人都信服的观点，仅从知网 CNKI 以“名物化”为关键词检索就有 1919 条文献。我们仅仅在此举了部分有代表性的文献以管中窥豹。

既然动词能产生名词的用法，名词自然也可以有类似变化，就是所谓的动词化（朱永生，2006）、名转动（宋作艳，2013、2018）、名词动用（王勇、徐杰，2018），即名词获得了一些动词的功能，如“百度一下”中的“百度”。

徐盛桓（2001）提出了“名动互含”假说，认为动词名词有共通的语义内容，认为这种语义上的共通是动词化的基础。在“名动互含”假说基础上，高芳、徐盛桓（2000a、2000b）从语用角度以推理方式进行了验证。朱永生（2006）认为动词化是一种语法隐喻（Halliday, 1985），在语义层指的是把不是过程的成分当作过程来处理，在形式层指的是用动词表达一般不是由动词表达的意义，并给动词化分了四个类别：把时间关系看作过程、把因果关系看作过程、把条件看作过程、把让步关系看作过程。周领顺（2000）对比了英汉名-动类词的转化机制和发展趋势。王勇、徐杰（2018）探讨了诗歌中的动词化，并认为诗歌是一种“语言特区”，名词动用能取得诗性效果。

宋作艳（2013）从生成词库理论出发，考察了汉语中十类常见的名转动现象，认为名转动是种逻辑转喻，可以用事件强迫来解释。宋作艳（2018）依托物性结构理论，探讨了名词转动词的语义基础，认为名转动是物性结构和论元结构互动的结果。骆琳、穆柔云（2019）从物性结构角度探讨了工具类名词的动词化，并认为不同词动词化的程度有轻重之分。

也有诸多学者从认知语言学角度出发对动词化现象进行了探索，王冬梅

(2001)认为动词化实际上是一种概念转喻;高航(2008)认可了概念转喻的观点,并讨论了汉语中表示不同语义角色的名词发生动词化的频率和成因;周统权、杨静(2010)以“山寨”为对象进行个案研究,认为“山寨”表面是名词,但实际上在不停向动词游移,表现出明显的动词化,认为社会心理认知中的追求时尚是动词化的动因。

特别值得说明的是,当前研究的一大热点是对动名关系的宏观理论探讨。沈家煊(2007、2009a、2009b、2010、2016、2023)提出“名动包含说”。这种观点认为汉语词类分布与印欧语不同,印欧语词类分立且有交叉,而汉语词类则是同心圆式的包含:动词包含在名词里面,因而动词拥有名词的语法功能也不足为奇。学者们对“名动包含”的态度各有不同,肯定批评兼有。张伯江(2012)肯定了“名动包含”,认为它有限地维护了名动有别,可以大大简化汉语的树形结构和句法操作。孙崇飞(2022)从逻辑推论上指出“名动包含”不存在逻辑错误,因而要承认汉语“名包动”的事实。

反对的意见也层出不穷。陆俭明(2022a、2022b)认为“名动包含”存在逻辑错误,且指出其理据(汉语动词有指称性,因而能做主语)的错误——“指称性”是世界多数语言中实词的共性,并非汉语特色。金立鑫(2022)也认为“名动包含”存在逻辑错误,具体体现为若认可该观点,动词作为名词的子集必然拥有动词的全部特征,而汉语中名词动词至少有13组完全对立的分布特征。吴义诚、戴颖(2022)认为“名动包含”的两个主要理据,即“动词形式可以做主宾语”和“名词成分可以做谓语”,是常见的跨语言现象,印欧语也有大量动名兼类词,并且汉语是孤立语,名词不会发生形态变化。因而指出“名动包含”在语言事实和概念使用上都存在问题。邵斌、杨静(2022)考察了英汉名动范畴边界渗透现象,结果表明,英语汉语的动名边界渗透都明显,英语汉语都属于柔性边界语言,不支持“名动包含”说。诸多学者(杨亦鸣等,2002;刘涛等,2008;夏全胜、彭刚、石锋,2014;Xia et al.,2016;吴铭,2021)的心理语言学实验也不支持“名动包含”。

从学者们正在持续中的论战来看,“名动包含”存在相当大的争议,汉语名词是否包含动词仍需要进一步的检验。

1.1.3 动名兼类的应用研究

动名兼类词不仅在语言学理论研究中存在诸多问题,这些问题也应用中有相

对表现。除了对动名兼类词和动名关系的纯粹理论性探讨，也有不少应用的动名兼类词研究，涉及的领域大致有语言教学、语言处理、词典编纂。

先回顾在语言教学，特别是汉语教学中的动名兼类词相关研究。李月华(1991)强调了兼类词与同音词、词类活用的区别，还指出了教材中兼类词的内容不够详细，需要教师在教学中结合维吾尔语补充词性，并且认为在低年级时不要过早过多引入兼类词的概念，当到了高年级，有了兼类词的概念后，才应该有意识地引导学生习得兼类词。付炜(1994)发现在维吾尔语母语的学生的汉语教学中，动名兼类词是个难点。学生们常会产出“老师做教育了我”之类的病句。付文认为这是母语的迁移所致，认为比较维汉句式，找到其中的不同点，才能让学生更好地习得动名兼类词。赵金铭等(2003)向《(汉语水平)词汇等级大纲》提出了意见，其中就强调了兼类词的词性排序问题。赵文认为兼类词的多个词性应当依照使用频次的高低进行排序，但这还需要依托于大规模标注语料库才能实现统计，进而排序。马宏基(2008)认为第二语言习得过程是个转换和注释的过程，需要学习者强制地把词的意义、使用存入大脑，因此教学中应该按照兼类词不同词性的使用频次排序，这样更有利于兼类词的习得。许红花(2009)讨论了《汉语水平词汇与汉字等级大纲》中对兼类词的分类标准和划类标准，认为不应当把实词和虚词、词和词缀划归为兼类，这不利于汉语学习者系统掌握汉语的词类体系和语法规范。朱志平(2013)认为把词汇量作为衡量(汉语作为)第二语言水平时应当充分考虑兼类问题，比如在统计“迷信”时，应当把动词“迷信”和名词“迷信”分开计算。赵雪莹(2013)认为汉语词汇具有兼容性，其中一点就反映在词的兼类上。一方面，兼类有利于语言习得，学生掌握一个词就能产出多个句子；另一方面，并不是所有词都是兼类词，兼类词的习得可能导致语内迁移，产出如“我分歧他的意见”之类的句子。

语言处理(即自然语言处理、计算语言学)领域中，动名兼类词也是一个不可回避、需要妥善考虑的问题。朱学峰等(1997)在研制“现代汉语语法词典”过程中，对五万余次进行了词类归类。对待兼类问题，采取了承认但限制规模的做法。温锁林、刘开瑛(1998)针对计算机是被自然语言的特点，提出了两种识别兼类词的方法。一个是“并列类推法”，即出现在并列结构中的两个词词性相同，比如“欢送会上热闹而感人的场面实在让人难忘”中的“热闹”和“感人”都是形容词；另一个是“同境类推法”，即连续出现的多个词，只要其句法环境和语法结构相同，其词性相同，比如“有人以为出大价钱就能搞到好材料、好设计、好装修”中“材料”、“设计”、“装修”都是名词。洪铭材等(2006)提出了一种基于 CRFs 模

型的中文词性标注方法。通过建立概率模型，该模型具有强大的推理能力，能够良好地降低兼类词对词性标注的影响。冯志伟(2001)指出在基于规则的方法中，最严重的问题就是兼类词，还统计了现代汉语中兼类词的比例，其中动名兼类词占比为 37.6%。詹卫东(2013)从自动句法分析角度审视了词类问题，并对宋柔(2003)针对“北大词类体系”中兼类词提出的该体系存在逻辑混乱和概念混乱做了回应。一方面肯定了其可能导致有的词无法进入词类体系的说法，另一方面又认为该问题在汉语的实际语境中问题不大。

在词典编纂中，动名兼类也是一个备受关注的问题。郭锐(1999)认为在语文词典的词性标注中，兼类词要全盘考虑，统一处理。在划类策略定下来后，就要坚持同一性原则，对词一视同仁。同时还要尽可能将兼类词的规模变小，规模大的要处理为“第三类词”，比如统计数量占比占 44%的“研究”类词应当被划为“名动词”，而不是兼类词。张晓颖(2005)肯定了《新世纪汉英大词典》中对于兼类词先设词类再分义项的做法。

王仁强(2010)考察了《现代汉语词典》(第 5 版)内的词类体系效度，认为词典在处理动名兼类词上可圈可点。词典把高频的动名兼类词处理为兼类词，把低频的动名兼类词处理为单词性词。这种做法得到了语法转喻理论(朱永生, 2003; 王仁强, 2009)和语言事实统计(安华林, 2005; 郭锐, 2002; 王仁强, 2006; 王仁强、周瑜, 2015)的支持。姜自霞等(2011)考察了《现代汉语词典》(第 5 版)的动名兼类词，认为词典有明确的兼类词意识，并且恪守了兼类词要少数的原则。李翔(2013)考察了《汉英词典》(第 3 版)中的兼类词，发现该词典的兼类词区分略宽，因而造成了兼类词相比其他词典更多，但李文也认为在划分标准为统一的情况下这种做法也是合理的，同时还有利于词类研究。方清明(2010)借助《现代汉语词典》(第 5 版)，统计并分析了动名兼类词的内部构成方式。尹洪波(2016)对比了《现代汉语词典》(第 5 版)和《现代汉语词典》(第 6 版)中的词类增删，并认为这能够支持“名动包含”假说。王仁强、黄昌宁(2017)认为坚持兼类要少的原则使得同类词条、相同词条在词典中处理不一致。

1.1.4 动名兼类研究评述

我们在 1.1.1 中简单介绍了词类划分的标准以及兼类词的界定原则，在 1.1.2 和 1.1.3 中分别介绍了理论和应用的动名兼类研究。尽管学者们确实取得了不少成果，但动名兼类研究中仍然有种种谜团正待解开。接下来，我们从资源和方法三个方面来评述动名兼类词研究中的发展趋势。

从研究资源来看,动名兼类研究中的数据源类型越来越丰富,数据来源可信度越来越高,但与其他研究领域相比仍存在更多发展空间。以往动名兼类词研究中的语料来源大致有以下几类:人工搜集(朱德熙、卢甲文、马真,1961)、词典和大纲(安华林,2005;尹洪波,2016)、语料库(周统权、杨静,2010)、树库(王璐,2022)。语言学研究早期计算机技术并不发达,或者说没有普及到语言学研究中,语言学研究的一大难点就是搜集语料,学者们通常通过摘录小说、报刊等的语料来支持研究,比较著名的就是卡片法(吕叔湘,1999)。学者们把收集的语料摘录在卡片上,并依照研究目的分好类别,以备使用。这种方法十分耗时耗力,是在技术不成熟时候的无奈之举。收集的难度也直接使得在这种方法下几定量研究的难度极大,所以我们很少能看到在这一时期内有什么太详备的动名兼类词计量研究。随着诸多词典和大纲的出版,如《现代汉语词典》、《(汉语水平)词汇等级大纲》,学者们又有了新的语料来源。这些资源是国家部门或者大学机构集合众力而成,具备一定的广度深度,可以作为研究的支撑。但短板也很明显,词典、大纲、教科书限于篇幅和使用目的,毕竟不可能全然地反映出语言事实,并且其指定原则也有诸多不同。在信息技术发展后,语料库成为动名兼类词的主要来源。语料库分为传统语料库和树库。前者是当前研究中的主流资源,通过设置检索条件即可检索出所需语料。后者是基于短语结构语法或者依存语法标注的语料库,其中存储着以树形结构表示的句子。树库是种基于真实语料的第一手语料资源,基于有丰富标注的树库才能更好地研究动名兼类词。当前树库已经较广泛地应用在计算语言学研究中,比如詹卫东(2012),但基于树库的汉语动名兼类词研究仍比较少见。树库是语言研究的重要知识源,在计算语言学对传统语言学研究的冲击下(刘海涛、林燕妮,2017;刘海涛、郑国锋,2021),树库的使用也必将越来越广泛。

从研究方法看,研究者们研究范式正逐渐从定性研究向定量研究转变。前文所述,在研究早期学者们由于现实条件的制约,大多都是以定性研究为主。随着信息技术的发展,动名兼类词的定量研究越来越多。定量研究的主要方法为实验和计量。心理语言学和神经语言学实验被广泛应用在动名兼类和动名关系的探讨中(杨亦鸣等,2002;刘涛等,2008;沈家焯、乐耀,2013;夏全胜、彭刚、石锋,2014;Xia et al., 2016;吴铭,2021;杨静,2022)。通过设计实验,研究者可以得到被试对于动名兼类的直观表现,进而帮助说明问题。计量方法也越来越多的应用在动名兼类词研究之中。从早期的简单统计兼类词的数量(陈爱文,1986;吴蔚天、罗建林,1994;胡明扬,1995a;程娟,2004;杨丽皎,2010;方

清明, 2010), 逐步发展到统计兼类词的概率分布(徐艳华、朱诗瑶, 2020; 王璐, 2022)。计量方法提供了一种研究动名兼类词的新视角。通过数学的计算, 词类间的关系, 特别是动词名词之间的关系与动名兼类词都能够实现以数学模型。通过对数学模型的分析探讨, 动名兼类研究更有可能走向深入。

以往研究的局限性比较明显。第一, 已有的本体研究大多是归纳法, 这种做法的问题在于得出的结论并不一定能真正反映实际。如归纳得出的“天鹅都是白色的”, 这一结论实际上是不准确的, 因为后来发现了黑天鹅。且因研究手段和资源所限, 学者们的语料也相对不够丰富, 因此会对结论产生影响。第二, 从实证出发的研究较少。实证的步骤应该是提出假设, 将假设数学化, 用统计学的方法检验假设, 得出结论, 经过检验之后的假设才可以被接受。尤其词类是词的集合, 能够提取和检验的数据有很多, 但很遗憾这方面的研究较为少见, 少数研究也是由心理语言学学者所做(夏全胜、彭刚、石锋, 2014; 吴铭, 2021)。第三, 计量手段不够丰富。有不少学者意识到计量在词类问题中的重要性, 但受技术手段所限, 往往只流于简单的统计数量, 而没有进一步的深入挖掘其他计量指标以及进行分析。

总的来说, 我们发现, 树库和定量研究在动名兼类问题中起到的作用越来越明显, 但如何将这二者结合并且运用在动名兼类研究中仍然是个研究者较少跨足的空间。

第二节 研究问题和研究意义

由于词类划分的标准不统一, 动名兼类词的归类划分问题仍未成定论。以往多是从理性主义出发, 从汉语语法本体研究来看待动名词类划分以及兼类问题, 但目前仍没有达成共识。既然这种视角难以取得统一, 我们不妨换个角度来看问题。从经验主义出发, 通过对大规模实际语料的统计, 用词在语料中实在的表现所得到的计量特征来看待它的词类归属, 可以对已有成果进行验证, 可以对动名兼类词研究提供新的认识。同时, 以往针对词类的计量研究多是停留于对语法单位频次的简单统计, 但近年来对词类计量的研究开始走向深入, 这说明越来越多学者注意计量方法在词类问题内的价值。在汉语词类研究使用自然语言处理的相关技术, 结合统计学方法, 对语言学问题进行合理的假设、论证、解释, 可以迎合数智时代的时代潮流, 促进汉语词类研究的进步。从具体内容来看, 我们将讨论以下几个问题:

(1) 如何借助概率配价模式理论, 计量动词、名词、动名兼类词的语法功能, 并将语法功能特征化?

(2) 如何计量动名关系和比较动名兼类词与动词、名词的相似度? 如何检验所使用的方法和指标?

(3) 所使用的方法和指标如何运用在应用(具体为语言习得研究)上?

以往的研究中, 在探讨动词、名词两词类关系时把动名兼类词纳入研究范围的并不多, 明确提出词类之间的关系就是连续统的研究也不多。在原型论者的论述中, 词类被视作一个个原型范畴, 每个词可以同时隶属于多个范畴, 并且有隶属度数值上的不同。截取动词和名词两个词类来看, 一个词对动词和名词词类的隶属值可能在 $[0, 1]$ 间波动。在对词类的计量特征进行提取时, 不加分辨地以所有名词为对象提取特征和以典型名词为对象提取特征有很大区别。前者反映地是一个模糊的类, 包含了核心成员和边缘成员, 特征代表的实际是所有成员共同的属性, 当然也包含了边缘成员身上不属于该范畴的属性。后者反映的是范畴内典型成员的特征, 是该范畴的典型特征。就理论上而言, 后者应当更具代表性。这点我们会在实际中检验。

本研究共计六章。第一章回顾了动名兼类词的相关研究历史并做评述, 阐明本研究的研究问题 and 研究意义。第二章介绍本研究使用的资源和方法。第三章依照概率配价模式计量了动词和名词的语法功能, 并区别了主要功能和次要功能。第四章依照概率配价模式计量了动名兼类词的语法功能, 并区别了主要功能和次要功能, 并且基于余弦相似度计算了动名兼类词与动词、名词的相似度, 做了讨论分析。第五章将基于概率配价模式的相似度公式应用在汉语学习者的动词、名词配价习得水平计量中。最后, 第六章总结了全文的工作, 并对当前工作的不足做了剖析, 对未来工作做了展望。

第二章 资源和方法

第一节 资源

句法计量研究的常见资源是依存树库，依存树库是按照依存语法理论（Mel’čuk, 1988）标注的语料库数据集。本研究采用的树库来自表层句法通用依存标注数据集（Surface Syntactic Universal Dependencies, SUD）。SUD 是基于句法原则标注的语料库资源（Gerdes et al., 2018; Osborne and Gerdes, 2019; Yan and Liu, 2019、2021），常被用做自然语言处理的知识源和语言学研究的数据来源。我们选择了 SUD 的现代汉语树库 GSDSimp 作为本文资源。GSDSimp 树库共计 4997 句，规模为 123291 例符（tokens）。其语料来源于维基，语体接近书面语。GSDSimp 树库的标注方案中包括 14 种词类和 42 种依存关系（不计入标点符号），原标签与汉语对照表可见于表 2.1。

表 2.1 词类与依存关系对照信息表¹

	原标签	汉语对照	原标签	汉语对照
词类	ADJ	形容词	NUM	数词
	ADP	介词	PART	小品词
	ADV	副词	PRON	代词
	AUX	助词	PROP	专有名词
	CCONJ	并列连词	SCONJ	从属连词
	DET	限定词	SYM	符号
	INTJ	叹词	VERB	动词
	NOUN	名词	X	其他
依存	appos	同位语修饰	discourse	话语标记词
关系	cc	并列词	discourse@sp	话语标记词@语气词
	clf	量词	dislocated	脱位词
	comp:aux	补足语: 助动词	flat@foreign	扁平固定搭配@外来语
	comp:sux@pass	补足语: 助动词@被动	flat@name	扁平固定搭配@名字
	comp:obj	补足语: 宾语	mod	修饰语
	comp:obj@adv	补足语: 宾语@副词	mod@relcl	修饰语@从句先行词
	comp:obj@rel	补足语: 宾语@先行词	mod@tmobj	修饰语@时间修饰语

¹ 具体内容可参见 SUD 官网 <https://surfacesyntacticud.github.io/guidelines/u/>。表中标签汉语对照参考借鉴了阎建玮（2022）的附录 M: UD 依存类型解释。

第二章 资源和方法

comp:obj@x	补足语: 宾语@不定式	orphan	孤词
comp:obl	补足语: 旁格	parataxis	并列式
comp:obl@x	补足语: 旁格@不定式	root	根节点
comp:pred	补足语: 系动词	subj	主语
compound	复合词	subj@pass	主语@被动
compound@ext	复合词@描述性补足语程度补足语	udep	属性或属格
conj	连词	udep@patient	属性或属格@受事
conj:cmb	连词: 嵌入式并列	unk	其他
det	限定词		

表 2.1 中的词类体系与语言学传统中的词类体系区别不大, 但依存关系代表的句法成分与语言学传统的主谓宾定状补体系区别较大: 把一些词类成分归入了句法成分, 如量词、连词; 宾语被归并为补足语的子类; 一些词成分被标注, 如复合词; 以“:”为子类标记, 以“@”为深层句法和语义信息标记。这样的安排一是为了让标注体系有普遍性, 能服务于所有语言, 二是更细粒度的划分有助于更详细地记录信息。依存语法是种功能语法, 处理句法层问题时也无法完全避开语义。

树库格式为 CoNLL-U, 如一个句子“随后爱斯基摩人和维京人相继定居于此。”以表 2.2 格式储存。部分内容空缺是因为汉语的这些特征不突出。“ID”为词在句子中的线性位置。“FORM”是词形, “LEMMA”是词元, 汉语无形态, 因此两列内容相同。“UPOS”是该词的词性。“HEAD”是该词的支配词的线性位置, 若此项为 0, 则该词是句子的根。“DEPREL”是依存关系²。

表 2.2 例句“随后爱斯基摩人和维京人相继定居于此。”

ID	FORM	LEMMA	UPOS	XPOS	HEAD	DEPREL
1	随后	随后	ADV	RB	8	mod
2	爱斯基摩	爱斯基摩	PROPN	NNP	3	compound
3	人	人	PART	SFN	8	subj
4	和	和	CCONJ	CC	6	cc
5	维京	维京	PROPN	NNP	6	compound
6	人	人	PART	SFN	3	conj
7	相继	相继	ADV	RB	8	mod
8	定居	定居	VERB	VV	0	root
9	于	于	ADP	IN	8	udep
10	此	此	PRON	PRD	9	comp:obj
11	。	。	PUNCT	.	8	punct

² 其他标签的详细内容可见于 Universal Dependencies 官网: <https://universaldependencies.org/format.html>.

依存语法把依存关系视作基本要素 (Tesnière, 1959; 刘海涛, 2009; Hudson, 2010)。依存关系是词与词之间有标记的不对称的二元关系, 处在依存关系两端的两个词中处于支配地位的叫支配词, 处于从属地位的从属词 (刘海涛, 2022)。因此句子可以视作由多对支配词 (governor) 和从属词 (dependent) 构成的图, 表 2.2 中例句可以被解析为图 2.1。

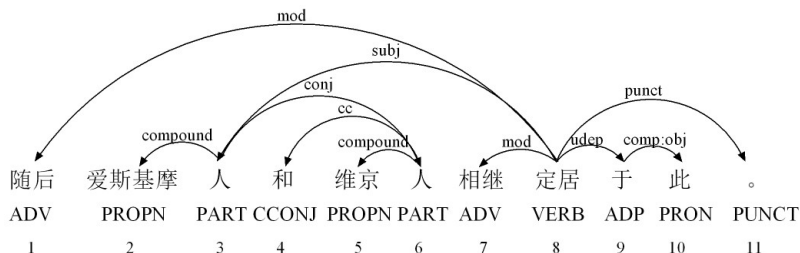


图 2.1 “随后爱斯基摩人和维京人相继定居于此。”的依存有向图

图 2.1 中, 例句“随后爱斯基摩人和维京人相继定居于此。”被划分为 10 个节点, 对应句中 10 个词, 词下方还标注了相应的词性标签。箭头从支配词指向从属词, 并以标签说明了两词之间依存关系的类型。

根据 Robinson (1970), 依存结构必须满足四个条件以保证其合理性, 接下来我们结合图 2.1 说明。第一, 一个句子只有一个根。充当根的词不从属于其他词, 并且能直接或间接支配其他词, 比如“定居”没有支配词, 但却可以直接或间接支配其他词, 因而“定居”是根。第二, 除根以外的所有词都必须从属于某词, 比如“随后”从属于“定居”, “相继”从属于“定居”。第三, 所有词最多只能有一个支配词, 即根无支配词, 其他词有一个支配词, 充当根的“定居”无支配词, 其他所有词都只有一个支配词。第四, 如果成分 A 直接从属于成分 B, 而成分 C 在句子中位于 A 和 B 之间, 那么, 成分 C 或者从属于 A, 或者从属于 B, 或者从属于 A 和 B 之间的某一成分, 比如处在“定居”和“此”之间的“于”从属于“定居”。根据这四条原则, 我们就能将一个句子很好地分析出来。

在依存语法视角下, 词的句法环境被分解为具有依存关系的词对, 描述词拥有的依存关系以及配价对象就可以得到其句法环境。比如例句中“定居”的句法环境可以表示为 mod、subj、udep 关系中的支配词, 还可以表示为 ADV、PART、ADP 的支配词。把词所有使用时的句法环境尽可能的汇总起来, 就能得到用具有依存关系的词对表示的词的语法功能。

总的来说, SUD 中的 GSDSimp 树库提供了规模较大、方便操作、准确度高

的语言研究资源，其更切合依存语法传统（Hudson, 2010; Mel'čuk, 1988）的标注方案也十分适用于句法指标的测量和计算。

第二节 研究方法

本节首先介绍本研究采用的分析工具，包括编程语言及其具体的相关库、包。然后重点说明本研究中使用的计量指标和计算方法。

2.2.1 分析工具

本研究中，我们采用 Python 编程语言，结合其第三方库挖掘和分析数据。Python 语言是荷兰人 Guido van Rossum 开发的一种解释型编程语言（van Rossum, 1995）。Python 支持多种编程范型，比如函数式、指令式等。Python 语法简洁易懂，并且拥有巨大的标准库和极为丰富的第三方库，这使得其功能极为强大、应用范围极为广阔。Python 是自然语言处理采用的主流编程语言之一，十分适合用于语言信息的处理。Python 基本可以适用于所有操作平台，适应性良好。

本研究中，Python 主要用于数据提取、数据分析和可视化，比如按照词性筛选统计对象以及其统计信息，将数据分析所得结果和模型可视化等。这些功能依托于具体的库实现，因此需要一一介绍。

Conllu 库是一个 CoNLL-U 解析器，能把 CoNLL-U 格式的字符串（string）转换为嵌套的字典（dictionary）格式。我们的树库资源是字符串格式，无法直接用于数据检索和统计。经 conllu 转换的字典能直接用来检索和统计，十分方便快捷。该库可以在 <https://github.com/EmilStenstrom/conllu/> 下载使用。

Scipy（Virtanen et al., 2020）主要用于数学、科学和工程计算，提供了大量数理统计相关功能，本研究借助其 stats 模块对数据进行统计检验。

Matplotlib（Hunter et al., 2007）是一个用于在 Python 中制作数组的 2D 图的库。Matplotlib 大量使用 Python 高效数据处理标准库 NumPy 和其他扩展代码，因此性能良好，是最著名的 Python 可视化库。本研究中数据分析结果的可视化都以 matplotlib 和基于它的相关库 seaborn（Bisong and Ekaba, 2019）实现。

2.2.2 计量指标与计算方法

词类间并不是泾渭分明的，在无形态的汉语中这种现象表现得尤为突出。连

续性观点下的词类研究认为词类是一个个有重叠的范畴,且在范畴内部存在典型成员和边缘成员之分。动名兼类词的存在证明了动词范畴和名词范畴存在一定程度的交叠,这点也被诸多研究证实。词类划分最可靠的标准是词的语法功能,可以用词的语法功能来判断它在某个词类内的典型性程度。那么如何将词的语法功能数值化并用所得数据判断其典型性程度呢?

正如 2.1 所说,词的句法环境可以用与其相关的具有依存关系的词对表达。我们提取某词的所有词对,计算处在词对中的词的信息,就可以得到该词各项语法功能的概率。我们将这一组有标签的概率视作词这个对象的特征,把这一组概率视作该词的特征向量。以此为基础,我们能够计算每个动名兼类词与动词、名词词类的相似度。概率性是语言的本质属性 (Liu, 2011),通过基于概率的特征我们就可以计算词在不同词类范畴内的位置。

2.2.2.1 基于概率配价模式的特征向量计算

根据概率配价模式理论 (刘海涛、冯志伟, 2007),词或词类的配价能力可以用其概率配价模式表示。词或词类的概率配价模式如图 2.2 所示。

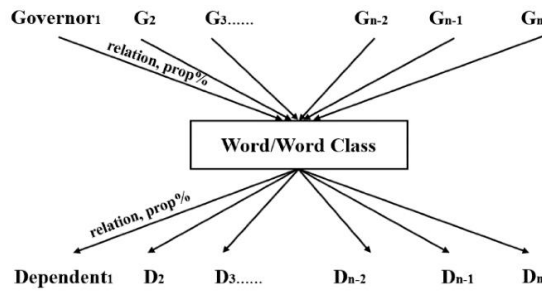


图 2.2 词或词类的概率配价模式

矩形为词或词类,上方的 G_1 、 G_2 、 G_3 G_{n-2} 、 G_{n-1} 、 G_n 是支配该词或词类的词或词类;下方的 D_1 、 D_2 、 D_3 D_{n-2} 、 D_{n-1} 、 D_n 是从属于该词或词类的词或词类。箭头代表依存关系,上面的标签代表依存关系的概率。

依存语法框架下,词的配价能力反映的就是词的语法功能 (Yan and Liu, 2022; Yang and Liu, 2022)。概率配价模式实际上提供了两组能够反映词的语法功能的特征集合:配价对象词性的特征集和依存关系的特征集。传统视角下,词(词类)能否与某词(词类)结合是衡量词的句法能力的重要标准,“能否结合”实际上就是两词(词类)能否建立依存关系。相比于离散视角下的“能不能”,连续视角下的“多么能”更深入一步。因而从这里我们可以推出两种特征获取方法,第一种关注与目标词(词类)的词(词类),第二种关注目标词(词类)所处的依存关系。

从词的视角出发时，选择词类（词）的配价对象的词类标签作为特征，出现该词类的概率为特征值。以上内容可以表示为：对于 n 个词类标签，词类标签的集合 $C = \{C_1, C_2, C_3, \dots, C_{n-2}, C_{n-1}, C_n\}$ 。对于集合 C 内的元素，其对应概率集合 $p(C) = \{p(C_1), p(C_2), p(C_3), \dots, p(C_{n-2}), p(C_{n-1}), p(C_n)\}$ ，集合 $p(C)$ 内的元素之和为 1。以此方法可以得到目标词（词类）做支配词时和做从属词时的两个集合，将两个集合内的元素以固定顺序排列，即可得到一组基于词的特征向量。对于例符数为 i 的数据集，其特征结构的形状为 $i \times 2n$ 。

仍以表 2.2 中例句为例。“定居”不做从属词，只做支配词。做支配词时，支配 2 个 ADV，1 个 PART，1 个 ADP，不支配其他词类。可得“定居”的特征向量为 $[0.5 (\text{gov_ADV}), 0.25 (\text{gov_PART}), 0.25 (\text{gov_ADP}), 0, \dots, 0]$ 。将文本扩大后，向量里的 0 将会减少，因而建立在一定数据量上的特征向量才具有意义。

从依存关系的视角出发时，选择依存关系作为特征，出现该依存关系的概率为特征值。以上内容可以表示为：对于 n 个依存关系，依存关系的集合 $R = \{R_1, R_2, R_3, \dots, R_{n-2}, R_{n-1}, R_n\}$ 。对于集合 R 内的元素，其对应概率集合 $p(R) = \{p(R_1), p(R_2), p(R_3), \dots, p(R_{n-2}), p(R_{n-1}), p(R_n)\}$ ，集合 $p(R)$ 内的元素之和为 1。以此方法也可以得到目标词（词类）做支配词时和做从属词时的两个集合，将两个集合内的元素以固定顺序排列，即可得到一组基于词的特征向量。对于例符数为 i 的数据集，其特征结构的形状为 $i \times 2n$ 。

同样以表 2.2 例句为例。“定居”不做从属词，只做支配词。做支配词时，与其他词建立了 2 个 mod 关系，1 个 subj 关系，1 个 udep 关系，因而其特征向量为 $[0.5 (\text{gov_mod}), 0.25 (\text{gov_subj}), 0.25 (\text{gov_udep}), 0, \dots, 0]$ 。

借助概率配价模式理论，我们可以将词和词类向量化，只有经过向量化的词和词类，才便于计算相似性。

2.2.2.2 基于特征向量的相似度计算

常见的监督学习和无监督学习都需要借助特征向量。机器学习的两大任务就是分类和聚类，前者是给定类别数量，把一组样本分入这几类；后者是不给定类别数量，把相似样本聚集成类，如语言分类（Liu and Li, 2010; Liu and Cong, 2013）、语体分类（陈芯莹、刘海涛，2012）。我们的研究目的在于讨论动名兼类词在动词和名词两个词类范畴内的地位，并不是给动名兼类词分类或聚类，但这不代表与分类聚类任务就毫无关系。

分类算法和聚类算法的实现通常要借助于计算样本间的距离或相似性。我们可以通过计算动名兼类词与名词或名词间的相似性来判断其在两词类上的位置。

余弦相似度 (Cosine Similarity) 是常见的相似度计算指标, 余弦相似度是一个向量空间中两个向量夹角间的余弦值, 能够用来衡量两个个体之间差异的大小, 余弦值接近 1, 夹角趋于 0 度, 表明两个向量越相似, 余弦值接近于 0, 夹角趋于 90 度, 表明两个向量越不相似。两个向量 A 和 B 间的计算公式为:

$$Similarity(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^n (A_i \times B_i)}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}} \dots\dots\dots (2.1)$$

如点 $A(1, 2)$ 和点 $B(1, 4)$ 之间的相似度为 $\frac{1 \times 2 + 1 \times 4}{\sqrt{1^2 + 2^2} \times \sqrt{1^2 + 4^2}} = 0.97$ 。若动名词类 V 的特征向量为 $[C_{V1}, C_{V2}, C_{V3}, \dots, C_{Vn-2}, C_{Vn-1}, C_{Vn}]$, 动名兼类词 H 的特征向量为 $[C_{H1}, C_{H2}, C_{H3}, \dots, C_{Hn-2}, C_{Hn-1}, C_{Hn}]$, 则动名兼类词 H 的特征与动词的相似度为 $\frac{\sum_{i=1}^n (C_{V_i} \times C_{H_i})}{\sqrt{\sum_{i=1}^n C_{V_i}^2} \times \sqrt{\sum_{i=1}^n C_{H_i}^2}}$ 。同理, 也可以计算出该词与名词的相似度。此时, 我们就

得到了一个相似性指标 $Similarity (Similarity \in [-1, 1])$ 。就我们的问题而言, 概率不存在负值, 因此相似度也不会存在负值, 此时的相似度为 $Similarity (Similarity \in [0, 1])$ 。 $Similarity$ 越接近 1, 两词 (词类) 的语法特征也就接近; $Similarity$ 越接近 0, 两词 (词类) 的语法特征就越有差异。

借助余弦相似度, 我们就可以从语法功能角度比较词 (词类) 之间的相似性了, 进而可以观察动名兼类词与动词、名词的关系。

第三章 动名兼类词的语法功能计量

不同语法体系下,词的语法功能可以用不同方式描述,比如结构主义框架下,可以用层次分析和分布分析描述;生成语法框架下,可以用短语结构树、生成词库描述。尽管流派不同,但词(词类)的语法功能不外乎在两个方面:一是词(词类)能与哪些词(词类)结合;二是词(词类)能做哪些句法成分。动名兼类词的语法功能计量也应该从这两个方面着手。

本研究采纳依存语法体系,词的语法功能大致可以等同于其配价能力,特别是词在句法层的配价能力(Yan and Liu, 2022; Yang and Liu, 2022)。依存语法把语法关系视为不对称、有标签的二元关系,即依存关系。词(词类)的语法功能即它能与哪些词(词类)建立哪些依存关系,获取这些词(词类)或关系就能得到一种描述语法功能的方法。概率配价模式理论在此基础上给词(词类)的语法功能加上了概率性特征,即词(词类)有多大可能与某词(词类)建立某依存关系,这提供了一种计量语法功能的可行手段。

第一节 动名兼类词的选择

作为研究对象,动名兼类词成分十分复杂,其内容并不都是体现其动词性功能和名词性功能的。此外,语料中动名兼类词数量是有限的,其出现频次也是不同的。频次太低,其统计意义也就值得商榷。因此,我们首先需要给研究对象定一个更明确的范围。

词的语法功能是连续的,因此有典型和非典型之分。动名兼类词内部同样存在典型动名兼类词和非典型动名兼类词。典型动名兼类词只具有名词性功能和动词性功能,而非典型动名兼类词还兼有其他词性的功能。

我们统计了每个词元拥有的所有词类标签,将其中只有名词和动词的记为典型动名兼类词,还有其他词类标签的记为非典型动名兼类词。然后经人工校验,剔除因字形相同而导致的误差。经统计,GSPSimp 树库中,典型动名兼类词 763 个。最后,以频次为筛选条件,剔除频次小于等于 5 的对象,共得典型动名兼类词 286 个。词表可见于附录 B。

本章中,我们把典型动名兼类词作为计量对象,从词和依存关系两个角度计

量其语法功能。本章中，如未特别说明，“动名兼类词”指典型动名兼类词。

第二节 基于词的语法功能计量

就 2.2.2.1 所述，我们可以通过计算动名兼类词的支配词、从属词情况，得到动名兼类词与词类的结合情况，并记录为概率化的特征向量。

首先，统计动名兼类词的所有支配词和从属词，记录其词性。需要注意的是，支配词和从属词当然也可能存在兼类，但在具体的句子中能相对明确地表现出某种词类的功能，因此被标记为该词类。然后，统计支配词和从属词的词性情况并归一化，得到各个词性及其概率。为了将成组概率转化为特征向量，按照固定的词类标签顺序排列作为特征值的概率。在基于词的语法功能计量中，我们默认词类标签的顺序如下（附对应例句）：

AUX（助动词）：部分“光棍”会选择各式各样的庆祝方式

ADJ（形容词）：可以根据剧情需要作不同选择

ADV（副词）：最后他选择从事基督教音乐事业

NUM（数词）：火提供了天然光源外的另二选择

VERB（动词）：他们的关系在德瑞克选择重回妻子身边后画下句点

NOUN（名词）：选择婴儿性别在印度属于违法

PRON（代词）：他选择前往欧洲球坛觅食

PART（小品词）：选择新皇后的时候根本没有把她考虑在内

ADP（介词）：为了保住她的生命，他们不得不选择流产

DET（限定词）：火提供了天然光源外的另一选择

CCONJ（并列连词）：可自创主角或者选择游戏提供的两个虚拟的主角

PROP（专有名词）：他们的关系在德瑞克选择重回妻子身边后画下句点

按以上方法，可以得到所有动名兼类词的语法特征向量。比如“选择”的特征向量为：

[0.03125000, 0.03125000, 0.00000000, 0.03125000, 0.21875000, 0.28125000, 0.0625000, 0.15625000, 0.09375000, 0.03125000, 0.03125000, 0.03125000]

[0.28571429, 0.00000000, 0.09523810, 0.00000000, 0.38095238, 0.00000000, 0.00000000, 0.19047619, 0.04761905, 0.00000000, 0.00000000, 0.00000000]

以上两组向量分别是“选择”做支配词和从属词时的特征向量。向量表示：“选择”有 3.1%的概率支配助动词，3.1%的概率支配形容词，0%的概率支配副词，3.1%的概率支配数词，21.9%的概率支配动词，28.1%的概率支配名词，6.2%的概率支配代词，15.6%的概率支配小品词，9.3%的概率支配介词，3.1%的概率支配限定词，3.1%的概率支配并列连词，3.1%的概率支配专有名词；有 28.6%的概率从属助动词，0%的概率从属形容词，9.5%的概率从属副词，0%的概率从属数词，38.1%的概率从属动词，0%的概率从属名词，0%的概率从属代词，19.0%的概率从属小品词，4.8%的概率从属介词，0%的概率从属限定词，0%的概率从属并列连词，0%的概率从属专有名词。因此，我们根据所结合的词类得到了两组具有语言学意义的向量。

2.2.2.1 介绍的方法对词和词类都适用。根据将统计对象视为动名兼类词这个类还是一个一个的动名兼类词，可以得到两种结果。我们以这两种结果相互佐证，从而区分出动名兼类词的主要语法功能和次要语法功能。

3.2.1 以类为单位的语法功能统计

将动名兼类词视作一个整体，统计其语法功能，此时的统计流程实际上等同统计词类的概率配价模式。首先遍历所有句子，记录每个词有的所有词类标签。然后找出同时具有且只具有动词和名词标签的词，并记录它们支配词和从属词的词性。随后，不区分词元，统计动名兼类词的支配词和从属词中每种词类的数量，转化为概率，按照顺序排列即可获得动名兼类词的特征向量。

经计算，动名兼类词的两种特征向量分别为：

[0.01868171, 0.02097286, 0.06415227, 0.00704970, 0.12354600, 0.39302080, 0.01850546, 0.16320056, 0.11579133, 0.00810716, 0.01727177, 0.04970039]

[0.11838219, 0.00568053, 0.07770961, 0.00022722, 0.36332652, 0.14519427, 0.00022722, 0.21063395, 0.07748239, 0.00000000, 0.00000000, 0.00113611]

为方便观察，我们将以上结果绘制于图 3.1。

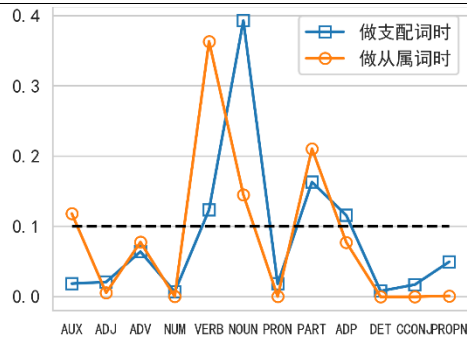


图 3.1 动名兼类词与词类结合的情况

我们将概率大于 0.1 的语法功能视做主要功能，其他视为次要功能。根据概率大小，动名兼类词的主要语法功能为：支配名词、小品词、动词、介词；从属动词、小品词、名词、助动词。

接下来，再以词为单位，做更为细致的统计。

3.2.2 以词为单位的语法功能统计

将动名兼类词视作一个集合，统计集合内成员的语法功能，此时的统计实际上等同统计词的概率配价模式。首先遍历所有句子，记录每个词有的所有词类标签。然后找出同时具有且只具有动词和名词标签的词，并记录它们支配词和从属词的词性。随后，区别词元，统计其支配词和从属词每种词类的数量，转化为概率，按照顺序排列即可获得每个动名兼类词的特征向量矩阵。我们将所有动名兼类词的语法功能以图 3.2 表示出来。

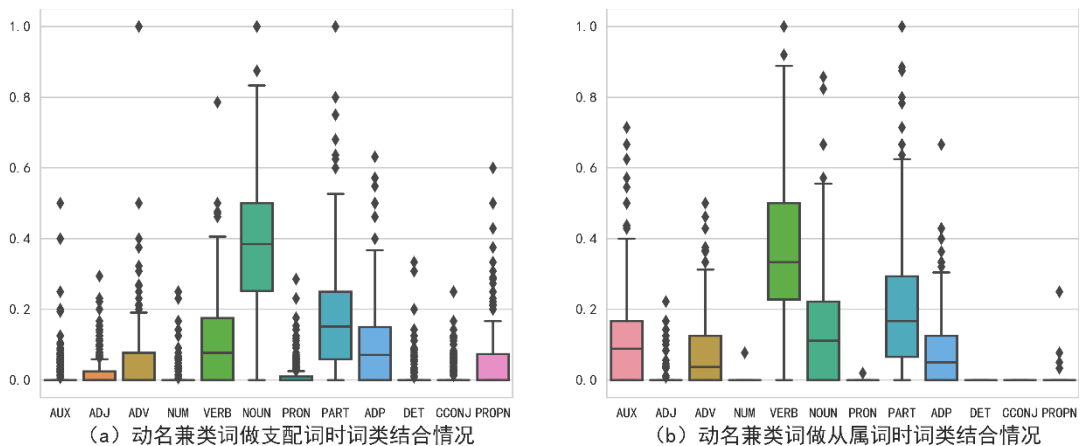


图 3.2 动名兼类词的词类结合情况

从动名兼类词的整体来看，各项语法功能的数值不相等，存在主次差异。我们可以将动名兼类词能结合的几种词类区分出主次。首先检验各项语法功能内部的分布是否符合正态。经 Kolmogorov-Smirnov 检验，全部数据都不服从正态分布 ($p < 0.05$)。随后两两比较所有数据的大小。由于数据全部为非正态分布，选用非参数检验的 Wilcoxon 检验进行两两比较， p 值可见于表 3.2³。

表中可见，绝大多数两两对比的 p 值都小于 0.05，表明对比的两项间存在显著差异。这足以让我们给动名兼类词的语法功能做出主要功能区分。

结合 3.2.1 中以类为对象的统计结果，我们可以得出：从结合的词类看，动名兼类词主要功能是支配名词、小品词、动词 ($\text{mean} > 0.1$)，从属动词、小品词、名词、助动词 ($\text{mean} > 0.1$)。

3.2.3 本节小结

结合以类为对象和以词为对象两种方法得到的结果，我们可以认为，动名兼类词的主要功能就是支配名词、小品词、动词、介词和从属动词、小品词、名词、助动词。我们得到了以下结论：

首先，除了不能做代词、限定词、连词的从属词，动名兼类词能与其他所有词类结合。这说明动名兼类词的可配价成分类型十分丰富，这种丰富性来自于动名兼类词兼有动词和名词两种词性的功能。在“他研究语言学”中，“研究”表现出动词词性，支配代词“他”、名词“语言学”。而在“他做语言学研究”中又表现出名词词性，从属于动词“做”，支配名词“语言学”。动名兼类词可以灵活地出现在动词或名词可出现的位置，因而能够与大多数词类结合。

其次，从属和支配的主要结合词类大致相同，只存在介词和助动词两个词类的区别。动词、名词、小品词的高频与他们在语言本身中的占比有关系。动词和名词的占比分别为 21.75% 和 39.29% (刘海涛, 2009: 251)，而小品词则是使用最为频繁的词类 (Liu, 2008)，这就使得他们更可能参与词类结合。那么支配和从属时的主要区别显然在于介词和助动词。介词和助动词都是封闭类，内部成员的数量是有限的，且明显少于作为开放类的动词和名词。当动名兼类词支配介词时，介词结构常充当补语 (张云徽, 2005) 和状语 (岳中奇, 2007)，比如“我走向北方”中介词“向”充当动词“走”的补语，“关于他，我不再提了”中介词“关于”充当动词“提”的状语；做从属词时，常充当助动词的宾语，如“我不想吃饭”中助动

³ 本节的表占篇幅较大，考虑到论文的可读性，表 3.1-3.4 见于附录 A，不在正文中登出。

词“想”支配动词“吃饭”。这两点上动名兼类词表现出明显的动词性功能偏好（高松、颜伟、刘海涛，2010），更能体现出其动词性功能。区别还体现在具体的概率上，动名兼类词最主要支配名词，最主要从属动词。支配名词是动词的最主要功能，而从属于动词是名词的最主要功能（详见第四章）。这也反映出了动名兼类词的兼类特性。

除了考虑可结合的词类，我们还可以把动名兼类词可进入的依存关系纳入考虑范围。

第三节 基于依存关系的语法功能计量

按照 2.2.2.2 所述方法，我们能计算出动名兼类词的依存关系情况，得到第二种概率化的语法特征向量。

首先，检索所有以动名兼类词为支配词或从属词的依存关系类型及数量。随后计算各项依存关系概率。为了将成组概率转化为特征向量，按照固定的依存关系标签顺序排列作为特征值的概率。在本研究基于依存关系的语法功能计量中，为了便于比较和追求准确，选取的依存关系标签是典型动(名)词(详见第四章)、动名兼类词可能存在的依存关系的并集。我们默认依存关系的顺序如下（附对应例句）：

Subj（主语）：选择 婴儿性别在印度 属于 违法

Subj@pass（主语@被动）：就算 数论的 应用 被找到了，也不会有人会因此罢黜这一数学的皇后

Root（根）：他以德裔身份取得了德国护照，选择 前往欧洲球坛觅食

Comp:obj（补足语：宾语）：这次游行最大的特色，在 于越来越多年轻人上街 游行，而且当中不乏行动激烈的躁少年

Comp:obj@adv（补足语：宾语@副词）：所以，有关成年人及老年人心理发展的研究开始如雨后春笋 的 发表

Comp:obj@rel（补足语：宾语@先行词）：学校的社会经济地位得分得自于从澳大利亚统计署人口普查中一个人口普查区中 选择 的 样本

Comp:obj@x（补足语：宾语@不定式）：1986 年开始，三星生命保险 开始 拓展 海外市场，并在日本东京和美国纽约开设办事处

Comp:obl（补足语：旁格）：红曲米，是一种 经过 红曲菌 发酵 加工的米，呈红色，用于食物上色，是一种天然的食品添加剂

Comp:obl@x（补足语：旁格@不定式）：弓弦乐器至最早在唐朝出现，可能 是 从弹拨乐器箏发展 出 以竹片拉奏的轧箏

Comp:aux (补足语: 助动词): 新东方网吧采用提问积分制, 同样 可以 选择 付出一定积分 (0-30 分) 作为回答者的奖励

Comp:aux@pass (补足语: 助动词@被动): 2001 年时, 旧议会大楼的遗迹在一次挖掘中 被 发现

Comp:pred (补足语: 系动词): 霸占该区域 是 最有利的 选择

Conj (连词): 可自创主角 或者 选择 游戏提供的两个虚拟的主角

Conj:emb (连词: 嵌入式并列): 在此之后, 地形图集也成为了现代化国家在规划基础设施和 矿产 探测 中的基础国家资源

Mod (修饰语): 诸葛亮 发明 八阵图 以 利立营练兵

Mod@relcl (修饰语@从句先行词): 《中药大辞典》是中华人民共和国 成 立 后出版的 第一部 大型中药专业工具书

Mod@tmod (修饰语@时间修饰语): 莫里逊组 过去 是 个多样化的动物群

Parataxis (并列式): 众多民工 选择 前往 内地的这些特大城市务工

Dislocated (脱位词): 不过, 由于 他在同年 9 月膝头 重伤, 使得他未能更进一步

Udep (属性或属格): 截至 2009 年, 千禧公园是全芝加哥人气第二高的 旅游 景点, 仅次于海军码头

Udep@tmod (属性或属格@时间修饰语): 小 时候, 因为擅长长笛的母亲影响, 使 她时常接触音乐,

Udep@patient (属性或属格@受事): 2006 年, 武昌区政府投入 800 万资金, 将原本三米宽的小街扩展到八米, 空中 网线 全部 入地, 并针对油烟污染等大力改善

Appos (同位语修饰): 全国青联的机关刊物是《中华儿女》, 另办有内部工作刊物《全国青联 通讯》

Compound (复合词): 英国航空宣布订购 15 架波音 777, 另加 15 架 选择 权, 采用奇异的发动机

Compound@ext (复合词@描述性补足语程度补足语): 他兢兢业业, 把文化馆 管理 得 有条不紊

Clf (量词): 最常见的情况是在该处安装 两 挺 步枪口径的机枪

Det (限定词): 该 专辑 并在 2007 年 5 月发行

Cc (并列词): 该路临近静安寺 和 曹家 渡 两段店铺较多

Discourse (话语标记词): 而且中国共产党 又 是中国人民 的 大救星

Orphan (孤词): 北京外城共有七门, 南面 三 门, 东西各一门, 此外还有两座便门

按前文所述方法, 可获得所有动名兼类词的语法特征向量。仍以动名兼类词“选择”为例, 其做支配词和从属词时的特征向量分别为:

[0.18750000,0.00000000,0.00000000,0.28125000,0.00000000,0.00000000,
0.25000000,0.00000000,0.00000000,0.00000000,0.00000000,0.00000000,
0.00000000,0.00000000,0.09375000,0.00000000,0.03125000,0.00000000,
0.00000000,0.09375000,0.00000000,0.00000000,0.00000000,0.00000000,
0.00000000,0.00000000,0.03125000,0.03125000,0.00000000,0.00000000]

[0.04545455,0.00000000,0.04545455,0.36363636,0.00000000,0.09090909,
0.00000000,0.00000000,0.00000000,0.22727273,0.00000000,0.04545455,
0.04545455,0.00000000,0.00000000,0.00000000,0.00000000,0.04545455,
0.00000000,0.00000000,0.00000000,0.00000000,0.00000000,0.09090909,
0.00000000,0.00000000,0.00000000,0.00000000,0.00000000,0.00000000]

以上这组特征向量表示：“选择”有 18.8% 的概率做 subj 关系中的支配词，0% 的概率做 subj@pass 关系中的支配词，0% 的概率做 root 关系中的支配词，28.1% 的概率做 comp:obj 关系中的支配词，0% 的概率做 comp:obj@adv 关系中的支配词，0% 的概率做 comp:obj@rel 关系中的支配词，25.0% 的概率做 comp:obj@x 关系中的支配词，0% 的概率做 comp:obl 关系中的支配词，0% 的概率做 comp:obl@x 关系中的支配词，0% 的概率做 comp:aux 关系中的支配词，0% 的概率做 comp:aux@pass 关系中的支配词，0% 的概率做 comp:pred 关系中的支配词，0% 的概率做 conj 关系中的支配词，0% 的概率做 conj:emb 关系中的支配词，0.9% 的概率做 mod 关系中的支配词，0% 的概率做 mod@relcl 关系中的支配词，0.3% 的概率做 mod@tmod 关系中的支配词，0% 的概率做 parataxis 关系中的支配词，0% 的概率做 dislocated 关系中的支配词，0.9% 的概率做 udep 关系中的支配词，0% 的概率做 udep@tmod 关系中的支配词，0% 的概率做 udep@patient 关系中的支配词，0% 的概率做 appos 关系中的支配词，0% 的概率做 compound 关系中的支配词，0% 的概率做 compound@ext 关系中的支配词，0% 的概率做 clf 关系中的支配词，0.3% 的概率做 det 关系中的支配词，0.3% 的概率做 cc 关系中的支配词，0% 的概率做 discourse 关系中的支配词，0% 的概率做 orphan 关系中的支配词。

“选择”有 0.5% 的概率做 subj 关系中的从属词，0% 的概率做 subj@pass 关系中的从属词，0.5% 的概率做 root 关系中的从属词，3.6% 的概率做 comp:obj 关系中的从属词，0% 的概率做 comp:obj@adv 关系中的从属词，0.9% 的概率做 comp:obj@rel 关系中的从属词，0% 的概率做 comp:obj@x 关系中的从属词，0%

的概率做 comp:obl 关系中的从属词, 0% 的概率做 comp:obl@x 关系中的从属词, 22.7% 的概率做 comp:aux 关系中的从属词, 0% 的概率做 comp:aux@pass 关系中的从属词, 0.5% 的概率做 comp:pred 关系中的从属词, 0.5% 的概率做 conj 关系中的从属词, 0% 的概率做 conj:emb 关系中的从属词, 0% 的概率做 mod 关系中的从属词, 0% 的概率做 mod@relcl 关系中的从属词, 0% 的概率做 mod@tmod 关系中的从属词, 0.5% 的概率做 parataxis 关系中的从属词, 0% 的概率做 dislocated 关系中的从属词, 0% 的概率做 udep 关系中的从属词, 0% 的概率做 udep@tmod 关系中的从属词, 0% 的概率做 udep@patient 关系中的从属词, 0% 的概率做 appos 关系中的从属词, 0.9% 的概率做 compound 关系中的从属词, 0% 的概率做 compound@ext 关系中的从属词, 0% 的概率做 clf 关系中的从属词, 0% 的概率做 det 关系中的从属词, 0% 的概率做 cc 关系中的从属词, 0% 的概率做 discourse 关系中的从属词, 0% 的概率做 orphan 关系中的从属词。

以上结果中依存关系颗粒度非常细。为了更贴近语言学传统中的句法成分, 我们可以合并一些具有相同标记的依存关系, 从而获得一组粗颗粒的依存关系。粗颗粒的操作更有利于我们观察和从语言学角度解释动名兼类词的特征。

首先, 标签中“@”后的成分是深层句法或语义信息, 在我们聚焦于表层句法信息时可以被忽略。其次, SUD 标注方案中用“:”标记的细化标签可以合并。按以上思想, 我们把 30 种依存关系标签合并为 15 种:

(1) root、parataxis、dislocated、appos、clf、det、cc、discourse、orphan 保持不变;

(2) subj@pass、subj 合并为 subj; comp:obj、comp:obj@rel、comp:obj@x、comp:obj@adv、comp:obl@x、comp:obl、comp:aux、comp:aux@passive 合并为 comp; mod@relcl、mod@tmod、mod 合并为 mod; conj:emb、conj 合并为 conj; udep@mod、udep@patient、udep 合并为 udep; compound、compound@ext 合并为 compound。

调整后的依存关系标签及其顺序: subj, root, comp, conj, mod, parataxis, dislocated, udep, appos, compound, clf, det, cc, discourse, orphan。

3.3.1 以类为单位的语法功能统计

基于依存关系计算语法功能也能以类为单位。首先遍历所有句子, 记录每个词有的所有词类标签。然后找出同时具有且只具有动词和名词标签的词, 并记录它们支配词和从属词的依存关系。随后, 不区分词元, 统计动名兼类词的支配词

和从属词中每种依存关系的数量，转化为概率，按照顺序排列即可获得动名兼类词的特征向量。

经计算，动名兼类词的特征向量为：

[0.15926758, 0.00000000, 0.25911211, 0.02573847, 0.32302643, 0.03765763, 0.00051822, 0.13231992, 0.00621869, 0.00069097, 0.01710140, 0.01951978, 0.01848333, 0.00034548, 0.00000000]

[0.07090221, 0.06652061, 0.48655646, 0.03206533, 0.21569408, 0.01892053, 0.00019916, 0.00059749, 0.00338578, 0.10515833, 0.00000000, 0.00000000, 0.00000000, 0.00000000, 0.00000000]

将向量所表示的概率绘制于图 3.3。

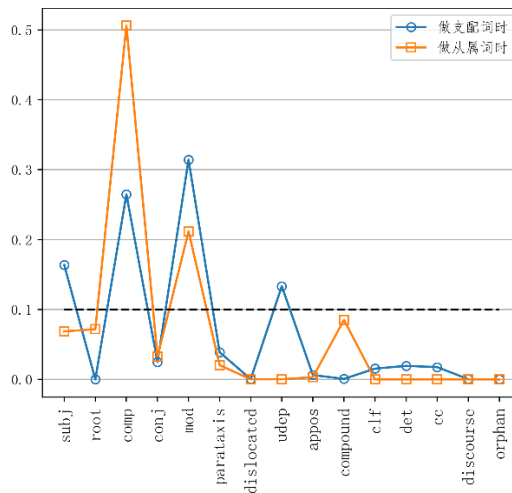


图 3.3 动名兼类词支配和充当各句法成分的情况

同样以概率 0.1 为分界，我们可以得出动名兼类词的主要功能是做 mod、comp、subj、udep 关系中的支配词，以及做 comp、mod 中的从属词。

接下来，以词为单位统计语法功能，与以上结果相互佐证。

3.3.2 以词为单位的语法功能统计

区分词元，统计动名兼类词的支配词和从属词中每种依存关系的数量，转化为概率，按照顺序排列即可获得动名兼类词的特征向量矩阵，将矩阵内容绘制于图 3.4。

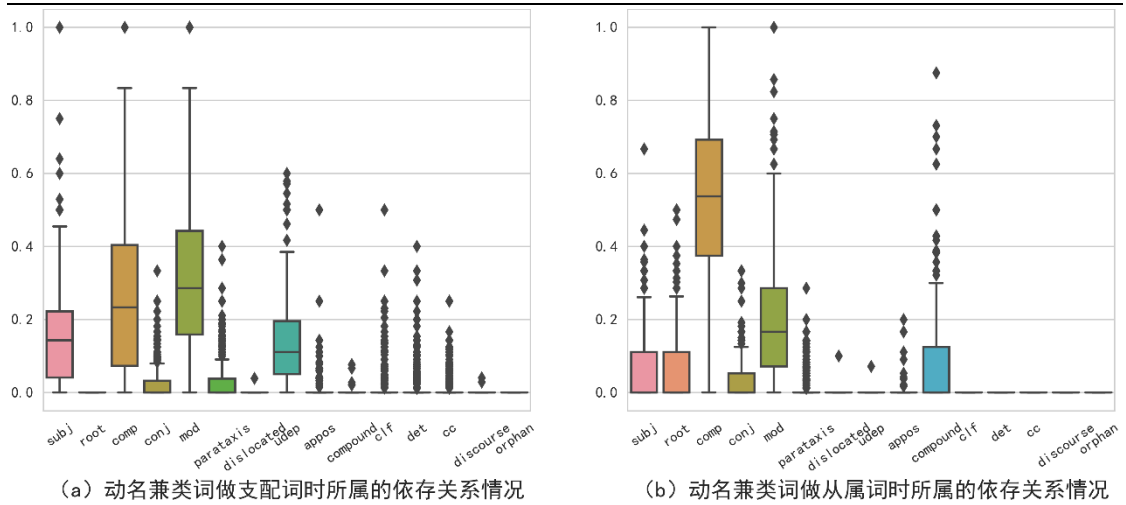


图 3.4 动名兼类词所属的依存关系情况

随之使用 Wilcoxon 检验进行两两比较，结果见于表 3.4。

经过比较，结合 3.2.1 中统计结果，我们可以认为：着眼于依存关系时，动名兼类词的主要功能是做 mod、comp、subj、udep 中的支配词和做 comp、mod 中的从属词。

3.3.3 本节小结

本节使用了以类为对象和以词为对象两种方法统计了动名兼类词可进入的依存关系，结果表明，动名兼类词的主要功能是做 mod、comp、subj、udep 中的支配词和做 comp、mod 中的从属词。这说明，动名兼类词常支配修饰语、宾语、主语、补语，常做宾语、修饰语。修饰语包括了传统意义上的定语和状语，SUD 的标注体系并未对此做出区分。就实际而言，动名兼类词常做的是定语，比如“研究内容”中的“研究”可以被分析为“内容”的定语；常支配的是定语和状语，比如“一篇出色的研究”中“出色”做“研究”的定语，“他认真地研究鸟”中“认真”做“研究”的状语。

结合莫彭龄、单青（1985），徐艳华（2006），高松、颜伟、刘海涛（2010）对动词语法功能的统计来看，动词可做的主要句法成分大致是谓语、宾语、定语。显然对应于谓语的 root 并没有出现在动名兼类的主要功能中。我们认为这是标注体系的影响。传统意义上的谓语并不只限于 root，多出的部分被纳入了其他依存关系，比如“我写了好看的小说，但他不喜欢”中充当 root 的是“写”，处在子句中的“喜欢”充当的其实也是谓语，但由于语法体系的

不同，“喜欢”充当了“但”的 comp: obj（即宾语）。因此如果把谓语也视作动名兼类词的主要功能，那么我们的结果与前人对动词语法功能的研究结果相符，这说明动名兼类词与动词在可充当的句法成分上有高度的相似性。

莫彭龄、单青（1985）和高松（2010）对名词语法功能的统计显示宾语、主语、定语是名词最常充当的句法成分，而动名兼类词可充当的主要句法成分中不包括主语。这也归结于动名兼类词略强的动词性质。虽然动词也可以做主语，但要受到种种限制，并非其主要用法（朴重奎，2003）。

高松、颜伟、刘海涛（2010）的统计表明动词常支配从属词构成状语、宾语、主语、复句四种关系。我们的统计认为修饰语、宾语、主语是动名兼类词常支配的成分，类别和顺序都和动词的功能相同，不同点在于复句关系和大致对应于补语的 udep 关系。但实际上，在高文的统计中补语在第五位，我们的统计中对应于复句关系的 parataxis 也在第五位，所以在这一点上区别也不大。从可支配的句法成分上看，动名兼类词与动词也高度相似。高松（2010）的统计中表明，名词最常支配的句法成分只有定语。远不如动名兼类词丰富。

就统计对比结果来看，动名兼类词的主要语法功能大致能覆盖名词，并且与动词的主要语法功能呈现出高度的相似性。

第四节 本章小结

本章从可结合的词、可进入的依存关系两个角度统计了动名兼类词的语法功能。首先，把只有名词性功能和动词性功能的动名兼类词区别为典型动名兼类词，还兼有其他词性功能的区别为非典型动名兼类词，然后把本章的研究对象精确化为频次大于 5 的典型动名兼类词。然后，从词和依存关系两个角度分别统计了动名兼类词可结合的词类与可充当的句法成分，并给词类、句法成分区分了主要次要。每个角度内还分别以类和词为对象进行了统计，所得结果大致相符。动名兼类词的主要功能是支配名词、小品词、动词、介词和从属动词、小品词、名词、助动词，以及支配修饰语、宾语、主语、补语和做宾语、修饰语。还把以上结果与前人和我们对动词、名词的语法功能计量做了对比，发现动名兼类词大致覆盖名词的主要功能，与动词的语法功能还高度相似。

下一章，我们将计算动词、名词的特征向量，并计算动名兼类词与动词和名词的相似度。

第四章 动名兼类词的动名相似度计量

动名兼类词位于动词和名词两个范畴的交叉上，因而兼有两种词类的特征和功能。在掌握动词特征和名词特征后，就可以比较动名兼类词与两个词类的相似程度。本节我们先以两种方法计算动词和名词两个词类的语法功能特征向量，然后比较动名兼类词与动词、名词的相似性。

计算动词和名词的语法功能，同样需要先精确化动词和名词的范围。

第一节 动词和名词的选择

用做对比对象的动名词类语法功能特征向量应当体现词类的特性。放眼于词类内的词，没有一个能不多不少地完美体现该范畴全部的特征（Ungerer and Schmid, 2006; Langacker, 2008）。整个范畴的特征是从范畴内成员的特征身上的出来的抽象集合，因而成员的典型性越高，就越能体现范畴的特征。这种典型性等级不仅存在于范畴交叠处，也存在于词类内部（Langacker, 2008; 金立鑫, 2022），典型性等级较高的成员持有更多该范畴的特征，更具有代表性。换言之，动词的特征应该从典型动词上抽取，名词的特征应该从典型名词上抽取。

参照 3.1 的先例，我们定义词类标签只有动词的词是典型动词，只有名词的词是典型名词。接下来以此二者为对象，计量动词和名词的语法功能并作比较。经统计，GSDSimp 树库中频次大于 5 的典型动词共计 243 个，典型名词共计 765 个。可见统计对象数量较为丰富，提取的特征能代表词类。

第二节 基于词的语法功能计量

将注意力放在典型动词、典型名词可支配或从属的词类，可以计算其语法功能。当把典型动词或典型名词作为一个整体时，先提取所有典型动词或典型名词的支配词和从属词，记录其词类标签，随后统计各种词类标签的概率并向量化即可。

经计算，动词做支配词和从属词时的特征向量分别为：

[0.02900763, 0.00610687, 0.08716169, 0.00652325, 0.20541291, 0.37973629,
0.01623872, 0.08480222, 0.11825121, 0.00152672, 0.00333102, 0.06190146]

[0.19593522, 0.00412829, 0.18132741, 0.00000000, 0.38678946, 0.02731026, 0.00031756, 0.14004446, 0.06160686, 0.00000000, 0.00000000, 0.00254049]

名词做支配词和从属词时的特征向量分别为：

[0.00728070, 0.04201754, 0.00736842, 0.11517544, 0.01833333, 0.36938596, 0.01035088, 0.21245614, 0.02377193, 0.04973684, 0.03175439, 0.11236842]

[0.09414777, 0.01461173, 0.03057920, 0.00180764, 0.29615124, 0.26225804, 0.00060255, 0.21586202, 0.07343526, 0.00030127, 0.00007532, 0.01016796]

为了便于观察和对比，我们将动词和名词的特征向量绘制于图 4.1。特征向量反映的就是词类结合情况。

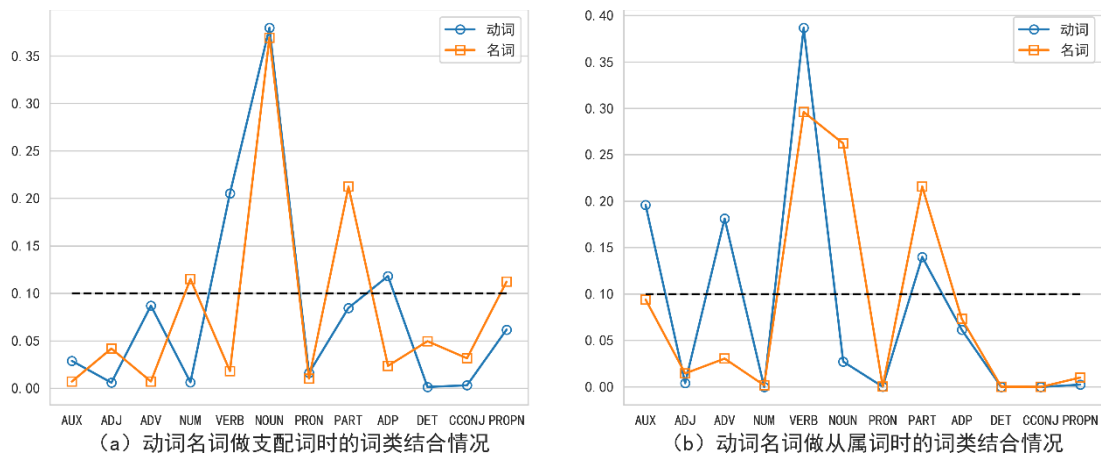


图 4.1 动词和名词的词类结合情况

接下来再以词为对象进一步细致地探讨以上所得动词和名词的语法功能。逐个计算所有典型动词和典型名词的特征向量，具体信息见于图 4.2 和表 4.1⁴。

⁴ 情况同表 3.1-3.4，本章中表 4.1-4.4 同样不在正文中登出，可见于附录 A。

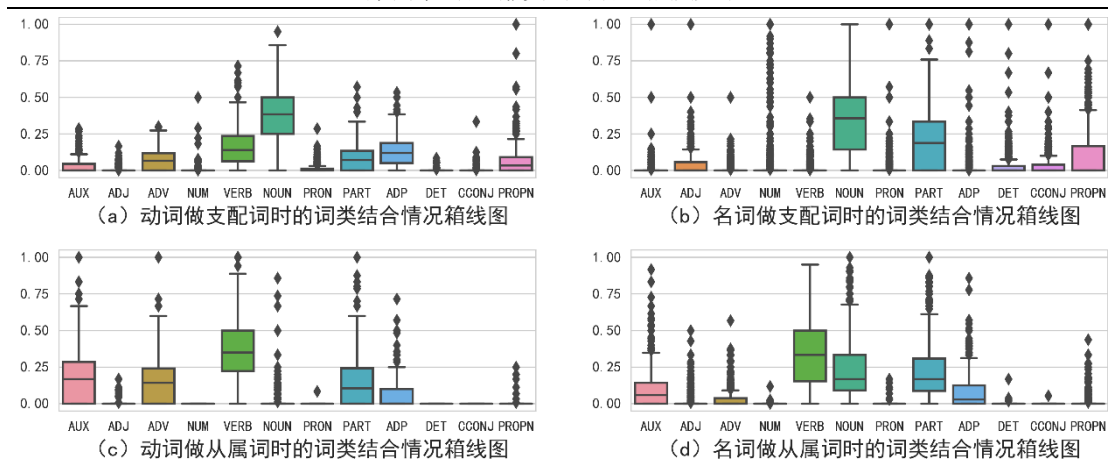


图 4.2 动词和名词的词类结合情况箱线图

图 4.2 可见，动词和名词的支配词、从属词词类结合情况同样有显著差异。结合表 4.2 中的统计检验结果，我们可以区分出动词和名词的主次要语法功能：动词主要支配名词、动词、介词，主要从属动词、助动词、连词、小品词；名词主要支配名词、小品词、专有名词，主要从属动词、名词、小品词。

结合以类为对象和以词为对象的结果，我们得到了动词和名词的主要语法功能，按概率大小，动词可支配的主要词类为名词、动词和介词，其他为次要；可从属的主要词类为动词、助动词、副词（连词）、小品词，其他为次要。名词可支配的主要词类为名词、小品词、专有名词、数词，其他为次要；可从属的主要词类为动词、名词、小品词。

图 4.1 和图 4.2 可见，名词和动词与词类结合的能力是不相同的。参照第三章中，我们仍以 0.1 为主次要功能的分界线。做支配词时，动词主要支配名词、动词、介词，而名词主要支配名词、小品词、专有名词、数词。名词是数量最多且使用频次很高的词类（李媛、段庭辉、刘海涛，2016），是填补头词所开辟空位的最常见成分。动名两词类都以名词为最主要的从属词，这是十分好理解的。动词支配动词和介词都是较为常见的，一棵存在多个动词的句法树（如表 2.2）中，做根的动词支配动词和介词以连接以动词和介词为中心的组块（Lu et al., 2015；刘海涛，2022），从而构造出句法树。当然诸多其他词类也能做组块的中心词，但在概率上，显然语法能力更强的动词和介词更具优势。名词支配小品词、专有名词和数词。汉语中使用频次最多的小品词就是“的”，而且“的”还是汉语中使用频次最高的词。GSDSimp 对“的”字结构（主要是“Adj+的+N”结构）的分析中，“的”常被视作名词的从属词，因此小品词的优势地位是可以理解的。专有名词是种特殊的名词，功能上与名词类似，也具有较高频次。数词从属名词常见于

数量关系，在依存句法分析中，数词与量词都从属于名词。而动词则较少与数词结合，与数词接触需要经过作为数词头词的名词。

做从属词时，动词主要从属于动词、助动词、副词、小品词，名词主要从属于动词、名词、小品词。相同之处是动词和小品词。如做支配词时名词占据最主要地位，动词在两种词类做从属词时占据最主要地位。动词是数量仅次于名词的词类，使用频次也相当高。尽管虚词比动词使用频次高不少，但它们属于封闭类，词的数量远不及动词，因而动词成为最主要支配动词和名词的词。小品词仍是以“的”为主，不过是另一种“的”字结构，动词/名词+“的”。在依存分析中，“的”支配前面的词。接下来谈一谈不同之处。动词还被助动词、副词支配。助动词是谓宾动词，支配动词不足为奇。副词支配动词，这一结果与我们的常识相去甚远。这是由于“ADV”中包括了不少我们常说的连词，比如“还”、“而且”，这些词连接两个句子，句子的头词自然大概率为动词。经实际的检索，也确实如此。检索到的语料绝大多数都是被归入“ADV”的连词。名词从属于名词，这是因为名词除了有强大的从属能力，还有强大的支配能力（Chen et al.,2015），而从当从属词的自然大概率是数量最多且使用最频繁的名词。

依存关系是词与词间不对称有标签的二元关系，体现了词可能充当的句法成分。接下来基于依存关系计量动词和名词的语法功能。

第三节 基于依存关系的语法功能计量

同样先以类对象，先提取所有典型动词或典型名词的依存关系，统计各种依存关系的概率并向量化。

经计算，动词做支配词和从属词时的特征向量分别为：

[0.20792501, 0.00000000, 0.43843204, 0.00610709, 0.20465843, 0.05127113,
0.00000000, 0.08663542, 0.00014203, 0.00028405, 0.00000000, 0.00000000,
0.00340861, 0.00113620, 0.00000000]

[0.00976668, 0.18068367, 0.54150841, 0.01166576, 0.17525773, 0.05887141,
0.00108519, 0.00027130, 0.00054259, 0.02034726, 0.00000000, 0.00000000,
0.00000000, 0.00000000, 0.00000000]

名词做支配词和从属词时的特征向量分别为：

[0.01157885, 0.00000000, 0.00134638, 0.04927744, 0.66690602, 0.00789875, 0.00107710, 0.07880801, 0.03123598, 0.00026928, 0.05385513, 0.06498519, 0.03258235, 0.00008976, 0.00008976]

[0.19186756, 0.00518767, 0.34093683, 0.04317974, 0.29203540, 0.00427220, 0.00015258, 0.00312786, 0.01441868, 0.06110772, 0.04371376, 0.00000000, 0.00000000, 0.00000000, 0.00000000]

将两种词类的特征向量绘制于图 4.3。为了便于观察，仍以动词的各项数值从大到小顺序为横轴。

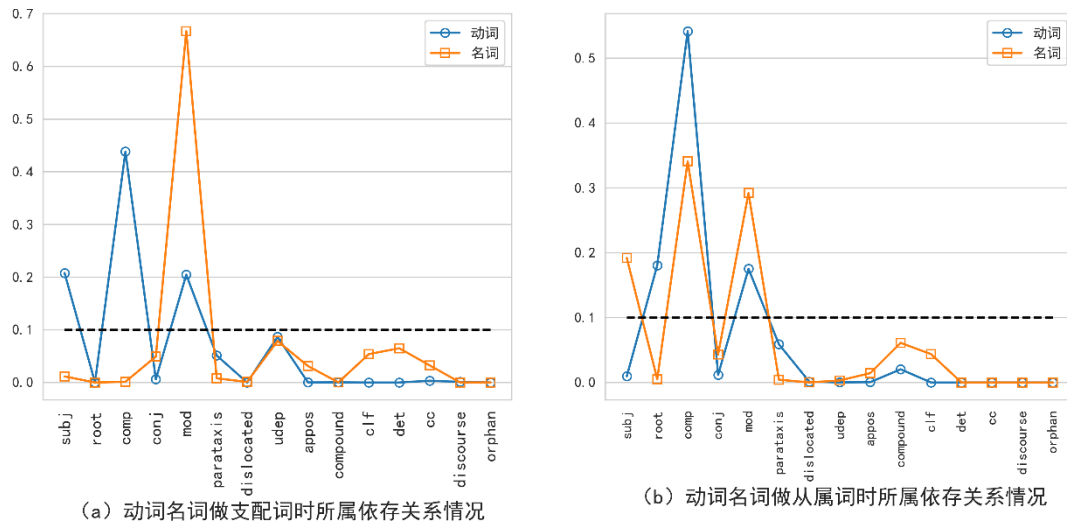


图 4.3 动词和名词所属的依存关系情况

图 4.3 可见，两种词类的所属依存关系概率分布呈现出显著的差异。尽管两种词类的概率分布都是不均衡的，我们把概率大于 0.1 的视做主要功能，小于 0.1 的视作次要功能。可知，动词主要 comp、subj、mod 中的支配词和 comp、root、mod 中的从属词；名词主要做 mod 中的支配词和 comp、mod、subj 中的从属词。

接下来，我们以词为对象，逐个计算其依存关系，以进一步计量和比较动词和名词的句法功能。具体信息见于图 4.4。

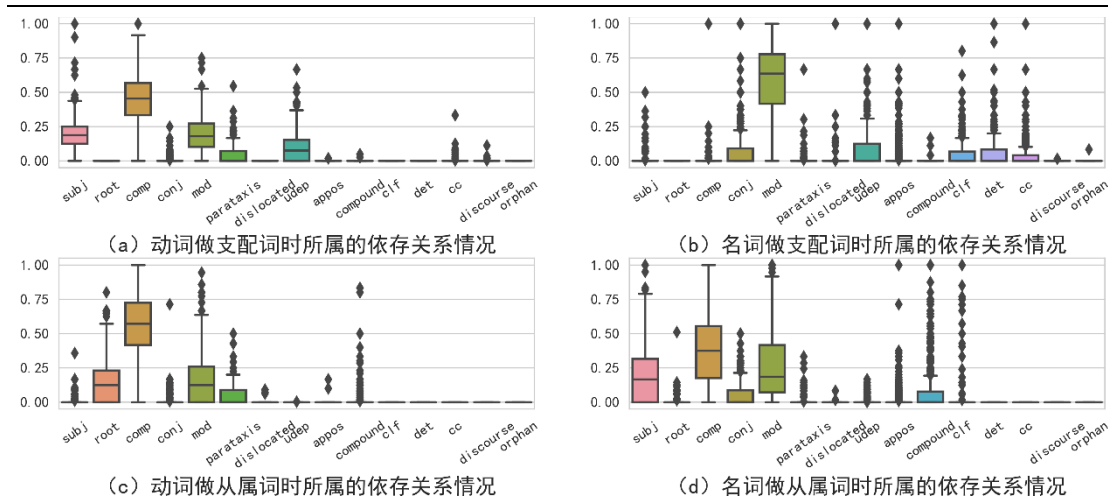


图 4.4 动词和名词所属的依存关系情况箱线图

图 4.3、表 4.3、表 4.4 所得结果与以类为对象时得到的结果一致。总体来看，在依存关系的视角下，动词主要做 comp、subj、mod 中的支配词；主要做 comp、root、mod 中的从属词。名词主要做 mod 中的支配词；主要做 comp、mod、subj 中的从属词。

首先讨论做支配词时的情况。动词和名词相同的主要功能都是做 mod 中的支配词。mod（修饰语）大致对应传统上的偏正关系（最大的不同是依存关系是不对等的），也就是说名词和动词的相同之处是做偏正关系中的中心语。这点和我们的语言常识相符，动词常做状中关系里的中心语，名词通常做定中关系里的中心语。但我们的标注体系里没做出区分，因此在表现上成为了共同点。动词还能做 comp（宾语）、subj（主语）关系中的支配词，这代表动词常做谓语。

再从做从属词来看。动词名词的相似之处在于都常做 comp 和 mod 中的从属词，大致对应动宾关系中的宾语和偏正关系中的修饰语。做宾语时，如“他做了研究”中“研究”充当“做”的宾语；做修饰语时，特别是定语。如“研究人员”中“研究”充当“人员”的定语。动词的独特之处是做 root 里的从属词，这大致对应谓语。名词的独特之处是做 subj 中的从属词，大致对应主谓关系里的主语。

由此可见，动词常做谓语、宾语、修饰语，名词常做宾语、修饰语、主语。莫彭龄、单青(1985)，徐艳华(2006)，高松、颜伟、刘海涛(2010)对动词功能的计量结果大致相同。莫文的统计结果前三位的是谓语、状语、定语，徐文和高文都为谓语、宾语、定语。在细节上，以上研究都区分出了宾语和状语。高松(2010)对名词的语法功能计量结果也符合我们的统计。

就主要的现代汉语教材（黄伯荣、廖序东，2002；胡裕树，1995；邢福义，

1991; 张斌, 2002) 来看, 动词做谓语, 带宾语, 被副词或状语修饰也是共识。但在动词是否能做宾语上, 只有张斌(2002)表明了认可。实际上动词做宾语并不罕见, 特别是做谓宾动词的宾语(朱德熙, 1982), 比如“我想吃饭”中“吃”做“想”的宾语。谓宾动词只能带谓词性宾语, 体谓宾动词可以带谓词性体词性宾语(宋玉柱, 1991)。这两类动词子类都是可以带动词做的宾语。

从词和依存关系两个视角我们都发现, 动词和名词不仅在语法项类别上存在显著差别, 且各项内部数值也差别显著。结合第三章中对动名兼类词语法功能的计量结果, 可证实动名兼类词兼具两种词类的特点。因此, 动名兼类词和动词、名词是有相似性的, 且这种相似性可以以一定的指标和方法计量出来并进行比较。接下来, 我们统计所有典型动名兼类词与动词和名词的相似度, 并做讨论。

第四节 动名兼类词的动名相似度计量

前文可见, 动词和名词两个词类的语法功能特征都有突出的特点, 但动名兼类词因兼有两种词类的特点导致其语法功能特征的特点不如词类那样突出。尽管我们已经计量了动名兼类词、动词、名词的语法功能, 但仍没有将动名兼类词放在动名连续统中观察。动名兼类词是兼有动词性功能和名词性功能的词, 那么对于单个的词来说, 究竟有多少动词或名词的功能呢? 换句话说, 单个的动名兼类词与动词、名词两词类有多接近。

第二章中, 我们已经给出了一种计算两个对象相似度的公式, 即以此二者在 N 维空间的坐标计算其余弦相似度。依次例, 将动名兼类词的语法特征向量视作其在 N 维空间内的坐标, 对应将动词和名词的语法特征向量也依此处理, 计算动名兼类词与动词、名词间的余弦相似度。

前文所有的计量中, 一个对象都有做支配词和做从属词两组向量, 这两组向量都反映了词或词类的部分语法功能, 两组合一才能反映词或词类的完整语法功能。因此我们把作从属词时的向量接在做支配词的向量之后, 合为一个新的向量。

为避免抽象, 我们举“选择”一词为例, 演示如何具体如何计算其动名相似度。在此我们姑且只以词的视角进行演算。“选择”一词的语法功能特征向量如下:

[0.03125000, 0.03125000, 0.00000000, 0.03125000, 0.21875000, 0.28125000, 0.0625000, 0.15625000, 0.09375000, 0.03125000, 0.03125000, 0.03125000, 0.28571429, 0.00000000, 0.09523810, 0.00000000, 0.38095238, 0.00000000, 0.

00000000, 0.19047619, 0.04761905, 0.00000000, 0.00000000, 0.00000000]

第二节中计算所得动词、名词的语法特征向量组合后分别如下：

[0.02900763, 0.00610687, 0.08716169, 0.00652325, 0.20541291, 0.37973629, 0.01623872, 0.08480222, 0.11825121, 0.00152672, 0.00333102, 0.06190146, 0.19593522, 0.00412829, 0.18132741, 0.00000000, 0.38678946, 0.02731026, 0.00031756, 0.14004446, 0.06160686, 0.00000000, 0.00000000, 0.00254049]

[0.00728070, 0.04201754, 0.00736842, 0.11517544, 0.01833333, 0.36938596, 0.01035088, 0.21245614, 0.02377193, 0.04973684, 0.03175439, 0.11236842, 0.09414777, 0.01461173, 0.03057920, 0.00180764, 0.29615124, 0.26225804, 0.00060255, 0.21586202, 0.07343526, 0.00030127, 0.00007532, 0.01016796]

依照公式（2.1），可得“选择”与动词和名词的相似度分别为 0.95 和 0.78。由相似度可见，“选择”与动词和名词都有不低的相似度，但明显更接近动词。

所有的典型动名兼类词的语法特征可以组成一个 324×24 的特征向量矩阵，以矩阵可逐个计算每个动名兼类词的动名相似度。在前文的语法功能计量中，我们一直都区分了基于词和基于依存关系，本节中我们仍然从这两个视角来计算相似度。

计算所有典型动名兼类词基于词和基于依存关系的动名相似度，以动词相似度为纵轴，以名词相似度为横轴绘制于图 4.5。

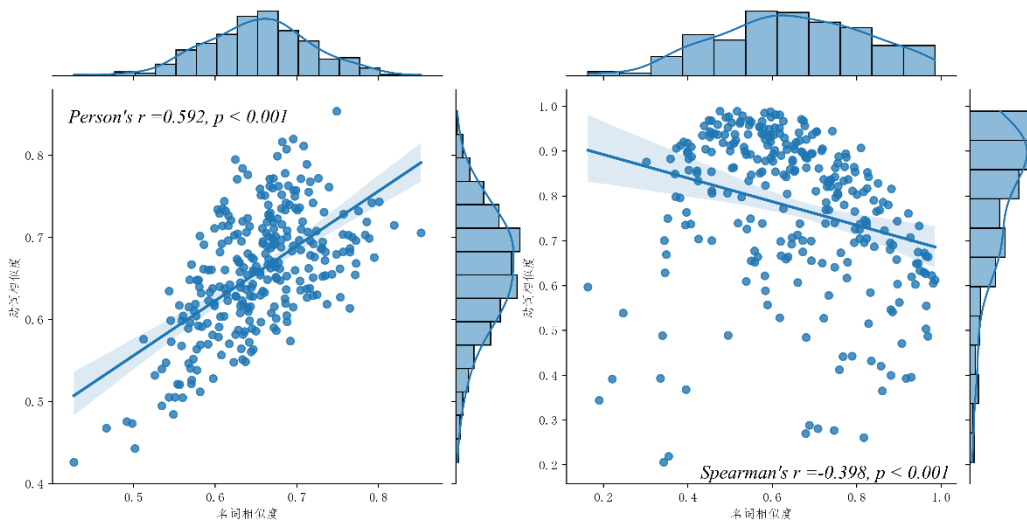


图 4.5 (a) 基于词的典型动名兼类词动名相似度; (b) 基于依存关系的典型动名兼类词动名相似度

图 4.5 可见, 两种视角下的动名相似度关系并不一致。子图 (a) 基于词的视角中, 动名兼类词的动词相似度和名词相似度成中度正相关 (Person's $r = 0.592$, $p < 0.001$), 这意味着若词的动词相似度高, 其名词相似度也倾向于高; 而子图 (b) 基于依存关系的视角中, 动名兼类词的动词相似度和名词相似度成中度负相关 (Spearman's $r = -0.398$, $p < 0.001$), 这意味着若词的动词相似度高, 其名词相似度倾向于低。无论是正相关还是负相关都能反映出典型动名兼类词与动词和名词都相对较高的相似度。这一点还能通过观察相似性大于 0.5 的对象占比来佐证: 子图 (a) 中有 97.6% 的词动词相似度大于 0.5, 98.6% 的词名词相似度大于 0.5; 子图 (b) 中有 91.3% 的词动词相似度大于 0.5, 79.4% 的词名词相似度大于 0.5。典型动名兼类词与动词和名词的相似程度绝大多数都很高, 可见相似度指标忠实地反映出了这一点。

动词相似度和名词相似度的关系是有趣的。Crystal (1967) 指出有一些词既有甲类词的特征也有乙类词的特征, 是甲乙两类词之间的“桥梁词类”。Ross (1973) 曾指出, 有些动词比其他动词更接近动词, 有些名词比其他名词更接近名词, 存在“名/动连续统”。“连续统”这种说法一般指的是甲乙两个范畴间存在中间状态, 从甲逐渐地过渡到乙。比如 Labov (1986) 的语体风格连续统, 说话人的语体从谨慎到放松是逐渐过渡的。连续统中甲范畴的特征越多, 乙范畴的就越少, 反之亦然。放到动名兼类这个问题中, 动词范畴和名词范畴是连续的, 中间存在动名兼类词作为过渡, 且动名兼类词存在有的相对更接近动词, 有的相对更接近名词。从动词到名词, 动词相似度递减而名词相似度递增——这是一种负相关关系。

子图(a)中有 97.6% 的动名兼类词的动名相似度都达到 0.5 以上。这意味着, 动名兼类词有高动词相似度的同时也拥有高名词相似度。这与“动名连续统”的推论是不符的——一个词不能同时出现在连续统两头。而子图 (b) 则只有 73.4% 的动名兼类词的动名相似度都大于 0.5, 再加之动名相似度的负相关关系, 基于依存关系的结果更符合“动名连续统”的假设。

总体上看, 基于词和基于依存关系都较好地反映了典型动名兼类词与动词、名词的相似程度。但这种指标能否较好地捕捉非典型动名兼类词以及其他词类的词需要进一步校验——相似度指标应当能捕捉当计量对象中掺进去典型动名兼类词之外的内容时的变化——此时的数据应当与以动名兼类词为对象所得的相似性数据有较大差异。接下来, 我们以典型动词和典型名词为对象计算动名相似

度。处在动名连续统两头的词类应当在相似性分布上有相当明显的偏好。为了方便观察，我们把不同对象以不同颜色在图 4.6 中标识。

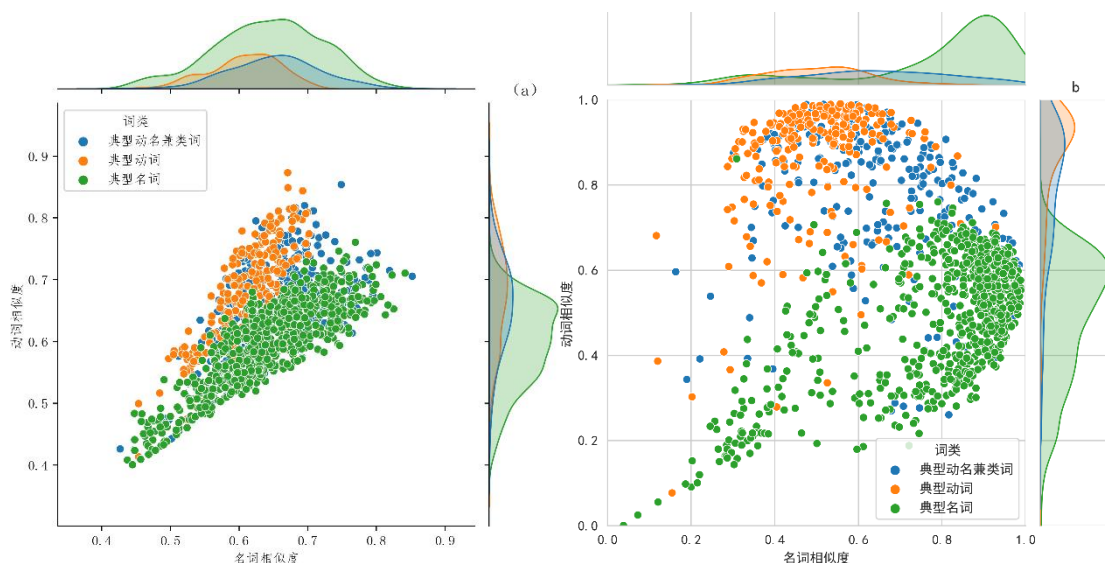


图 4.6 (a) 基于词的典型动名兼类词、典型动词、典型名词的动名相似度；(b) 基于依存关系的典型动名兼类词、典型动词、典型名词的动名相似度

图 4.6 中每个点都代表一个词，横纵坐标为该词的动名相似度。由于我们已经设定了有三种词——典型动名兼类词、典型动词、典型名词，所以可以根据它们的数据结果来评估基于词和基于依存关系两种方法的优劣。典型动词拥有典型的动词用法，其动词相似度都应该远高于其名词相似度；同样典型名词的名词相似度也应远大于其动词相似度。而典型动名兼类词介于动词和名词之间，其动词相似度应该小于典型动词而大于典型名词，其名词相似度应大于典型名词而小于典型动词。经比较，子图 (a) 中，典型动词的动词相似度显著大于其名词相似度 ($p < 0.001$)，典型名词的名词相似度显著大于其动词相似度 ($p < 0.001$)；子图 (b) 中同样典型动词的动词相似度显著大于其名词相似度 ($p < 0.001$)，典型名词的名词相似度显著大于其动词相似度 ($p < 0.001$)。这说明，基于词和基于依存关系都能捕捉到动词和名词的特征，典型动词和典型名词切实存在区别。再把典型动名兼类词纳入比较范围。子图 (a) 中典型动名兼类词的动词相似度显著小于典型动词 ($p < 0.001$)，且显著大于典型名词 ($p < 0.001$)；名词相似度显著大于典型动词 ($p < 0.001$)，且显著大于典型名词 ($p < 0.001$)。子图 (b) 中典型动名兼类词的动词相似度显著小于典型动词 ($p < 0.001$)，且显著大于典型名词 ($p < 0.001$)；名词相似度显著大于典型动词 ($p < 0.001$)，且显著小于典型名词 ($p < 0.001$)。虽然都 $p < 0.001$ ，但子图 (a) 中的平均值之

差却远不及子图 (b) 中, 加之考虑到动名连续统的假设, 基于依存关系的方法捕捉语法信息的能力更强大。

无论是从“动名连续统”假设的支持程度还是典型词类检验结果来看, 基于依存关系的方法都明显表现更好。我们认为有两方面的原因。一方面是特征数量的不同, 词类共计 22 个特征, 依存关系共计 30 个特征。从数学上来看, 特征越多描述能力就强, 描写的内容也就越精细, 因此得到更接近事实的相似度。另一方面则是词类与依存关系的本质性的不同。依存语法认为, 依存关系是语言的基本要素 (Tesnière, 1959; 刘海涛, 2009; Hudson, 2010)。依存关系是二元不对称的, 有标签的, 这说明关系双方有主有从, 关系有类型的区分。依存关系实质上反映的就是从属词做支配词的什么句法成分。而词类是根据词能充当的句法成分聚合而成的模糊的类, 词类本身并不代表词的具体语法功能, 只是给了一个大概的范围。由此可见, 依存关系反映的内容比词类更加细致且切合语言本质。这两方面的因素叠加, 就是的基于依存关系的方法取得了更好的结果。

第五节 本章小结

本节中我们同样以词和依存关系为角度计量了动词和名词的语法功能, 并根据量的差异区别了主要功能、次要功能和其他功能, 动词和名词各有其特色的用法。基于词的视角下, 动词主要支配名词、动词、介词, 主要从属动词、助动词、副词、小品词; 名词主要支配名词、小品词、专有名词、数词, , 主要从属动词、名词、小品词。在依存关系的视角下, 动词主要做 comp、subj、mod 中的支配词; 主要做 comp、root、mod 中的从属词。名词主要做 mod 中的支配词; 主要做 comp、mod、subj 中的从属词。

经过比较, 动名兼类词确实兼具动词和名词两种词类的功能。我们进而计量了动名兼类词与两种词类的相似度。经过比较, 基于依存关系的方法优于基于词的方法。典型动词的动词相似度最高, 名词相似度最低; 典型名词的动词相似度最低, 名词相似度最高; 而典型动名兼类词的两种相似度都居于中位。典型动名兼类词的动词相似度和名词相似度呈负相关, 证实了“动名连续统”的假说。

相似度指标给词类的“柔性处理” (史有为, 1994) 提供了一种切实可行的方案, 它可以计算出词与词、词类与词类、词与词类的关系。相似度不仅可以用来处理理论语言学中的词类关系问题, 这种连续化、定量化的方法还可以进一步发散用来评估语言学习者在具体语言项目上的习得情况, 这将在下一章中讨论。

第五章 基于相似度的汉语中介语动词、名词配价习得情况 计量

语言是渐变的 (Altmann, 1983; Wang, 1969; Langacker, 1987), 语言习得 (包括第二语言习得) 作为语言演化的微观尺度 (石锋, 2018) 自然也符合这一点。中介语假说 (Selinker, 1972) 认为, 在习得过程中, 学习者接受目的语输入后会形成一种既不同于其母语, 也不同于目的语的语言, 即中介语。中介语会随着学习进度逐步向目的语靠拢。换一种说法, 从初学至习得, 中介语与目的语的相似程度整体上是越来越高的。具体到词类上, 以母语者为参照, 学习者的词类相似度整体上是越来越高的。第二章引入了概率配价模式理论 (刘海涛、冯志伟, 2007), 作为一种广义配价理论, 概率配价模式理论可以更好的展现词或词类总体的结合能力, 而不必拘泥于必有成分和可有成分 (Tesnière, 1959) 之分。基于该理论, 高松等 (2010) 和高松 (2010) 分别计算了汉语动词和名词的概率配价模式, 并区分了两词类的主要功能和次要功能; 郝瑜鑫等 (2021) 等采用计量视角研究了不同年级英语母语学习者的汉语动词配价发展情况; 郝瑜鑫等 (2022) 定量分析了四个年级汉语名词配价的习得情况。相比于二语习得中的传统方法, 定量配价的视角能更加清晰细致地追溯到习得的动态过程。但以上研究更着眼于具体的细节, 即更偏好词类所处的具体依存关系情况。那么如何从词类的全局视角评估其配价习得情况呢?

第三章和第四章的结果说明, 相似度可以作为词与词类相似程度的计量指标。当我们把相似度的比对双方由动名兼类词和词类替换为中介语的词类与目的语的词类, 再按照第二章中的方法即可评估学习者对某词类的习得程度。本章我们将借助有多个习得等级的汉语中介语树库, 观察学习者习得汉语不同词类的整体趋势。

第一节 资源和方法

目前中介语树库仍然较少, 汉语中介语树库则更为稀有。SUD 中的汉语中介语树库 CFL 规模较小且多不区分习得阶段, 难以满足我们的需求。本章中我们使用的树库⁵包括七个等级的汉语学习者的中介语, 语料来自学习者的作文。

⁵ 此树库来于华侨大学郝瑜鑫教授, 特此感谢郝教授的无私帮助。

七个等级依照作文的分值设定，分别为：40-55、60、65、70、75、80、 ≥ 85 ，共计 2262 句，44047 词。

该树库标注参照刘海涛（2009：106）中的标注体系（“现代汉语依存关系”，以下简称为 MCDR），此外还增加了部分表示偏误的标注。该树库的标注体系同样按照依存语法，但与 SUD 有一定差异。为了方便比较，我们将 MCDR 中的依存关系转换为 SUD 中的依存关系，具体可见表 5.1。为了便于转换，没有考虑 SUD 中语义层的相关信息，即第二章中以“@”为首的信息。

表 5.1 MCDR 与 SUD 中依存关系的转换

MCDR		SUD	MCDR		SUD
s	谓语	root	qc	量词补足语	mod
subj	主语	subj	adva	状语	mod
obj	宾语	comp:obj	atr	定语	mod
obj2	间接宾语	comp:obj	auxr	助词附着关系	comp:aux
subobj	兼语	subj	coor	并列关系	conj
sentobj	小句宾语	comp:obj	cr	复句关系	mod
obja	助动词宾语	comp:aux	csr	连带关系	mod
comp	补语	udep	epa	同位语	appos
soc	兼语补语	comp:obj	esa	句末附加语	discourse
pobj	介词宾语	comp:obj	fc	方位结构补语	udep
baobj	“把”字句宾语	comp:obj	ina	插入语	mod
beis	“被”字句	mod	ma	数词结构	mod
dec	“的”字结构补足语	comp:obj	ta	时态附加语	mod
dic	“地”字结构补足语	comp:obj	top	主题	subj
dfc	“得”字结构补足语	comp:pred	va	连动句	mod
plc	名词复数	compound	c-	并列词	cc
oc	序数补足语	mod			

经过转换，我们就可以进行比对。由于本章我们聚焦在具体词类的习得情况，因此要采用第三章和第四章中以类为对象的方法。首先计算目的语词类的特征向量，然后分别计算七个等级的中介语词类的特征向量，最后依次计算中介语七个水平的词类与母语者词类的相似度并分析讨论即可。

本章中的特征与第三、四章略有不同。由于不把深层语义信息纳入考虑范围以及标注体系的差别，两种体系中的部分依存关系难以一一对应，因此对特征进行删减，只取以下特征值并归一化作为比对项。特征依次为：subj（主语），root（根），comp:obj（补足语：宾语），comp:aux（补足语：助动词），comp:pred（补足语：预格），conj（连词），mod（修饰语），udep（属性或属格），appos

(同位语修饰), compound (合成词), cc (并列词)。

我们将依此得来的特征向量记为 $vec1$ 。以往研究(刘海涛、冯志伟, 2008; 高松、颜伟、刘海涛, 2010; 高松, 2010; 郝瑜鑫、王雪琳、刘海涛, 2021; 郝瑜鑫、王燕子、刘海涛, 2022)和第三、四章都表明, 词类做支配词时与做从属词时的语法功能有很大差异。因此除了以不区分支配从属的向量 $vec1$ 计算相似度外, 还以支配向量 $vec2$ 、从属向量 $vec3$ 计算区分这两种情况时的习得情况。依这三种向量计算可以得到相似度, 我们将三组动词相似度对应命名为 $v1$ 、 $v2$ 、 $v3$, 三组名词相似度对应命名为 $n1$ 、 $n2$ 、 $n3$ 。

比如, 母语者的动词做支配词时的特征向量为:

[0.22202002, 0.00000000, 0.38974826, 0.06111617, 0.00000000, 0.00652108, 0.21853200, 0.09660297, 0.00015165, 0.00045496, 0.00363967, 0.00121322]

40-55 分等级的中介语动词的特征向量为:

[0.18324982, 0.02863278, 0.26163207, 0.03328561, 0.00000000, 0.00644238, 0.42090193, 0.02648533, 0.00286328, 0.01002147, 0.02040086, 0.00608447]

因而该等级上相似度约为 0.796。

在计算词类时, 我们不再区分典型与非典型, 而是统计在句中出现时为指定词类的情况作为替代。这样选择的原因是一方面我们不再需要格外关注词与词类之间的关系, 另一方面可以扩大统计数量, 从而更加全面地反映词类的习得情况。

第二节 动词配价和名词配价的习得情况计量

动词和名词是汉语的主要词类, 本节中计算这两种词类的习得情况。七个等级的中介语与目的语的动词名词相似度可见于图 5.1。

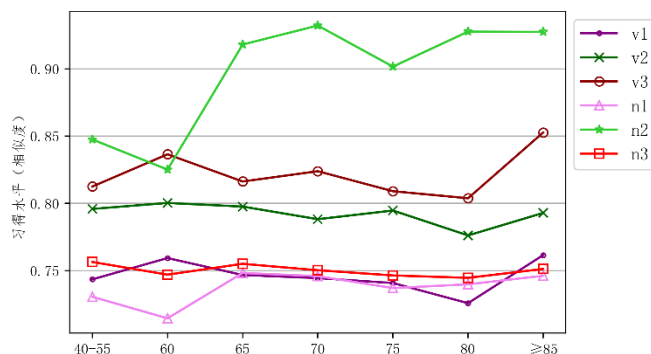


图 5.1 学习者与母语者的动词、名词配价相似度

总体来看，中介语的动词、名词配价习得有如下特点：

(1) 动词与名词的整体配价（不区分支配从属，即 $vec1$ ）习得水平没有显著差异（ $p_{v1-n1} > 0.05$ ）。做支配词时（即 $vec2$ ），动词配价习得水平低于名词配价（ $mean_{v2} < mean_{n2}, p_{v2-n2} < 0.05$ ）；做从属词时（即 $vec3$ ），动词配价习得水平高于名词配价（ $mean_{v3} > mean_{n3}, p_{v3-n3} < 0.05$ ）。

(2) 动词和名词配价内部的习得水平也存在差异。动词、名词的整体配价习得水平都是最低的（ $mean_{v1} < mean_{v2} \& mean_{v3}, p_{v1-v2} \& p_{v1-v3} < 0.05; mean_{n1} < mean_{n2} \& mean_{n3}, p_{n1-n2} \& p_{n1-n3} < 0.05$ ）。动词做从属词时配价习得水平高于做支配词时的习得水平（ $mean_{v2} < mean_{v3}, p_{v2-v3} < 0.05$ ）；而名词做从属词时配价习得水平则低于做支配词时的习得水平（ $mean_{n2} > mean_{n3}, p_{n2-n3} < 0.05$ ）。

(3) 动词、名词配价的习得并不是个线性的过程，整体配价、做支配词时配价、做从属词时配价都有回退、化石化、跳跃式进步等现象。经一元一次回归分析，六列数据所得结果的 p 值都小于 0.95（ $v1: y = -0.001x + 0.748, R^2 = 0.015; v2: y = -0.002x + 0.799, R^2 = 0.333; v3: y = 0.001x + 0.817, R^2 = 0.046; n1: y = 0.003x + 0.728, R^2 = 0.315; n2: y = 0.013x + 0.851, R^2 = 0.586; n3: y = -0.001x + 0.753, R^2 = 0.251$ ）。这说明线性模型并不能适配图中的习得过程。

(4) 第二级（60分）和第六级（80分）是两个重要的节点，两等级前后的变化最为显著。具体说，动词的整体配价、做支配词时配价、做从属词时配价于第六级前后发生较大变化；名词的整体配价、做支配词时配价、做从属词时配价于第二级前后发生较大变化。

(5) 一方面，相似度总是在 0.7 以上，并不是从 0 开始变化；另一方面，相似度即使在最高分段习得情况最好的项目（名词做支配词时的配价）上也没有达到 1。

从以上的几条特点可以看出,动词与名词配价的习得情况是有较清晰的区别的,而且词类配价内部的支配与从属也存在不同。为何配价的习得会有以上特点?我们接下来尝试从目的语本身的特点入手分析。

第三节 讨论

首先来看动词和名词两词类间的配价习得情况对比。有学者指出,多数名词与外语对应性很强,名词习得总体来看不是难点;而动词的表义对应差异较大以及更多的句法使用限制使得学习者会产生一系列偏误(卢福波,2010)。这似乎是在说动词的习得难度要大于名词。但仅就配价而言,动词整体配价和名词整体配价的习得水平并没有显著差异,这也说明从整体来看,两词类配价的习得难度不相上下。但在区分支配和从属后,动词配价和名词配价习得表现出了显著差别。做支配词时动词配价习得难度大于名词,做从属词时则相反。

这种结果可以从词类的价入手考察。词类的平均价(Yan and Liu, 2022)是不同的,词类 w 的平均价可以以公式 5.1 计算而来:

$$\text{mean valency } (w) = \frac{1}{n} \sum_{i=1}^n |W_i| \quad (5.1)$$

公式中, n 是词类为 w 的词的数量, W_i 是第 i 个词类为 w 的词的从属词的数量。比如,表 2.2 句子“随后爱斯基摩人和维京人相继定居于此。”中只有一个动词“定居”,过滤掉标点符号后,它支配“随后”、“人”、“相继”、“于”,共计四个词,因而在这个句子中动词的平均价为 $4/1=4$ 。计算结果表明,目的语中动词平均价大于名词平均价($1.58 > 0.84$),这意味着动词可能支配更多数量的词,比如动词更经常与“着”、“了”、“过”等表示时体的助词结合(戴耀晶,1997),还可以与表达情态的情态动词结合(彭利贞,2007)。即使词义相近,使用价大的词时也要在补足语位置、补足语数量、补足语是否能省略上做更多考虑(方绪军,2003)。这自然会加大使用和习得动词做支配词时的难度。

具体来看,动词和名词的价分布是不同的。目的语中动词和名词的价分布如图 5.2。

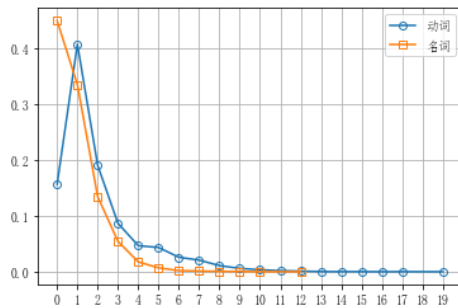


图 5.2 动词和名词的价分布

由图 5.2 可见，动词和名词的价分布有显著差异，动词价服从的最佳分布为柯西分布 ($loc = 1.106, scale = 0.599, ksvalue = 0.0$)，名词价服从的最佳分布为伽马分布 ($shape = 0.564, loc = 0.000, scale = 0.833, ksvalue = 0.0$)。最明显的差别在于零价词，动词的零价词占比低于一价词和二价词，而名词的零价词占比则最大。零价词实际上就是在句子里只做从属词，而不支配其他词的词。而从一价到最高价，动词价的占比一直大于等于名词。这种不同也造成了作支配词时动词配价习得难度大于名词配价。

但平均价和价分布实际上都是以支配词为中心的指标，因而对做从属词时动词名词配价习得难度的差异解释能力有限——这是我们仍需解释的点。从前文的平均价其实能够看出，动词的支配能力更强 ($1.58 > 1$)，而名词的从属能力更强 ($0.84 < 1$)。这一点可以从词类的主次要功能中看出来，第四章中统计显示，做支配词时，名词的主要功能只有充当 mod 关系中的支配词；而做从属词时，主要功能为充当 comp、mod、subj 关系里的从属词，高松等 (2010)、郝瑜鑫等 (2022) 的统计也支持以上结果。做从属词时的依存关系类型明显更为丰富，因而作为重点学习的内容会更多。此外，名词做从属词时，还有名词做状语 (孙德金, 1995; 李锡江、刘永兵, 2020; 李雯雯、王文斌, 2022) 等特殊用法。这些因素都提高了名词做从属词时配价的习得难度。

动词名词配价的习得难度差异还可以在依存距离上得到佐证。依存距离是存在依存关系的两个词在句子里的线性距离，常用于计算语法结构的复杂性 (Liu, 2008)。依存距离有最小化倾向 (Liu, 2008; Temperley, 2008; Ferrer-i-Cancho, 2014, Liu et al., 2017)，依存距离越长，其复杂性和认知难度就越大 (Jiang and Liu, 2015; Lu et al., 2015)，依存距离最小化反映了人类认知能力的制约，自然也包括语言能力，因此我们可以用这一指标评估动词配价和名词配价的习得难度。词类 w 的平均依存距离 (Mean Dependency Distance, MDD) 可以由公式 5.2

计算得出：

$$MDD_1(w) = \frac{1}{n-1} \sum_{i=1}^n |DD_i| \dots \dots \dots (5.2)$$

其中 n 是树库中词类为 w 的词的数量， $|DD_i|$ 是支配词词性为 w 的第 i 个依存关系中的依存距离。比如表 2.2 句子“随后爱斯基摩人和维京人相继定居于此。”中，“定居”支配“随后”，依存距离为 7；支配“人”，依存距离为 5；支配“相继”，依存距离为 1；支配“于”，依存距离为 1；因此其 MDD_1 为 $(7+5+1+1)/4 = 3.5$ 。

但如同平均价，通常意义上的 MDD （即 MDD_1 ）实际上是从支配词出发，遍历其从属词，求取线性距离得来。在比对支配和从属时，我们还需要从从属词出发的指标，公式 5.2 无法满足这一需求。当我们从从属词出发，找到其唯一的支配词，就可以求取到一个偏好从属词的线性距离（若没有支配词，则为 1），可见于公式 5.3⁶。

$$MDD_2(w) = \frac{1}{n-1} \sum_{j=1}^n |DD_j| \dots \dots \dots (5.3)$$

其中 n 是树库中词类为 w 的词的数量， $|DD_j|$ 是从属词词性为 w 的第 j 个依存关系中的依存距离。比如表 2.2 句子中，副词“随后”和“相继”都从属于“定居”，依存距离分别为 7 和 1，因此在此句中副词的 MDD_2 为 $(7+1)/2 = 4$ 。

为了区别，我们把偏好支配词的 MDD 称为 MDD_1 ，偏好从属词的称为 MDD_2 。经计算，目的语中动词的 MDD_1 为 4.98， MDD_2 为 4.19；名词的 MDD_1 为 1.92， MDD_2 为 3.06。结果显示，动词的 MDD_1 大于名词的 MDD_1 ，动词的 MDD_2 小于名词的 MDD_2 。这说明，做支配词时动词配价的习得难度大于名词配价；而做从属词时，动词配价的习得难度小于名词配价，与我们在 5.2 节中得到的结果一致。两种词类在做支配词和从属词时表现出了不同的 MDD ，更长的 MDD 造成了更高的习得难度。

其次在词类内部配价上，习得水平也存在差异。虽然整体配价都是习得水平最低的项目，但动词做支配词时的配价习得水平低于做从属词时，名词则相反。配价就是词潜在的结合能力，根据其方向可以分为向心和离心（Liu, 2009），因而配价有支配和从属之分。整体配价考虑了支配、从属两个方面，因此相对于考虑单一情况时更为复杂，习得的难度也更大。如果说整体配价的习得难度与做支配词、从属词时配价的习得难度的差异更多是数学上的问题，那么做支配词、从属词时配价习得难度的不同显然与语言本身更为密切。

我们仍可以 MDD 中管中窥豹。前文的统计结果显示，动词的 MDD_1 大于

⁶ 这其实也只是对 Liu（2008）中 MDD 公式的化用。

MDD₂ (4.98 > 4.19), 而名词的 MDD₁ 则小于 MDD₂ (1.92 < 3.06)。MDD 越大, 就越违背省力原则 (Zipf, 1949), 大脑在处理时就需要更大的认知负担和工作记忆容量 (Xu and Liu, 2022), 进而造成习得困难。结合平均价的表现, 我们认为词类内部配价的差异实际上直接对应于词类本身的配价能力特点: 动词支配能力强, 而名词从属能力强。动词充当支配词需要考虑诸多补足语的语序位置、能否省略 (方绪军, 2003), 以动词为中心的词语搭配和框式结构一直是动词习得的难点 (冯丽萍、盛双霞, 2004; 肖奚强、周文华, 2009; 杨圳、施春宏, 2013)。而名词做支配词时, mod 关系 (大致对应传统上的偏正关系) 占了绝对主流, 需要考虑的情况很有限。在做从属词时, 可进入的主要关系类型 (comp、mod、subj) 更为丰富 (蔺璜, 2005; 郝瑜鑫等, 2022), 这说明名词可以充当的补足语类型丰富, 甚至兼有可有补足语和必有补足语两类。因而在使用时需要考虑得更为复杂, 这也加大了习得的难度。

最后, 二语习得充满回退、停滞、跳跃式进步等现象 (Larsen-Freeman, 1997), 动词和名词的配价习得都表现出这一点。习得水平的各种波动都说明了动词和名词的配价习得是动态的、复杂的 (Larsen-Freeman and Comeron, 2008), 习得水平并不是随时间而一直提高。动词做支配词时的配价在各个阶段的变化最不明显 ($std = 0.004$), 呈现出停滞。这可能与学习者的母语迁移有关, 其他语言与汉语的动词对应性较差 (卢福波, 2010)。比如汉语中主语还可以省略或者缺失 (吕叔湘, 1986; White, 1987), 动词可以支配趋向补语 (杨德峰, 2003) 等, 这些汉语习得难点都可能导致配价习得的停滞。动词做从属词时的配价习得发展波动最显著 ($std = 0.017$), 呈现出“N”形发展。习得过程中有诸多因素会影响习得水平, 进而产生波动。学习者可能受母语迁移, 在汉语中动词能够做主语和宾语 (汪涛, 2011), 而且主语宾语有时候还可以省略 (沈阳, 1994; 刘正光, 2007)。而学习者母语中主语宾语受的限制较大 (Li and Thompson, 1976), 汉语宾语出现与否则更为灵活 (吴静、石毓智, 2005)。学习者还可能采取一些可能阻碍习得学习策略和产出策略 (Elis, 1997), 比如习得使役构式时出现的过度泛化 (王保昌等, 2022)。但适当的教学策略 (赵金铭, 2010) 和纠错练习又可以帮助提升习得水平。因此在诸多因素的动态博弈下, 习得水平也随之表现出动态特征。

60 和 80 两等级前后的变化最为显著。我们以公式 5.4 计算变化的显著程度 (significant degree):

$$\text{significant degree} = |S_i - S_{i-1}| + |S_i - S_{i+1}| (i \in [2,6], i \text{ 为整数}) \dots\dots (5.4)$$

在此式中， i 为等级的顺序， S_i 为第 i 个等级上的相似度（即习得水平）。经计算，动词在 80 分等级前后变化最大，名词则在 60 分等级前后变化最大。具体而言，动词从 80 分到 ≥ 85 分这个阶段上出现跳跃式发展，而名词则在 40-55 分到 60 分阶段上出现了倒退，随后又在 60 分到 65 分阶段上发展，因而呈现出“U”形。这显然与词类的差别有关系。

动词配价能力强，还可以带上介词短语、补语、时态附加语等附加成分，习得难度大（杨圳、施春宏，2013）。因而在 80 分等级前，习得水平较差。成绩在中低分段时，学习者常会选择采取回避、泛化、换用上位用法等策略（刘颂浩，2003；张宝林，2010），这会导致使用不足，进而可能引起“化石化”（Selinker，1972），使得动词配价在中低分段的习得水平出现停滞甚至倒退。但随成绩提高，一方面学习者会经历越来越多的练习和纠正（郝红艳，2015）；另一方面，学习者也会对关键词汇和语法结构有更深刻的理解（Swain and Lapkin，2000），进而语言能力会在高分时有跳跃式的提高。而名词配价能力相对较弱，并且与学习者母语的对应性较强，因而在低分阶段就能良好习得。随成绩提高和学习进度深入，一些较难的用法开始出现，照搬母语经验已经无法追上学习进度，比如汉语二价名词习得中重点难点在于语块配位（蔡淑美、施春宏，2014），而母语知识无法起到太大帮助，因此在 60 分时出现明显的倒退。但经过系统的训练和纠正，学习者又可以较好的处理这些偏误，而动词的这个过程要远滞后于名词。

相似度总是在 0.7 以上，并非从 0 开始变化。这是在说，二语习得和母语习得是不同的，二语习得并非是从头开始。从 L2 语言网络表现出的复杂性来看，L2 一开始就有与母语者语言网络相同的特征——无尺度特性（Jiang et al. 2019），而这一点并没有出现在母语习得之中（Solé, 2005; Corominas-Murtra et al., 2009）。这与我们的结果是相符的。这种现象可以归结于二语习得中学习者是有语言知识背景（即母语）的。学习者借助母语与目的语间的句法相似性，在学习一定的词汇后就可以产出句子。比如英语母语者可以较轻松地理解 SVO 语序的句子（寮菲，1998）。母语的负迁移使得低分的学习者也能产出基础的目的语句子。

相似度始终不能达到 1。这说明，学习者的动词配价和名词配价始终无法达到母语者水平。母语的迁移不总是正向的，负迁移制约着语言能力的发展，学习者还是倾向于使用与其母语相同的结构产出句子，即使到了高分段，负迁移仍然发挥着作用。比如少见于其他语言的补语结构就是个难点，趋向补语的习得了高级阶段仍然有较大问题（杨德峰，2003）。

第四节 本章小结

随着学习进度深入和语言能力的提高,学习者的中介语会逐渐向目的语过渡,与目的语的相似性会越来越高。借助目的语依存树库和中介语依存树库,我们提取了其中的动词和名词概率配价模式并转化为了三组特征向量,通过计算概率配价模式间的相似度,对动词、名词配价的习得水平进行了定量分析,有以下几点发现。首先,动词配价和名词配价组间的习得水平是不同的。从句法层面来看,这要归结于两种词类本身不同的配价表现和依存距离。动词支配其他词的能力比名词更强大,更可能有更多的句法结构和语法表现。动词的 MDD_1 也远大于名词,做支配词时的习得水平低于名词。而名词的从属能力强, MDD_2 大于动词,因而做从属词时名词配价习得水平更低。其次,动词配价和名词配价组内的习得水平也存在区别。动词、名词的整体配价习得水平都是最低的。动词做从属词时配价习得水平高于做从属词时的习得水平,而名词做从属词时配价习得水平则低于做支配词时的习得水平。兼顾支配、从属的配价习得难度最大的成因更多来自数学上,而做从属词时配价和做支配词时配价的习得难度差别来自与词类配价能力和依存距离的差异。再次,动词配价和名词配价的习得都不是线性的过程,而是充满着倒退、停滞、跳跃式进步的动态过程。这可能来自多方面的因素,如母语迁移、学习者策略等。再者,动词配价和名词配价的习得都存在关键节点,但节点的位置不同。动词配价在 80 分到 ≥ 85 分段出现明显的跳跃式进步,而名词配价在 60 分前后出现“U”形发展。前者由于难度较大,学习者在高分段才会出现明显的进步;而后者的难度较小,在初学时就可以用母语知识产出一些正确的句子,但随着学习深入,母语知识已经不足以适配学习者的进度,当受到纠正训练后又会出现明显的进步。最后,无论哪个阶段,配价习得水平都不低于 0.7,但又没有达到 1。学习者在母语正迁移作用下,已经有了一定的 L2 基础,但这又导致他们会依赖于母语知识,加之母语负迁移作用,会对 L2 的习得产生制约。

语言是个人驱复杂适应系统 (Cong and Liu, 2014),这也体现在二语习得的复杂性上 (Larsen-Freeman, 1997)。L2 习得水平是学习内容本身、学习者母语迁移、学习者策略等诸多因素综合博弈后的结果。本章中,我们给出了一种从全局视角考察词类配价习得情况的方法,能够直接地考察语言能力的发展和变化。一方面能够从习得角度进一步观察汉语动词和名词的差别,另一方面也丰富了语言系统复杂性和语言习得复杂性的研究。

第六章 结论

第一节 主要内容总结

基于依存语法和概率配价模式理论,本研究提出了一种从语法功能角度计算词(类)与词(类)间相似度的方法,并且计算了汉语中的高频动名兼类词与动词、名词两个词类范畴的相似度,还用借助相似度计量了以汉语为第二语言的学习者动词配价和名词配价的习得水平发展情况。主要成果有三项:动词、名词、动名兼类词的语法功能统计与区分,动名兼类词与动词和名词的相似度计算,汉语中介语中的动词配价、名词配价习得水平发展评估。接下来我们以这三项具体说明本研究成果。

首先,动词、名词、动名兼类词的语法功能统计与区分。基于表层句法通用依存标注数据集(SUD)中的汉语树库 GSDSimp,我们做了典型动词、典型名词、典型兼类词与非典型动词、非典型名词、非典型动名兼类词的区分,并将统计对象做了频率上的限制。随后以基于词语与基于依存关系、以类为对象与以词为对象两个维度上计量了动词、名词、动名兼类词的语法功能,并对语法功能做了主次要的区分。结果显示,在基于词和基于依存关系时,以类为对象和以词为对象得到的结果高度一致。这一方面说明了结果可信,另一方面也说明了基于词和基于依存关系是存在一定区别的。经统计,基于词时,动词主要支配名词、动词、介词,主要从属动词、助动词、副词、小品词;名词主要支配名词、小品词、专有名词、数词,主要从属动词、名词、小品词。基于依存关系时,动词主要支配宾语、主语、修饰语,主要做谓语、宾语、修饰语;名词主要支配修饰语,特别是定语,主要宾语、修饰语、主语。动名兼类词则兼具动词和名词的特点,动名兼类词的主要功能是支配名词、小品词、动词、介词和从属动词、小品词、名词、助动词,以及支配修饰语、宾语、主语、补语和做补语、修饰语。

然后,我们依托动词、名词、动名兼类词的概率配价模式,借助余弦相似度计算出每个动名兼类词与动词、名词的相似度,并对比了基于词和基于依存关系所得结果。结果表明,基于依存关系更有利于展现词与词类间的关系。动名兼类词的动词相似度和名词相似度呈弱负相关,这代表一个动名兼类词的越接近动词,就越远离名词,证实了动词和名词是个连续统。这一研究结果表明动名兼类词是个研究动名关系的好窗口。

最后,我们利用相似度指标计量了汉语作为第二语言的学习者习得动词配价和名词配价的具体情况。通过计算目的语和中介语中的动词、名词的相似度,我们得到了七个等级的学习者配价习得情况。第一,动词与名词的整体配价习得水平没有显著差异。做支配词时,动词配价习得水平低于名词配价;做从属词时,动词配价习得水平高于名词配价。动词和名词配价内部的习得水平也存在差异。动词、名词的整体配价习得水平都是最低的。动词做从属词时配价习得水平高于做从属词时的习得水平;而名词做从属词时配价习得水平则低于做支配词时的习得水平。这来自于价和依存距离两方面因素的制约。动词平均价更大且价分布服从柯西分布, MDD_1 大, MDD_2 小,这使得动词做支配词时习得难度较大,做从属词时习得难度小;而名词平均价较小,价分布服从伽马分布, MDD_1 小, MDD_2 大,使得名词做支配词时习得难度小,做从属词时习得难度较大。第二,动词、名词配价的习得也不是个线性的过程,整体配价、做支配词时配价、做从属词时配价都有回退、化石化、跳跃式进步等现象。母语迁移、学习策略、教学方案等因素的动态博弈造成了动词配价和名词配价的这种非线性、动态特征。第三,名词配价习得在60分前后变化最显著,动词在80分前后变化最显著。动词和名词不同的习得难度造成了这一点,动词可参与的依存关系更为丰富,学习者需要较长的时间学习、纠错,才能良好习得;而名词难度较低,学习者依据母语知识就能产出相关结构,当学习深入,母语知识就不足以支持产出,但经过适时的纠错和练习,习得水平又能够提高。第四,动词配价和名词配价习得并不是从零开始的,且习得始终不能达到母语者水平。学习者母语的负迁移能帮助学习者在学习初期就产出一些结构正确的句子;但母语的负迁移也会导致中介语始终达不到目的语水平。

总的来说,我们提出了一种切实可行的计量词与词、词与词类、词类与词类间关系的指标,并且用这个指标测量了学习者的习得情况。对词类理论研究来说,这一方法有助于明晰词类之间的关系,对构建更合理的词类体系有一定帮助。此外对于语言习得研究来说,计算相似度也是一种很好的考察学习者习得进度的好方法,在第二语言习得中可以发挥出作用。

第二节 不足与展望

我们的研究也存在一些明显的问题。

首先,我们使用的主要资源 GSDSimp 树库标注是存在一些问题的,这些问

题在施佳璐等(2022)提到。虽然在大规模统计下,误差的影响会被尽可能缩小,从结果来看,这些误差不足以影响我们的观点。但不可否认,它仍然存在。我们选择 GSDSimp 的原因是因为它是 SUD 汉语树库中规模最大的。另外,在计量动词配价和名词配价的习得水平时,我们选择的中介语树库标注体系并不是 SUD,因而进行了标注体系转换。转换过程中,也会一定程度上使数据失真。虽然从结果看,数据的解释性较强,但我们同样不能忽略这一点。在这些问题上,更大规模和同标注体系的资源仍然值得期待。

其次,在比较基于词和基于依存关系的相似度结果时,我们认为依存关系比词类更有利于计量相似度。诚然,如我们文章中说的那样,依存关系更能体现词类在句法上的特征。但有一点也很重要——特征的数量。词类特征有 24 ($12*2$) 个,而依存关系有 30 ($15*2$) 个。更多种类的特征也有利于结果的精确。我们未来也会想办法排除特征数量对结果的影响,以检验依存关系是否真的更有利于描述词类。

再次,词在句法层面的功能不仅有其结合能力和可充当的句法成分,词的语序限制也是很重要的内容,但显然我们没有把语序考虑进来,这也是我们未来的进步空间之一。

最后,就机器学习常用的特征维度而言,我们的特征维度并不算多。这或许也会造成一定程度上的结果模糊。从概率配价模式理论来看,我们还可以把词类和依存关系组合起来,把特征维度丰富到 576 ($12*24*2$) 个。这种方法下得出来的特征或许会高度稀疏,但经过适当操作,或许也可以让我们的结果更加精确。这也是未来我们会尝试的内容。

参考文献

- [1]安华林.从两种词表看名、动、形兼类的处理[J].语言教学与研究,2005(04):31-39.
- [2]蔡淑美,施春宏.基于汉语中介语语料库的二价名词习得研究[J].语言文字应用,2014(02):85-95.
- [3]陈爱文.汉语词类研究和分类实验[M].北京:北京大学出版社,1986.
- [4]陈承泽.国文法草创[M].北京:商务印书馆,1922.
- [5]陈芯莹,刘海涛.句法复杂网络作为语体分类的知识源研究[J].计算机工程与应用,2013,49(08):32-36.
- [6]陈望道.文法简论[M].上海:上海教育出版社,1978.
- [7]程工.名物化与向心结构理论新探[J].现代外语,1999(02):131-144.
- [8]程娟.《现代汉语词典》词语功能义项计量考察[J].语文研究,2004(01):45-51.
- [9]戴耀晶.现代汉语时体系统研究[M].杭州:浙江教育出版社,1997.
- [10]翟淑英.谈现代汉语中词的兼类问题[J].内蒙古大学学报(人文社会科学版),2000(S1):90-92.
- [11]方光焘.体系与方法[J].语文周刊,1939(28).
- [12]方清明.《现代汉语词典》“动、名兼类”计量考察[J].辞书研究,2010(04):30-40.
- [13]方绪军.近义词的配价差异及其对习得的影响[J].海外华文教育,2003(04):8-13+23.
- [14]冯丽萍,盛双霞.外国学生中文三价动词的习得规律研究[J].云南师范大学学报,2004(03):30-34.
- [15]冯志伟.汉字和汉语的计算机处理[J].当代语言学,2001(01):1-21+77.
- [16]付炜.汉语词汇教学的两个难点——兼类词、介词[J].新疆大学学报(哲学社会科学版),1994(04):109-112.
- [17]高芳,徐盛桓.名动转用语用推理的认知策略[J].外语与外语教学,2000(04):13-16+20.
- [18]高芳,徐盛桓.名动转用与语用推理[J].外国语(上海外国语大学学报),2000(02):7-14.
- [19]高航.概念物化与名词化[J].解放军外国语学院学报,2007(06):14-17+86.
- [20]高航.现代汉语名动互转的认知语法考察[D].中国人民解放军外国语学院,2007.

- [21]高航.动词化机制的认知语法考察[J].解放军外国语学院学报,2008(05):20-25.
- [22]高航.认知语法视角下的汉语兼类问题考察[J].汉语学习,2009(02):17-24.
- [23]高航.名词化的概念组织层面:从认知语法的视角[J].解放军外国语学院学报,2009,32(03):7-12.
- [24]高名凯.汉语语法论[M].北京:商务印书馆,1986.
- [25]郭锐.语文词典的词性标注问题[J].中国语文,1999(02):150-158.
- [26]郭锐.现代汉语词类研究[M].北京:商务印书馆,2002.
- [27]高松,颜伟,刘海涛.基于树库的现代汉语动词句法功能的计量研究[J].汉语学习,2010(05):105-112.
- [28]高松.基于依存树库的现代汉语名词语法功能的计量研究[J].华文教学与研究,2010(02):54-60.
- [29]郝红艳.“读后续写”练习对汉语名量词促学促用效果研究[J].广东外语外贸大学学报,2015,26(06):90-94.
- [30]郝瑜鑫,王雪琳,刘海涛.基于句法标注语料库的汉语中介语动词配价发展计量研究[J].语言文字应用,2021(01):29-41.
- [31]郝瑜鑫,王燕子,刘海涛.基于依存树库的汉语中介语名词配价发展研究[J].云南师范大学学报(对外汉语教学与研究版),2022,20(04):84-92.
- [32]何元建,王玲玲.论汉语中的名物化结构[J].汉语学习,2007(01):13-24.
- [33]胡明扬.动名兼类的计量考察[J].语言研究,1995(02):91-99.
- [34]胡明扬.现代汉语词类问题考察[J].中国语文,1995(05):381-389.
- [35]胡明扬.词类问题考察[M].北京:北京语言学院出版社,1996.
- [36]胡明扬.现代汉语的词类问题[C]//第六届国际汉语教学讨论会论文选.,1999:451-460.
- [37]胡明扬.汉语词类兼类研究[J].语言文字应用,2000(01):19-24.
- [38]胡裕树,范晓.动词形容词的“名物化”和“名词化”[M].中国语文,1994(2):81-85.
- [39]胡裕树.现代汉语(重订版)[M].上海:上海教育出版社,1995.
- [40]黄伯荣,廖序东.现代汉语(增订三版)[M].北京:高等教育出版社,2002.
- [41]洪铭材,张阔,唐杰,李涓子.基于条件随机场(CRFs)的中文词性标注方法[J].计算机科学,2006(10):148-151+155.
- [42]姜自霞,丁崇明,侯燕.《现代汉语词典》(第5版)中的双音节动名兼类词[J].辞书研究,2011(03):37-50+71.
- [43]金立鑫.普通话“NP的VP”的句法语义实质[J].语言教学与研究,2019(04):80-93.

- [44]金立鑫.汉语动词的典型性等级[J].语言教学与研究,2022(02):41-52.
- [45]金立鑫.“名包动”理论的逻辑问题[J].外国语(上海外国语大学学报),2022,45(01):2-13.
- [46]金兆梓.国文法之研究[M].北京:商务印书馆,1983.
- [47]黎锦熙.新著国语文法[M].北京:商务印书馆,2000.
- [48]李立成.自指的“的”字短语[J].语言教学与研究,1999(03):73-84.
- [49]李翔.《汉英词典》(第3版)词类标注评析[J].辞书研究,2013(05):53-59. [1]李宇明.所谓的“名物化”现象新解[J].华中师范大学学报(哲学社会科学版),1986(03):117-120+73.
- [50]李锡江,刘永兵.中国学生英语时空状语语序分布与母语概念迁移[J].外语教学与研究,2020,52(01):90-102+159.
- [51]李雯雯,王文斌.中国英语学习者时间状语从句语序习得发展的思维模式迁移[J].外语电化教学,2022(05):66-72+114.
- [52]李媛,段庭辉,刘海涛.名词分布是人类语言的不变量吗?——以德语书面语中名词分布为例[J].浙江大学学报(人文社会科学版),2019,49(06):39-48.
- [53]李月华.有关词的兼类问题[J].新疆大学学报(哲学社会科学版),1991(03):112-116+95.
- [54]寮菲.第二语言习得中母语迁移现象分析[J].外语教学与研究,1998(02):60-66+82.
- [55]蔺璜.定语位置上名词的句法表现及其语义特征[J].山西大学学报(哲学社会科学版),2005(02):95-99.
- [56]刘海涛,冯志伟.自然语言处理的概率配价模式理论[J].语言科学,2007,(6).
- [57]刘海涛.依存语法的理论与实践[M].北京:科学出版社,2009.
- [58]刘海涛,林燕妮.大数据时代语言研究的方法和趋向[J].新疆师范大学学报(哲学社会科学版),2018,39(01):72-83.
- [59]刘海涛,郑国锋.大数据时代语言学理论研究的路径与意义[J].当代外语研究,2021(02):5-18+31+2. [1]陆俭明.关于词的兼类问题[J].中国语文,1994(01):28-34.
- [60]刘海涛.依存关系与语言网络[M].北京:科学出版社,2022.
- [61]刘顺.现代汉语动词的零形式名词化[J].山东师范大学学报(人文社会科学版),2003(01):32-37.
- [62]刘颂浩.论“把”字句运用中的回避现象及“把”字句的难点[J].语言教学与研究

- 究,2003(02):64-71.
- [63]刘涛,杨亦鸣,张辉,张珊珊,梁丹丹,顾介鑫,胡伟.语法语境下汉语名动分离的ERP研究[J].心理学报,2008(06):671-680.
- [64]刘正光.宾语隐形时的语义与语用特征[J].外语教学与研究,2007(01):43-50+81.
- [65]陆俭明.再论汉语词类问题——从沈家煊先生的“名动包含”观说起[J].东北师大学报(哲学社会科学版),2022(04):1-15.
- [66]陆俭明.再议“汉语名动包含说”[J].外国语,2022,45(5):2-14.
- [67]卢福波.汉语语法教学理论与方法[M].北京:北京大学出版社,2010.
- [68]卢英顺.词类划分:(广义)形态和原型理论的嫁接[A],范晓、李熙宗、戴耀晶等.语言研究的新思路[M],上海:上海教育出版社,1998.
- [69]骆琳,穆柔云.从物性结构看工具类名词的动词化现象[J].语言文字应用,2021(03):80-90.
- [70]吕叔湘.中国语法要略[M].北京:商务印书馆,1947.
- [71]吕叔湘.关于词类的一些原则性问题[J].中国语文,1954(09/10).
- [72]吕叔湘.汉语语法论文集[M].北京:商务印书馆,1984.
- [73]吕叔湘.汉语句法的灵活性[J].中国语文,1986,(1).
- [74]吕叔湘.语法研究入门[M].北京:商务印书馆,1999.
- [75]马建忠.马氏文通[M].北京:商务印书馆,2008.
- [76]马宏基.对外汉语教材中兼类词的标注问题[J].科学决策,2008(10):86-88.
- [77]莫彭龄,单青.三大类实词句法功能的统计分析[J].南京师大学报(社会科学版),1985(03):55-63.
- [78]聂志平,杨文全.从同一性理论看汉语词的兼类问题[J].云南师范大学学报(对外汉语教学与研究版),2016,14(06):64-74.
- [79]彭利贞.现代汉语情态研究[M].北京:中国社会科学出版社,2007.
- [80]朴重奎.单个动词作主语的语义语法考察[J].汉语学习,2003(06):25-31.
- [81]邵斌,杨静.英汉名动范畴边界渗透的类型学考察[J].外国语(上海外国语大学学报),2022,45(03):2-10.
- [82]沈家煊.汉语里的名词和动词[J].汉藏语学报,2007(1):27-47.
- [83]沈家煊.我看汉语的词类[J].语言科学,2009,8(01):1-12.
- [84]沈家煊.我只是接着向前跨了半步--再谈汉语里的名词和动词[J].语言学论丛,2009(40):3-22.
- [85]沈家煊.英汉否定词的分合和名动的分合[J].中国语文,2010(05):387-399+479.

- [86] 沈家煊, 乐耀. 词类的实验研究呼唤语法理论的更新[J]. 当代语言学, 2013, 15(03): 253-267+376.
- [87] 沈家煊. 名词和动词[M]. 北京: 商务印书馆, 2016.
- [88] 沈家煊. 名词的定义问题[J/OL]. 现代外语: 1-13[2023-03-03].
- [89] 沈阳. 动词的句位和句位变体结构中的空语类[J]. 中国语文, 1994(02): 139-148.
- [90] 石锋. 演化语言学的宏观史、中观史和微观史[J]. 南开学报(哲学社会科学版), 2018(04): 65-71.
- [91] 师佳璐, 罗昕宇, 杨麟儿, 肖丹, 胡正升, 王一君, 袁佳欣, 余婧思, 杨尔弘. 汉语学习者依存句法树库构建[J]. 中文信息学报, 2022, 36(01): 39-46.
- [92] 史有为. 多元·柔性·主体——80~90年代语法研究大势之我见[J]. 世界汉语教学, 1991(04): 204-210.
- [93] 史有为. 呼唤柔性 走向柔性[J]. 汉语学习, 1993(01): 23-24.
- [94] 宋柔. 从语言工程看汉语词类[J]. 语言学论丛, 2003(50).
- [95] 宋玉柱. 关于体宾动词和谓宾动词[J]. 世界汉语教学, 1991(02): 90-91.
- [96] 宋作艳. 逻辑转喻、事件强迫与名词动用[J]. 语言科学, 2013, 12(02): 117-129.
- [97] 宋作艳. 名词转动词的语义基础——从动词视角到名词视角[J]. 中国语文, 2018(03): 295-310+383.
- [98] 宋作艳. 名词转动词的语义基础——从动词视角到名词视角[J]. 中国语文, 2018(03): 295-310+383.
- [99] 孙崇飞. 名动包含理论存在逻辑问题吗?[J]. 外国语(上海外国语大学学报), 2022, 45(05): 15-23.
- [100] 孙德金. 现代汉语名词做状语的考察[J]. 语言教学与研究, 1995(04): 88-98.
- [101] 汪涛. 二语习得中的错误分析研究现状及其发展趋势[J]. 河北理工大学学报(社会科学版), 2011, 11(04): 101-103.
- [102] 王保昌, 洪霞, 李荣宝. 固化、占位和动词语义对英语使役构式过度泛化的抑制作用[J]. 外语教学与研究, 2022, 54(03): 375-387+479.
- [103] 王冬梅. 现代汉语动名互转的认知研究[D]. 中国社会科学院研究生院, 2001.
- [104] 王冬梅. 动词的控制度和谓宾的名物化之间的共变关系[J]. 中国语文, 2003(04): 315-328+384.
- [105] 王力. 中国现代语法[M]. 北京: 商务印书馆, 1943.
- [106] 王璐. 基于语料库的汉语兼类词计量研究[A], 黄伟. 词汇与句法计量研究[M]. 杭州: 浙江大学出版社, 2022.

- [107]王仁强. 认知视角的汉英词典词类标注实证研究[D].广东外语外贸大学,2006.
- [108]王仁强.语法隐喻与汉语词典自指义项的设立——一项基于语料库的研究[J].外国语文,2009,25(01):100-108.
- [109]王仁强.现代汉语词类体系效度研究——以《现代汉语词典》(第5版)词类体系为例[J].外语教学与研究,2010,42(05):380-386+401.
- [110]王仁强,周瑜.现代汉语兼类与词频的相关性研究——兼评“简略原则”的效度[J].外国语文,2015,31(02):61-69.
- [111]王仁强,黄昌宁.从双层词类范畴化理论看现代汉语自指词项的兼类问题[J].外国语文,2017,33(01):87-96.
- [112]王勇,徐杰.现代汉语诗歌中的名词动用[J].华中师范大学学报(人文社会科学版),2018,57(05):106-114.
- [113]温锁林,刘开瑛.汉语名、动、形兼类词的两种鉴别方法[J].语文研究,1998(01):31-35.
- [114]吴静,石毓智.英汉主宾语性质的差别及其对话题化的影响[J].四川外语学院学报,2005(05):74-79.
- [115]吴铭.汉语是否存在名动包含关系?[J].解放军外国语学院学报,2021,44(03):60-67.
- [116]吴蔚天,罗建林.汉语计算语言学——汉语形式语法和形式分析[M].北京:电子工业出版社,1994.
- [117]吴义诚,戴颖.有标记语言现象与语法特性研究[J].现代外语,2022,45(03):306-317.
- [118]夏全胜.汉语名词、动词和动名兼类词语义加工的ERP研究[D].南开大学,2012. [1]夏全胜,彭刚,石锋.汉语名词、动词和动名兼类词语义加工的偏侧化现象——来自ERP的研究[J].心理科学,2014,37(06):1333-1340.
- [119]项梦冰.论“这本书的出版”中“出版”的词性——对汉语动词、形容词“名物化”问题的再认识[J].天津师大学报(社会科学版),1991(04):75-80.
- [120]肖奚强,周文华.外国学生汉语趋向补语句习得研究[J].汉语学习,2009(01):70-81.
- [121]邢福义.现代汉语[M].北京:高等教育出版社,1991.
- [122]邢福义.词类辨难[M].北京:商务印书馆,2003.
- [123]许红花.关于《HSK大纲》兼类词的范围问题[J].汉语学习,2009(04):94-99.

- [124]徐艳华. 现代汉语实词语法功能考察及词类体系重构[D]. 南京师范大学, 2006.
- [125]徐艳华, 朱诗瑶. 基于统计的《HSK 词汇等级大纲》动名兼类词标注研究[J]. 云南师范大学学报(对外汉语教学与研究版), 2020, 18(04): 51-57.
- [126]阎建玮. 基于大规模树库的跨语言语序类型计量研究[D]. 浙江大学, 2022.
- [127]杨成凯. 动词作主宾语是汉语的语法特点吗?——汉语语法特点散论之一[J]. 汉语学习, 1991(06): 9-12.
- [128]杨德峰. 英语母语学习者趋向补语的习得顺序——基于汉语中介语语料库的研究[J]. 世界汉语教学, 2003(02): 52-65+2-3.
- [129]杨丽姣. 词类区分词义计量分析[J]. 云南师范大学学报(哲学社会科学版), 2010, 42(01): 49-54.
- [130]杨同用. 从语法搭配看动名兼类的复杂性——兼论语文词典的词性标注[J]. 辞书研究, 2008(02): 27-36.
- [131]杨同用, 赵金广. 动词、形容词与名词兼类的复杂性[J]. 河北师范大学学报(哲学社会科学版), 2010, 33(03): 111-114.
- [132]杨亦鸣, 梁丹丹, 顾介鑫, 翁旭初, 封世文. 名动分类: 语法的还是语义的——汉语名动分类的神经语言学研究[J]. 语言科学, 2002(01): 31-46.
- [133]杨圳, 施春宏. 汉语准价动词的二语习得表现及其内在机制[J]. 世界汉语教学, 2013, 27(04): 558-573.
- [134]尹洪波. 《现代汉语词典》词类标注变化的统计与评析[J]. 中国社会科学院研究生院学报, 2016(03): 108-114.
- [135]俞士汶, 朱学锋, 刘扬. 参照生成词库理论对动名兼类现象再讨论[J]. 辞书研究, 2020(04): 1-8+125.
- [136]袁毓林. 词类范畴的家族相似性[J]. 中国社会科学, 1995(01): 154-170.
- [137]袁毓林. 一个汉语词类的准公理系统[J]. 语言研究, 2000(04): 1-28.
- [138]袁毓林. 基于隶属度的汉语词类的模糊划分[J]. 中国社会科学, 2005(01): 164-177+209.
- [139]袁毓林. 现代汉语虚词模糊划分的隶属度量表[J]. 汉语学报, 2005(04): 12-21+95.
- [140]岳中奇. 句首前置定语和状语的位移性质思辨[J]. 汉语学习, 2007(06): 12-19.
- [141]詹卫东. 计算机句法结构分析需要什么样的词类知识——兼评近年来汉语词类研究的新进展[J]. 中国语文, 2013(02): 178-190+192.

- [142]张爱朴.汉语动词是名词的一个次类吗?[J].汉语学习,2012(03):43-49.
- [143]张宝林.回避与泛化——基于“HSK 动态作文语料库”的“把”字句习得考察[J].世界汉语教学,2010,24(02):263-278.
- [144]张斌.新编现代汉语[M],上海:上海外语教育出版社,2002.
- [145]张伯江.词类活用的功能解释.中国语文,1994(5):339-346.
- [146]张伯江.双音化的名词性效应[J].中国语文,2012(04):338-346+384.
- [147]张晓颖.《新世纪汉英大词典》词类标注法评析[J].外语教学,2005(03):94-96.
- [148]张云徽.汉语“动、介”组合及其他[J].云南民族大学学报(哲学社会科学版),2005(02):150-152.
- [149]赵金铭,张博,程娟.关于修订《(汉语水平)词汇等级大纲》的若干意见[J].世界汉语教学,2003(03):73-83+4.
- [150]赵金铭.对外汉语教学法回视与再认识[J].世界汉语教学,2010,24(02):243-254.
- [151]赵雪莹.浅论汉语词汇的兼容性及其对对外汉语教学的影响[J].思想战线,2013,39(S2):351-352.
- [152]赵元任.北京口语语法[M].北京:中国青年出版社,1955.
- [153]赵元任.汉语口语语法[M].北京:商务印书馆,1979.
- [154]赵元任.中国话的文法[M].香港:香港中文大学出版社,2002.
- [155]钟棫.汉语词典标注词性问题[J].辞书研究,1980(01):68-93.
- [156]周领顺.英汉名—动转类词对比研究[J].外语教学与研究,2000(05):340-344+399.
- [157]周统权,杨静.风行“山寨”的认知发展[J].外语研究,2010(02):14-19.
- [158]朱德熙,卢甲文,马真.关于动词形容词“名物化”的问题[J].北京大学学报(人文科学),1961(04):53-66.
- [159]朱德熙.语法讲义[M].北京:商务印书馆,1982.
- [160]朱学锋,俞士汶,王惠.现代汉语五万词语归类的实践[J].语言文字应用,1997(04).
- [161]朱永生.名词化、动词化与语法隐喻[J].外语教学与研究,2006(02):83-90+160.
- [162]朱志平.汉语二语教学中词汇计量的维度[J].语言文字应用,2013(02):90-98.
- [163] Altmann G. A Law of Change in Language, Historical Linguistics[J]. Quantitative Linguistics, 1983, 18: 104-115.
- [164] Chen X, Liu H, Gerdes K. Classifying syntactic categories in the Chinese dependency network[C]//Proceedings of the Third International Conference on

- Dependency Linguistics (Depling 2015). 2015: 74-81.
- [165] Cong J, Liu H. Approaching human language with complex networks[J]. *Physics of life reviews*, 2014, 11(4): 598-618.
- [166] Corominas-Murtra B, Valverde S, Solé R. The ontogeny of scale-free syntax networks: phase transitions in early language acquisition[J]. *Advances in Complex Systems*, 2009, 12(03): 371-392.
- [167] Crystal D. 'English', in word classes[J]. *Lingua*, 1967, (17): 24-56.
- [168] Ellis R. *Second language acquisition*[J]. The United States: Oxford, 1997: 98.
- [169] Ferrer-i-Cancho R. A stronger null hypothesis for crossing dependencies[J]. *Europhysics Letters*, 2014, 108(5): 58003.
- [170] Gerdes K, Guillaume B, Kahane S, et al. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD[C]//*Proceedings of the second workshop on Universal Dependencies (UDW 2018)*. 2018: 66-74.
- [171] Halliday, M.A.K., *An introduction to functional grammar*[M]. London: Edward Arnold.
- [172] Hudson R. *An introduction to word grammar*[M]. Cambridge University Press, 2010.
- [173] Hunter J D. Matplotlib: A 2D graphics environment[J]. *Computing in science & engineering*, 2007, 9(03): 90-95.
- [174] Jiang J, Liu H. The effects of sentence length on dependency distance, dependency direction and the implications—based on a parallel English–Chinese dependency treebank[J]. *Language Sciences*, 2015, 50: 93-104.
- [175] Jiang J, Yu W, Liu H. Does scale-free syntactic network emerge in second language learning?[J]. *Frontiers in Psychology*, 2019, 10: 925.
- [176] Labov W. The social stratification of (r) in New York City department stores[M]//*Dialect and language variation*. Academic Press, 1986: 304-329.
- [177] Langacker R W. *Foundations of cognitive grammar: Volume I: Theoretical prerequisites*[M]. Stanford university press, 1987.
- [178] Langacker R W. *Cognitive Grammar: A Basic Introduction*[M]. Oxford University Press, 2008.
- [179] Larsen-Freeman D. Chaos/complexity science and second language acquisition[J]. *Applied linguistics*, 1997, 18(2): 141-165.

- [180] Larsen-Freeman D, Cameron L. Complex systems and applied linguistics[M]. Oxford: Oxford University Press, 2008.
- [181] Li C N, Thompson S A. Subject and topic: A new typology of language. In Charles N. Li(ed.) Subject and Topic[M]. Austin: University of Texas Press, 1976.
- [182] Liu H. The complexity of Chinese syntactic dependency networks[J]. Physica A: Statistical Mechanics and its Applications, 2008, 387(12): 3048-3058.
- [183] Liu H. Dependency distance as a metric of language comprehension difficulty[J]. Journal of Cognitive Science, 2008, 9(2): 159-191.
- [184] Liu H. Probability distribution of dependencies based on a Chinese dependency treebank[J]. Journal of Quantitative Linguistics, 2009, 16(3): 256-273.
- [185] Liu H, Li W. Language clusters based on linguistic complex networks[J]. Chinese Science Bulletin, 2010, 55: 3458-3465.
- [186] Liu H. Linguistic Complex Networks: a new approach to language exploration[J]. Grundlagenstudien Aus Kybernetik Geisteswissenschaft (grkg/Humankybernetik), 2011, 52: 151-170.
- [187] Liu H, Cong J. Language clustering with word co-occurrence networks based on parallel texts[J]. Chinese Science Bulletin, 2013, 58: 1139-1144.
- [188] Liu H, Xu C, Liang J. Dependency distance: A new perspective on syntactic patterns in natural languages[J]. Physics of life reviews, 2017, 21: 171-193.
- [189] Lu Q, Xu C, Liu H. The influence of chunking on dependency crossing and distance[J]. arXiv preprint arXiv:1509.01310, 2015.
- [190] Mel'cuk I A. Dependency syntax: theory and practice[M]. New York: SUNY press, 1988.
- [191] Osborne T, Gerdes K. The status of function words in dependency grammar: A critique of Universal Dependencies (UD)[J]. Glossa: a journal of general linguistics (2016-2021), 2019.
- [192] Robinson J J, Dependency Structures and Transformational Rules[J]. Language, 1970, (2).
- [193] Ross J R. Nouniness[C]/O. Fujimura. Three Dimensions of Linguistic Theory. Tokyo: TEC, 1973.137-257.
- [194] Selinker L. Interlanguage[J]. IRAL.1972,10(3): 209-231.
- [195] Solé R. Syntax for free?[J]. Nature, 2005, 434(7031): 289-289.

- [196] Swain M, Lapkin S. Task-based second language learning: The uses of the first language[J]. *Language teaching research*, 2000, 4(3): 251-274.
- [197] Temperley D. Dependency-length minimization in natural and artificial languages[J]. *Journal of Quantitative Linguistics*, 2008, 15(3): 256-282.
- [198] Tesnière L. *Elements de la Syntaxe Structurale*[M]. Paris: Klincksieck, 1959.
- [199] Ungerer F, Schmid H J. *An introduction to cognitive linguistics*[M]. Routledge, 2013.
- [200] Van Rossum G, Drake F L. *Python reference manual*[M]. Amsterdam: Centrum voor Wiskunde en Informatica, 1995.
- [201] Virtanen P, Gommers R, Oliphant T E, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python[J]. *Nature methods*, 2020, 17(3): 261-272.
- [202] Wang L. Part-of-speech studies in Chinese[J]. *Journal of Quantitative Linguistics*, 2016, 23(3): 235-255.
- [203] Wang W S Y. Competing changes as a cause of residue[J]. *Language*, 1969: 9-25.
- [204] Waskom M L. Seaborn: statistical data visualization[J]. *Journal of Open Source Software*, 2021, 6(60): 3021.
- [205] White L. Markedness and second language acquisition: The question of transfer[J]. *Studies in second language acquisition*, 1987, 9(3): 261-285.
- [206] Xia Q, Peng G. The roles of object and action, and concreteness and imageability, in the distinction between nouns and verbs: An ERP study on monosyllabic words in Chinese[J]. *Journal of Neurolinguistics*, 2022, 61: 101026.
- [207] Xu C, Liu H. The role of working memory in shaping syntactic dependency structures. In John Schwieter & Edward Wen. (eds.) *The Cambridge Handbook of Working Memory and Language*[M]. Cambridge University Press. p. 343-367, 2022.
- [208] Yan J, Liu H. Which annotation scheme is more expedient to measure syntactic difficulty and cognitive demand?[C]//*Proceedings of the first workshop on quantitative syntax (Quasy, SyntaxFest 2019)*. 2019: 16-24.
- [209] Yan J, Liu H. Quantitative Analysis of Chinese and English Verb Valencies Based on Probabilistic Valency Pattern Theory[C]//*Chinese Lexical Semantics: 22nd Workshop, CLSW 2021, Nanjing, China, May 15–16, 2021, Revised Selected Papers, Part II*. Cham: Springer International Publishing, 2022: 152-162.

- [210] Yang M, Liu H. The role of syntax in the formation of scale-free language networks[J]. Europhysics Letters, 2022, 139(6): 61002.
- [211] Zipf, G. K. Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Cambridge, MA: Addison-Wesley Press, 1949.

附录 A 统计信息附表

在第三章和第四章中，因表 3.1-3.4 和表 4.1-4.4 占空间较大，为提高正文的可读性，我们把它们在此展示。表名前标记了表在正文中出现的页码。

p.21: 表 3.1 动名兼类词与词类的结合情况描述性统计⁷

词类	频次	平均值	标准差	最小值	25%	50%	75%	最大值
AUX	286.0	0.017	0.051	0.0	0.000	0.000	0.000	0.500
	286.0	0.119	0.137	0.0	0.000	0.089	0.167	0.714
ADJ	286.0	0.023	0.048	0.0	0.000	0.000	0.025	0.294
	286.0	0.008	0.030	0.0	0.000	0.000	0.000	0.222
ADV	286.0	0.058	0.110	0.0	0.000	0.000	0.077	1.000
	286.0	0.079	0.103	0.0	0.000	0.037	0.125	0.500
NUM	286.0	0.010	0.036	0.0	0.000	0.000	0.000	0.250
	286.0	0.000	0.005	0.0	0.000	0.000	0.000	0.077
VERB	286.0	0.108	0.122	0.0	0.000	0.077	0.176	0.786
	286.0	0.359	0.205	0.0	0.228	0.333	0.500	1.000
NOUN	286.0	0.393	0.196	0.0	0.252	0.385	0.500	1.000
	286.0	0.145	0.155	0.0	0.000	0.111	0.222	0.857
PRON	286.0	0.016	0.037	0.0	0.000	0.000	0.010	0.286
	286.0	0.000	0.001	0.0	0.000	0.000	0.000	0.019
PART	286.0	0.180	0.165	0.0	0.059	0.152	0.250	1.000
	286.0	0.205	0.189	0.0	0.065	0.167	0.294	1.000
ADP	286.0	0.103	0.124	0.0	0.000	0.071	0.150	0.632
	286.0	0.084	0.107	0.0	0.000	0.050	0.125	0.667
DET	286.0	0.010	0.038	0.0	0.000	0.000	0.000	0.333
	286.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
CCONJ	286.0	0.019	0.043	0.0	0.000	0.000	0.000	0.250
	286.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
PROP	286.0	0.054	0.092	0.0	0.000	0.000	0.073	0.600
	286.0	0.001	0.016	0.0	0.000	0.000	0.000	0.250

p.21: 表 3.2 词类结合情况 Wilcoxon 检验两两对比 p 值表

	AUX	ADJ	ADV	NUM	VERB	NOUN	PRON	PART	ADP	DET	CCONJ
ADJ	.037										
		.000									
ADV	.000	.000									

⁷ 每个词类都有两行数据，第一行代表做支配词时的统计性描述，第二行则是做从属词时的统计性描述。以下相关类似表格如未做说明都循此例。

附录 A 统计信息附表

	.000	.000								
NUM	.019	.000	.000							
	.000	.000	.000							
VERB	.000	.000	.000	.000						
	.000	.000	.000	.000						
NOUN	.000	.000	.000	.000	.000					
	.039	.000	.000	.000	.000					
PRON	.771	.061	.000	.005	.000	.000				
	.000	.000	.000	.655	.000	.000				
PART	.000	.000	.000	.000	.000	.000	.000			
	.000	.000	.000	.000	.000	.000	.000			
ADP	.000	.000	.000	.000	.462	.000	.000	.000		
	.001	.000	.534	.000	.000	.000	.000	.000		
DET	.048	.000	.000	.662	.000	.000	.008	.000	.000	
CCONJ	.232	.303	.000	.001	.000	.000	.378	.000	.000	.002
PROPN	.000	.000	.670	.000	.000	.000	.000	.000	.000	.000
	.000	.001	.000	.279	.000	.000	.080	.000	.000	

p.27: 表 3.3 动名兼类词支配和充当的句法成分情况描述性统计

依存关系	频次	平均值	标准差	最小值	25%	50%	75%	最大值
Subj	286.0	0.154	0.143	0.0	0.041	0.143	0.222	1.000
	286.0	0.068	0.098	0.0	0.000	0.000	0.111	0.667
Root	286.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	286.0	0.070	0.104	0.0	0.000	0.000	0.111	0.500
Comp	286.0	0.249	0.202	0.0	0.073	0.233	0.404	1.000
	286.0	0.521	0.227	0.0	0.375	0.537	0.692	1.000
Conj	286.0	0.029	0.056	0.0	0.000	0.000	0.032	0.333
	286.0	0.032	0.058	0.0	0.000	0.000	0.053	0.333
Mod	286.0	0.317	0.224	0.0	0.159	0.286	0.443	1.000
	286.0	0.199	0.184	0.0	0.071	0.167	0.286	1.000
Parataxis	286.0	0.031	0.063	0.0	0.000	0.000	0.038	0.400
	286.0	0.022	0.048	0.0	0.000	0.000	0.000	0.286
Dislocated	286.0	0.000	0.002	0.0	0.000	0.000	0.000	0.039
	286.0	0.000	0.006	0.0	0.000	0.000	0.000	0.100
Udep	286.0	0.138	0.132	0.0	0.051	0.111	0.196	0.600

附录 A 统计信息附表

	286.0	0.000	0.004	0.0	0.000	0.000	0.000	0.071
Appos	286.0	0.008	0.038	0.0	0.000	0.000	0.000	0.500
	286.0	0.004	0.023	0.0	0.000	0.000	0.000	0.200
Compound	286.0	0.001	0.006	0.0	0.000	0.000	0.000	0.077
	286.0	0.084	0.143	0.0	0.000	0.000	0.125	0.875
Clf	286.0	0.020	0.057	0.0	0.000	0.000	0.000	0.500
	286.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
Det	286.0	0.023	0.056	0.0	0.000	0.000	0.000	0.400
	286.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
Cc	286.0	0.019	0.043	0.0	0.000	0.000	0.000	0.250
	286.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
Discourse	286.0	0.000	0.003	0.0	0.000	0.000	0.000	0.040
	286.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
Orphan	286.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	286.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000

p.27: 表 3.4 句法成分 Wilcoxon 检验两两对比 p 值表

	Subj	Root	Comp	Conj	Mod	Parataxis	Dislocated	Udep	Appos	Compound	Clf	Det	Cc
Root													
		.862											
Comp		.000											
		.000	.000										
Conj		.000	.000										
		.000	.000	.000									
Mod		.000	.011	.000									
		.000	.000	.000	.000								
Parataxis		.000	.000	.064	.000								
		.000	.000	.015	.000								
Dislocated		.000	.000	.000	.000	.000							
		.000	.000	.000	.000	.000	.000						
Udep		.184	.000	.000	.000	.000	.000						
		.000	.000	.000	.000	.000	.655						
Appos		.000	.000	.000	.000	.000	.000	.000					
		.000	.000	.000	.000	.000	.009	.006					
Compound		.000	.000	.000	.000	.000	.225	.000	.000				
		.417	.524	.008	.000	.000	.000	.000	.000	.000			
Clf		.000	.000	.035	.000	.017	.000	.000	.000	.000			

附录 A 统计信息附表

Det	.000	.000	147	.000	106	.000	.000	.000	.000	293		
Cc	.000	.000	.020	.000	.021	.000	.000	.000	.000	.827	.662	
Discourse	.000	.000	.000	.000	.000	.593	.000	.000	.463	.000	.000	.000

p.31: 表 4.1 动词和名词的词类结合情况描述性统计⁸

词性	频次	平均值	标准差	最小值	25%	50%	75%	最大值
AUX	243.0	0.030	0.052	0.0	0.000	0.000	0.045	0.286
	765.0	0.006	0.045	0.0	0.000	0.000	0.000	1.000
	243.0	0.192	0.190	0.0	0.000	0.167	0.286	1.000
	765.0	0.092	0.119	0.0	0.000	0.059	0.143	0.917
ADJ	243.0	0.006	0.021	0.0	0.000	0.000	0.000	0.167
	765.0	0.043	0.087	0.0	0.000	0.000	0.058	1.000
	243.0	0.005	0.023	0.0	0.000	0.000	0.000	0.167
	765.0	0.016	0.051	0.0	0.000	0.000	0.000	0.500
ADV	243.0	0.075	0.070	0.0	0.000	0.067	0.118	0.300
	765.0	0.006	0.030	0.0	0.000	0.000	0.000	0.500
	243.0	0.160	0.163	0.0	0.000	0.143	0.240	1.000
	765.0	0.030	0.062	0.0	0.000	0.000	0.037	0.567
NUM	243.0	0.007	0.042	0.0	0.000	0.000	0.000	0.500
	765.0	0.048	0.160	0.0	0.000	0.000	0.000	1.000
	243.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	765.0	0.000	0.004	0.0	0.000	0.000	0.000	0.117
VERB	243.0	0.166	0.148	0.0	0.062	0.139	0.235	0.714
	765.0	0.016	0.048	0.0	0.000	0.000	0.000	0.500
	243.0	0.394	0.241	0.0	0.222	0.350	0.500	1.000
	765.0	0.330	0.222	0.0	0.154	0.333	0.500	0.950
NOUN	243.0	0.387	0.188	0.0	0.250	0.385	0.500	0.950
	765.0	0.348	0.256	0.0	0.143	0.356	0.500	1.000
	243.0	0.024	0.098	0.0	0.000	0.000	0.000	0.857
	765.0	0.233	0.200	0.0	0.091	0.167	0.333	1.000

⁸ 每个词类有四组数据，从上到下依次为动词做支配词、名词做支配词、动词做从属词、名词做从属词。

附录 A 统计信息附表

PRON	243.0	0.016	0.035	0.0	0.000	0.000	0.013	0.286
	765.0	0.014	0.069	0.0	0.000	0.000	0.000	1.000
PART	243.0	0.000	0.005	0.0	0.000	0.000	0.000	0.083
	765.0	0.001	0.010	0.0	0.000	0.000	0.000	0.167
ADP	243.0	0.094	0.096	0.0	0.000	0.071	0.133	0.571
	765.0	0.211	0.196	0.0	0.000	0.188	0.333	1.000
DET	243.0	0.155	0.189	0.0	0.000	0.105	0.243	1.000
	765.0	0.213	0.191	0.0	0.087	0.167	0.308	1.000
CCONJ	243.0	0.130	0.110	0.0	0.050	0.120	0.188	0.533
	765.0	0.030	0.117	0.0	0.000	0.000	0.000	1.000
ADP	243.0	0.067	0.120	0.0	0.000	0.000	0.100	0.714
	765.0	0.076	0.109	0.0	0.000	0.029	0.125	0.857
DET	243.0	0.002	0.009	0.0	0.000	0.000	0.000	0.083
	765.0	0.035	0.084	0.0	0.000	0.000	0.030	1.000
CCONJ	243.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	765.0	0.000	0.006	0.0	0.000	0.000	0.000	0.167
CCONJ	243.0	0.005	0.027	0.0	0.000	0.000	0.000	0.333
	765.0	0.038	0.088	0.0	0.000	0.000	0.040	1.000
PROPN	243.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	765.0	0.000	0.002	0.0	0.000	0.000	0.000	0.053
PROPN	243.0	0.074	0.125	0.0	0.000	0.033	0.091	1.000
	765.0	0.113	0.191	0.0	0.000	0.000	0.167	1.000
CCONJ	243.0	0.003	0.025	0.0	0.000	0.000	0.000	0.250
	765.0	0.009	0.040	0.0	0.000	0.000	0.000	0.438

p.31: 表 4.2 动词和名词的词类情况 Wilcoxon 检验两两对比 p 值表

	AUX	ADJ	ADV	NUM	VERB	NOUN	PRON	PART	ADP	DET	CCONJ
ADJ	.000										
	.000										
	.000										
	.000										
ADV	.000	.000									
	.259	.000									
	.067	.000									
	.000	.000									
NUM	.000	.179	.000								

附录 A 统计信息附表

	.000	.000	.000							
	.000	.000	.000							
VERB	.000	.000	.000	.000						
	.000	.000	.000	.001						
	.000	.000	.000							
	.000	.000	.000	.000						
NOUN	.000	.000	.000	.000	.000					
	.000	.000	.000	.000	.001					
	.000	.002	.000		.000					
	.000	.000	.000	.000	.000					
PRON	.001	.034	.000	.000	.000	.000				
	.000	.000	.003	.007	.016	.000				
	.000	.006	.000		.000	.000				
	.000	.000	.000	.221	.000	.000				
PART	.000	.000	.061	.000	.000	.000	.000			
	.000	.000	.000	.000	.000	.000	.000			
	.030	.000	.577		.000	.000	.000			
	.000	.000	.000	.000	.000	.000	.000			
ADP	.000	.000	.001	.000	.026	.000	.000	.000		
	.000	.000	.000	.059	.080	.000	.000	.000		
	.000	.000	.000		.000	.000	.000	.000		
	.007	.000	.000	.000	.000	.000	.000	.000		
DET	.000	.002	.000	.053	.000	.000	.000	.000	.000	
	.000	.013	.000	.151	.000	.000	.000	.000	.000	
	.000	.000	.000	.508	.929	.197	.600	.000	.000	
CCONJ	.000	.264	.000	.922	.000	.000	.000	.000	.000	.049
	.000	.083	.000	.020	.000	.000	.000	.000	.000	.560
	.000	.000	.000	.123	.123	.036	.465	.000	.000	.345
PROPN	.000	.000	.062	.000	.000	.000	.000	.002	.000	.000
	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	.000	.522	.000		.000	.000	.050	.000		.000
	.000	.002	.000	.000	.000	.000	.000	.000	.000	.000

p.34: 表 4.3 动词和名词支配和充当的依存关系情况描述性统计

依存关系	频次	平均值	标准差	最小值	25%	50%	75%	最大值
Subj	243.0	0.204	0.142	0.0	0.125	0.188	0.250	1.000
	765.0	0.004	0.033	0.0	0.000	0.000	0.000	0.500
	243.0	0.009	0.035	0.0	0.000	0.000	0.000	0.357

附录 A 统计信息附表

	765.0	0.200	0.187	0.0	0.000	0.167	0.316	1.000
Root	243.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	765.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	243.0	0.155	0.167	0.0	0.000	0.125	0.231	0.800
	765.0	0.002	0.022	0.0	0.000	0.000	0.000	0.511
Comp	243.0	0.429	0.219	0.0	0.333	0.455	0.567	1.000
	765.0	0.003	0.039	0.0	0.000	0.000	0.000	1.000
	243.0	0.567	0.242	0.0	0.417	0.571	0.726	1.000
	765.0	0.378	0.238	0.0	0.176	0.375	0.556	1.000
Conj	243.0	0.008	0.026	0.0	0.000	0.000	0.000	0.250
	765.0	0.075	0.158	0.0	0.000	0.000	0.091	1.000
	243.0	0.014	0.056	0.0	0.000	0.000	0.000	0.714
	765.0	0.050	0.078	0.0	0.000	0.000	0.087	0.500
Mod	243.0	0.194	0.139	0.0	0.102	0.179	0.273	0.750
	765.0	0.568	0.300	0.0	0.417	0.636	0.778	1.000
	243.0	0.177	0.196	0.0	0.000	0.125	0.258	0.944
	765.0	0.266	0.253	0.0	0.071	0.185	0.417	1.000
Parataxis	243.0	0.046	0.070	0.0	0.000	0.000	0.071	0.545
	765.0	0.003	0.029	0.0	0.000	0.000	0.000	0.667
	243.0	0.047	0.079	0.0	0.000	0.000	0.088	0.500
	765.0	0.003	0.022	0.0	0.000	0.000	0.000	0.333
Dislocated	243.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	765.0	0.004	0.054	0.0	0.000	0.000	0.000	1.000
	243.0	0.001	0.010	0.0	0.000	0.000	0.000	0.091
	765.0	0.000	0.003	0.0	0.000	0.000	0.000	0.083
Udep	243.0	0.106	0.113	0.0	0.000	0.074	0.152	0.667
	765.0	0.082	0.130	0.0	0.000	0.000	0.125	1.000
	243.0	0.000	0.000	0.0	0.000	0.000	0.000	0.003
	765.0	0.003	0.018	0.0	0.000	0.000	0.000	0.167
Appos	243.0	0.000	0.001	0.0	0.000	0.000	0.000	0.018
	765.0	0.033	0.093	0.0	0.000	0.000	0.000	1.000
	243.0	0.001	0.012	0.0	0.000	0.000	0.000	0.167
	765.0	0.016	0.062	0.0	0.000	0.000	0.000	1.000
Compound	243.0	0.000	0.003	0.0	0.000	0.000	0.000	0.048
	765.0	0.000	0.007	0.0	0.000	0.000	0.000	0.167
	243.0	0.029	0.110	0.0	0.000	0.000	0.000	0.833

附录 A 统计信息附表

	765.0	0.069	0.151	0.0	0.000	0.000	0.077	1.000
Clf	243.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	765.0	0.046	0.085	0.0	0.000	0.000	0.067	0.800
	243.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	765.0	0.012	0.085	0.0	0.000	0.000	0.000	1.000
Det	243.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	765.0	0.051	0.099	0.0	0.000	0.000	0.083	1.000
	243.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	765.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
Cc	243.0	0.005	0.027	0.0	0.000	0.000	0.000	0.333
	765.0	0.039	0.089	0.0	0.000	0.000	0.042	1.000
	243.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	765.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
Discourse	243.0	0.001	0.008	0.0	0.000	0.000	0.000	0.111
	765.0	0.000	0.001	0.0	0.000	0.000	0.000	0.014
	243.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	765.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
Orphan	243.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	765.0	0.000	0.003	0.0	0.000	0.000	0.000	0.083
	243.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000
	765.0	0.000	0.000	0.0	0.000	0.000	0.000	0.000

p.34: 表 4.4 句法成分 Wilcoxon 检验两两对比 p 值表

	Subj	Root	Comp	Conj	Mod	Parataxis	Dislocated	Udep	Appos	Compound	Clf	Det	Cc	Discourse
Root														
	184													
	.000													
	.000													
Comp	.000													
	.000													
	.000	.000												
	.000	.000												
Conj	.000		.000											
	.000		.000											
	183	.000	.000											
	.000	.000	.000											
Mod	793		.000	.000										
	.000		.000	.000										

附录 A 统计信息附表

	.000	268	.000	.000								
	.000	.000	.000	.000								
Parataxis	.000		.000	.000	.000							
	110		897	.000	.000							
	.000	.000	.000	.000	.000							
	.000	638	.000	.000	.000							
Dislocated												
	269		954	.000	.000	946						
	.001	.000	.000	.000	.000	.000						
	.000	.001	.000	.000	.000	.000						
Udep	.000		.000	.000	.000	.000						
	.000		.000	.001	.000	.000	.000					
	.000	.000	.000	.000	.000	.000	.000	078				
	.000	053	.000	.000	.000	239	.000					
Appos	.000		.000	.000	.000	.000	.000					
	.000		.000	.000	.000	.000	.000	.000				
	.000	.000	.000	.000	.000	.000	.000	916	285			
	.000	.000	.000	.000	.000	.000	.000	.000	.000			
Compound	.000		.000	.000	.000	.000	.000	.000	.000	285		
	.001		.025	.000	.000	.018	.033	.000	.000			
	.012	.000	.000	125	.000	.000	.000	.000	.000	.000		
	.000	.000	.000	578	.000	.000	.000	.000	.000	.000		
Clf												
	.000		.000	.002	.000	.000	.000	.000	.000	.000	.000	
	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	
	.000	.023	.000	.000	.000	.029	.000	498	.000	.000	.000	
Det												
	.000		.000	.044	.000	.000	.000	.000	.000	.000	.000	322
	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	
	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	
Cc	.000		.000	.040	.000	.000		.000	.000	.001		
	.000		.000	.000	.000	.000	.000	.000	.049	.000	.005	.000
Discourse												
	.000		.000	.000	.000	.000	.000	.043	401			.004
	.000		.001	.000	.000	.001	.006	.000	.000	144	.000	.000
Orphan												
	.000		.003	.000	.000	.002	.005	.000	.000	273	.000	.000
											.000	655

附录 B 动名兼类词词表及相似度⁹

词	词频	动词相似度 ₁	名词相似度 ₁	动词相似度 ₂	名词相似度 ₂
过去	13	0.638	0.597	0.797	0.750
操作	8	0.468	0.467	0.392	0.917
复兴	10	0.636	0.704	0.589	0.584
作战	19	0.693	0.658	0.821	0.729
扩建	6	0.635	0.658	0.888	0.408
表演	11	0.687	0.731	0.790	0.840
检查	7	0.610	0.640	0.673	0.762
抵达	7	0.645	0.631	0.981	0.617
说服	8	0.696	0.616	0.956	0.438
评论	8	0.426	0.427	0.270	0.679
出土	7	0.608	0.668	0.939	0.575
上映	10	0.623	0.615	0.739	0.519
建议	13	0.669	0.672	0.676	0.672
输入	7	0.573	0.642	0.489	0.340
来源	7	0.629	0.625	0.743	0.822
计算	9	0.532	0.526	0.392	0.221
涉及	12	0.758	0.696	0.940	0.462
肯定	6	0.614	0.639	0.805	0.635
成年	7	0.522	0.571	0.489	0.495
收藏	9	0.618	0.634	0.769	0.727
运输	17	0.620	0.686	0.656	0.692
突破	10	0.599	0.545	0.746	0.818
导演	8	0.629	0.643	0.769	0.854
毕业	27	0.610	0.574	0.699	0.692
解释	12	0.778	0.686	0.921	0.596
派遣	10	0.576	0.512	0.937	0.579
分析	11	0.604	0.577	0.803	0.755
参赛	11	0.541	0.569	0.484	0.680
交易	14	0.630	0.724	0.585	0.832
培养	8	0.733	0.686	0.941	0.663
欣赏	7	0.564	0.577	0.699	0.619
联络	9	0.615	0.567	0.901	0.534
胜利	6	0.628	0.625	0.739	0.726
预期	6	0.591	0.605	0.901	0.716
掌握	6	0.714	0.616	0.913	0.589

⁹ 相似度₁指基于词得到的相似度，相似度₂指基于依存关系得到的相似度，

附录 B 动名兼类词词表及相似度

意味	9	0.617	0.556	0.880	0.648
提名	7	0.569	0.582	0.854	0.796
持有	9	0.658	0.655	0.913	0.382
差异	11	0.639	0.631	0.696	0.794
收敛	7	0.593	0.579	0.655	0.550
爆炸	9	0.643	0.646	0.707	0.647
传说	8	0.638	0.611	0.858	0.654
知道	13	0.744	0.611	0.968	0.674
抵抗	6	0.616	0.672	0.849	0.369
写作	8	0.637	0.712	0.650	0.945
调整	11	0.692	0.672	0.958	0.545
了解	8	0.711	0.599	0.905	0.653
支援	8	0.699	0.670	0.892	0.538
防御	7	0.484	0.549	0.219	0.354
防守	6	0.582	0.625	0.723	0.865
兴起	7	0.643	0.618	0.664	0.637
发射	14	0.604	0.611	0.764	0.397
观测	7	0.570	0.591	0.659	0.632
探测	8	0.505	0.560	0.597	0.163
销售	6	0.631	0.614	0.677	0.558
演化	6	0.590	0.672	0.750	0.352
吸收	6	0.589	0.567	0.883	0.402
突变	6	0.660	0.646	0.443	0.789
限制	16	0.705	0.627	0.903	0.635
幻想	8	0.495	0.535	0.400	0.881
雕刻	7	0.579	0.604	0.669	0.349
宣言	7	0.664	0.635	0.780	0.875
代表	75	0.719	0.785	0.886	0.775
决定	43	0.729	0.605	0.960	0.471
行动	23	0.772	0.726	0.707	0.927
生活	46	0.687	0.765	0.680	0.879
想像	8	0.614	0.640	0.920	0.656
应用	19	0.710	0.705	0.864	0.793
竞争	9	0.548	0.623	0.515	0.828
希望	19	0.693	0.608	0.883	0.359
改编	8	0.667	0.665	0.851	0.421
参考	6	0.725	0.666	0.932	0.452
用	122	0.817	0.682	0.899	0.456
设计	53	0.715	0.819	0.811	0.790
促进	8	0.714	0.609	0.878	0.583
通讯	9	0.614	0.765	0.503	0.964

附录 B 动名兼类词词表及相似度

交流	15	0.654	0.735	0.666	0.964
开发	17	0.731	0.733	0.885	0.682
带领	9	0.640	0.580	0.954	0.569
企图	6	0.635	0.543	0.891	0.677
表达	13	0.784	0.633	0.930	0.629
工作	84	0.704	0.757	0.673	0.960
指挥	18	0.701	0.688	0.856	0.422
演出	28	0.733	0.700	0.786	0.862
发现	74	0.777	0.640	0.989	0.517
意识	7	0.619	0.576	0.891	0.620
纪念	28	0.608	0.630	0.704	0.425
占地	10	0.557	0.607	0.629	0.346
组织	53	0.706	0.852	0.622	0.977
创立	14	0.763	0.667	0.937	0.694
解决	22	0.716	0.682	0.952	0.606
进攻	18	0.689	0.663	0.926	0.680
组成	31	0.723	0.676	0.882	0.565
协助	20	0.702	0.581	0.971	0.669
成长	9	0.690	0.695	0.769	0.734
发明	9	0.552	0.573	0.899	0.669
运动	53	0.627	0.713	0.471	0.906
保险	13	0.571	0.617	0.396	0.929
现存	10	0.626	0.685	0.772	0.627
交换	8	0.585	0.636	0.834	0.738
扩大	7	0.617	0.631	0.910	0.408
生产	33	0.668	0.772	0.834	0.769
出现	53	0.808	0.687	0.937	0.595
反应	13	0.719	0.707	0.603	0.896
破坏	12	0.732	0.775	0.803	0.801
下辖	26	0.505	0.551	0.858	0.402
播出	12	0.687	0.668	0.863	0.680
服务	47	0.696	0.772	0.649	0.941
重建	16	0.695	0.632	0.829	0.850
改变	16	0.740	0.665	0.905	0.719
建立	50	0.719	0.663	0.944	0.452
保留	18	0.695	0.658	0.966	0.525
哺乳	6	0.528	0.597	0.206	0.343
栖息	16	0.599	0.559	0.557	0.588
上升	9	0.724	0.640	0.877	0.648
作用	25	0.676	0.722	0.687	0.841
起源	11	0.688	0.681	0.739	0.576

附录 B 动名兼类词词表及相似度

使用	130	0.854	0.749	0.982	0.607
标志	15	0.596	0.590	0.823	0.583
指导	7	0.648	0.679	0.831	0.731
活动	47	0.696	0.703	0.634	0.949
分支	11	0.538	0.577	0.667	0.621
研究	89	0.744	0.801	0.817	0.824
完成	17	0.759	0.655	0.944	0.704
规定	12	0.672	0.734	0.806	0.754
测试	10	0.673	0.665	0.658	0.894
管辖	26	0.596	0.569	0.864	0.508
设置	9	0.740	0.691	0.917	0.509
连接	12	0.708	0.688	0.937	0.471
计划	43	0.789	0.712	0.816	0.882
教学	17	0.512	0.581	0.365	0.860
合作	18	0.665	0.650	0.726	0.796
合并	22	0.706	0.638	0.897	0.671
办公	16	0.581	0.597	0.412	0.758
出版	30	0.742	0.719	0.874	0.755
加工	6	0.548	0.637	0.288	0.688
污染	9	0.676	0.795	0.761	0.922
创造	9	0.654	0.655	0.833	0.424
提高	10	0.703	0.623	0.926	0.583
传播	15	0.686	0.730	0.843	0.748
死亡	15	0.742	0.686	0.809	0.816
成立	83	0.811	0.708	0.843	0.633
批评	10	0.680	0.640	0.811	0.703
游戏	61	0.703	0.763	0.627	0.963
投资	16	0.615	0.696	0.740	0.748
变化	20	0.669	0.629	0.766	0.847
联系	10	0.633	0.593	0.821	0.737
供应	7	0.633	0.660	0.862	0.710
建造	19	0.768	0.716	0.963	0.503
支持	29	0.772	0.705	0.940	0.617
统治	21	0.654	0.771	0.843	0.828
运行	10	0.678	0.717	0.701	0.456
领导	24	0.599	0.642	0.700	0.672
创作	23	0.682	0.765	0.936	0.604
反映	12	0.751	0.661	0.905	0.442
进口	6	0.573	0.546	0.393	0.335
保护	49	0.655	0.743	0.811	0.457
管理	51	0.699	0.789	0.863	0.749

附录 B 动名兼类词词表及相似度

爆发	19	0.750	0.698	0.950	0.593
发生	73	0.757	0.699	0.965	0.550
入侵	11	0.662	0.679	0.894	0.602
命令	11	0.637	0.638	0.931	0.593
改革	16	0.660	0.698	0.690	0.886
制造	27	0.707	0.713	0.766	0.795
象征	9	0.639	0.644	0.969	0.545
开放	9	0.640	0.605	0.966	0.575
存在	38	0.772	0.671	0.916	0.633
批准	8	0.682	0.606	0.893	0.548
调查	22	0.647	0.704	0.675	0.945
处理	15	0.695	0.732	0.865	0.771
选举	25	0.682	0.757	0.664	0.969
融合	7	0.568	0.559	0.862	0.505
播放	12	0.628	0.646	0.900	0.500
纪录	15	0.658	0.647	0.603	0.939
经历	11	0.736	0.698	0.924	0.664
参谋	7	0.605	0.687	0.541	0.885
训练	13	0.582	0.634	0.744	0.896
报告	14	0.685	0.705	0.620	0.930
建设	16	0.653	0.731	0.733	0.948
公布	12	0.601	0.674	0.854	0.575
同意	17	0.795	0.624	0.962	0.523
实验	20	0.597	0.688	0.432	0.834
治疗	20	0.638	0.716	0.555	0.964
打击	10	0.648	0.577	0.886	0.679
选择	22	0.820	0.695	0.939	0.455
注意	11	0.652	0.674	0.713	0.660
描述	14	0.752	0.681	0.895	0.774
记录	13	0.706	0.662	0.837	0.779
生长	13	0.623	0.592	0.528	0.617
制作	40	0.742	0.792	0.952	0.579
报道	10	0.674	0.647	0.737	0.807
命名	28	0.714	0.642	0.854	0.455
改造	6	0.664	0.674	0.877	0.601
来往	11	0.680	0.749	0.975	0.490
出生	28	0.711	0.657	0.664	0.565
教育	50	0.574	0.692	0.421	0.857
运作	12	0.650	0.675	0.692	0.646
效力	14	0.676	0.644	0.739	0.665
讨论	12	0.755	0.714	0.957	0.676

附录 B 动名兼类词词表及相似度

转变	7	0.710	0.672	0.735	0.394
代理	7	0.662	0.714	0.834	0.544
规划	13	0.703	0.784	0.822	0.756
呼吸	6	0.547	0.567	0.261	0.816
拍摄	15	0.760	0.670	0.938	0.511
相信	15	0.661	0.585	0.985	0.581
继承	17	0.762	0.739	0.919	0.483
教授	12	0.728	0.757	0.562	0.901
战斗	30	0.627	0.753	0.612	0.985
移民	13	0.630	0.655	0.888	0.757
主张	14	0.667	0.571	0.951	0.617
威胁	13	0.691	0.673	0.912	0.650
行为	29	0.662	0.752	0.605	0.981
投票	8	0.629	0.720	0.441	0.767
研发	8	0.526	0.571	0.921	0.390
轰炸	9	0.653	0.688	0.848	0.627
证明	8	0.660	0.591	0.897	0.398
抗议	9	0.648	0.569	0.853	0.825
营运	17	0.649	0.664	0.924	0.513
自治	26	0.476	0.492	0.277	0.746
进步	7	0.697	0.681	0.726	0.539
射击	7	0.564	0.646	0.578	0.733
灭亡	6	0.547	0.540	0.806	0.457
贡献	12	0.603	0.606	0.706	0.824
居住	19	0.699	0.670	0.786	0.727
殖民	14	0.505	0.544	0.344	0.190
袭击	10	0.651	0.720	0.696	0.818
扩展	8	0.671	0.628	0.935	0.434
下降	11	0.719	0.671	0.847	0.550
宣传	6	0.561	0.644	0.546	0.880
提升	9	0.721	0.619	0.943	0.704
容纳	8	0.604	0.561	0.918	0.399
重视	6	0.727	0.616	0.967	0.626
监督	12	0.692	0.720	0.735	0.824
冲突	11	0.617	0.599	0.641	0.747
协调	6	0.701	0.674	0.837	0.738
帮助	16	0.728	0.665	0.979	0.529
辅助	6	0.443	0.502	0.539	0.246
定义	9	0.683	0.694	0.725	0.806
飞行	14	0.597	0.656	0.652	0.778
保障	6	0.619	0.737	0.780	0.770

附录 B 动名兼类词词表及相似度

种植	7	0.696	0.716	0.894	0.694
购买	8	0.660	0.612	0.913	0.501
历任	6	0.565	0.632	0.609	0.538
现有	7	0.473	0.498	0.796	0.381
比赛	33	0.700	0.768	0.630	0.959
访问	10	0.674	0.626	0.938	0.543
签约	8	0.620	0.598	0.607	0.660
考虑	11	0.732	0.612	0.907	0.373
确认	7	0.691	0.586	0.968	0.456
预测	6	0.663	0.668	0.622	0.582
攻击	23	0.736	0.703	0.837	0.722
延伸	6	0.648	0.579	0.798	0.371
遭遇	6	0.751	0.656	0.969	0.449
请求	6	0.654	0.639	0.798	0.527
判决	6	0.724	0.667	0.885	0.596
统计	16	0.586	0.675	0.650	0.718
编号	7	0.521	0.556	0.368	0.395
缺乏	7	0.591	0.584	0.876	0.302
咨询	6	0.551	0.634	0.527	0.722
承认	18	0.735	0.609	0.989	0.476
占领	22	0.730	0.667	0.976	0.597
消费	10	0.521	0.554	0.281	0.708
分类	6	0.669	0.678	0.738	0.407
介绍	12	0.799	0.676	0.937	0.487
反对	28	0.771	0.667	0.977	0.497
转移	6	0.632	0.583	0.851	0.570
记载	15	0.658	0.630	0.860	0.642
移动	8	0.670	0.672	0.955	0.570
安排	10	0.734	0.603	0.902	0.605
拍卖	7	0.616	0.643	0.617	0.776
贸易	9	0.708	0.724	0.487	0.968
申请	7	0.771	0.627	0.855	0.652
建国	7	0.586	0.650	0.749	0.548
设定	13	0.589	0.645	0.704	0.807
组合	7	0.637	0.687	0.696	0.879
推动	13	0.762	0.677	0.953	0.495
主持	17	0.702	0.665	0.854	0.607
学习	24	0.735	0.681	0.907	0.810
结合	10	0.623	0.633	0.806	0.494
称呼	6	0.587	0.628	0.652	0.558
欢迎	8	0.540	0.535	0.701	0.344

附录 B 动名兼类词词表及相似度

指定	7	0.680	0.661	0.900	0.399
翻译	9	0.682	0.735	0.895	0.692
邀请	14	0.720	0.633	0.945	0.742
对抗	6	0.661	0.595	0.909	0.533
发表	22	0.763	0.689	0.988	0.595

附录 C 动名兼类词语料例证

动名兼类词	动词性用法	名词性用法
代表	他们代表的是你周遭的亲朋好友 后三位字母代表本班列车停靠的车站	以梁思成为代表 法人代表
决定	由议会党团决定 决定用五角星来象征它	收紧移民的决定 做出了决定
行动	依次行动 即使是白天也能行动	这项行动 保守派已经迅速采取行动
生活	生活于巴黎蒙马特 人类到达它们生活的岛屿之前它们没有天敌	生活资讯 经常描写平民生活
想象	为什么生活会与她想象的不一样 激发玩家想象他们行动可能带来的效果。	比想象中多 决定将各种印象和想象化成音乐作为表达
应用	可以应用这种逻辑系统 广泛地应用于有机合成	有很广泛的应用
竞争	两队竞争乔治哈拉斯杯	竞争期间 在激烈竞争下价格比外地便宜不少
希望	希望考进外交部工作 我们只希望，借着这个历史上真实人物的一生，能带给观众一些启示	充满希望
改编	改编成真人电影 由二人转的传统剧目改编而来	电影改编版权
参考	角色名字参考日本知名漫画家 几乎参考了所有有关武则天的资料	用为震荡电路的频率参考
用	用竹子作的 同样的手法被用于现代的摄影技术中	被改回教堂之用
设计	由荷兰建筑师设计 海德克始终在设计新相机	该站为岛式站台设计 拉斐尔也曾涉及毡幕图稿的设计
促进	促进经济复苏 促进龋齿的形成	起到促进作用

通讯	手机无法通讯	通讯快捷 促进不同系统之间的通讯交流
交流	当时交流的主要形式是主办中日科技座谈	与其他地区的交流。
开发	元朗区是两地交流的通道之一 心境感应是索尼所开发的一个 音乐分类、播放及管理的软体 开发能够解析生物医学文献的 系统	鼓励文化交流 此机制的开发 开发小组
带领	带领日本走向维新 一支由白人军官带领的黑人民 兵洗劫了卡梅隆庄园	在加恩卡纳的带领下
企图 表达	企图推翻巴蒂斯塔政权 始终不愿向对方坦然表达爱意	渡过塞尼奥河的任何企图 作为表达 这是很清楚的表达
工作	在事务所工作的第一年里 毛恒凤用《中华人民共和国劳 动法》起诉了自己工作的公司	教学工作 另办有内部工作刊物《全国青 联通讯》
指挥	邵恩曾指挥过莱夫斯的作品 整个抢金库计划由他一手指挥	首位指挥 意大利钢琴家梅帕器接任指挥
演出	会客串演出 但未获得演出的机会	流行音乐与古典音乐相互结合 的演出 前季主角江口洋介没有参与演 出
发现	发现了此行星 旧议会大楼的遗迹在一次挖掘 中被发现。 静华后来发现叶广义不爱自己 便偷取美丽的设计创意投靠云 想集团。	最激动人心的发现
意识	医生终于意识到外星人们只想 要他的歌声	宗教意识浓厚 强烈而无意识的冲动
纪念 占地 组织	纪念唐朝诗人杜甫 公园占地 320 亩 组织民兵抗击 中国大陆地区的学校通常在暑 期或 8 月末对新报到的学生组 织军训	森林草坪纪念公园 其中占地面积为 6100 平方米 组织解散了 许多国际国家组织对此表示了 不满
创立	次年司马炎创立西晋 由许荣茂先生创立	地理学的创立时期

解决	此法能解决问题	一种解决方法 得到解决
进攻	从东进攻 乘瑞典军进攻日耳曼南方之际	游击队开始进攻行动 奥斯曼帝国苏丹苏里曼一世的进攻
组成	物业由两幢高层办公塔楼及一座低层建筑组成 组成的字母序列要便于发音	这样的组成模式 30 岁左右的丈夫与 10 几岁妻子的组合
协助	共同组成独立的马来西亚 河流生物学家协助研究大陆水系沉淀物流失	得到 JAXA 的协助
成长	协助盟友塞尔维亚 法兰克便在当地成长 体色会随成长而改变	在善用兵的一益协助下 在木尼河区的成长经历 原始扰动的成长形成今天我们所观察到的所有结构
发明 运动	诸葛亮发明八阵图 他每天都运动	酒精饮料的发明 人气最旺的运动 世界三项全能运动协会
保险 现存 交换	被保险的人 现存的双吉寺建筑 双方交换了友好文书	保险公司 现存还有 300 多个 互惠互利的合理交换 第三次社会大分工首先在商品交换最为发达的地区出现
扩大	印刷的范围扩大到其他经典 丘吉尔则同意扩大在缅甸的军事行动以稳固蒋介石在中国的地位	交换的不断扩大 在教总理事扩大会议上
生产	生产文体教玩具 生产出来的橙	葡萄酒的生产中心 商品生产出现
出现	东突厥内部出现分裂 商品生产出现	N 解药的出现
反应	汞可以与硫化氢反应	这个反应 与五氧化二氮催化臭氧分解的反应有关
破坏	跑道被破坏 破坏水生的生态系统和鸟类	生长与破坏形成了一种平衡 造成很大破坏
下辖	明中乡是下辖的一个行政单位	荷花为中国湖南省浏阳市下辖街道
播出	在本集播出后 于晚间 10 时 00 分播出	延长播出时间

服务	为人民服务 主要服务于西岛及附近地区的 华人社区	最高水准的酒店服务 提供资讯和娱乐服务
重建	1599 年重建	此桥的重建 科学家们最早的重建揭示出一 个没有头的人形雕像
改变	改变信仰 从哥特式改变为巴洛克式教堂	发生改变 组织文化与管理结构上的改变
建立	建立了日本与新西班牙的正式 外交关系 张鸿南建立的产业	南京临时政府的建立
保留	汉字语音保留了中古音的入声 体系 这两座塔都被保留	持保留意见
哺乳 栖息	这个妇女在给婴儿哺乳 本鱼栖息于潟湖或珊瑚礁 使该地成为商人和他们货物的 重要栖息之处	野生哺乳动物 栖息地点
上升	产生于君权上升之时 上升了大约 25%	受到上升作用
作用	与原子核相互作用 这种原理作用于他身上	氢和氧之间的氢键作用 对君主的制约作用
起源	一些大学起源于教会学校	它的起源
使用	她常使用博多方言 拜占庭军所使用的其中一种武 器	奈米技术的使用 魔石的使用次数有限
标志	这项行动标志着冲突开始急剧 升级	这正是往日屯门海岸线之标志 家庭成立的标志
指导 活动	指导总体经济体制改革 成群活动 易于活动	挑选他作为指导师傅 民族识别活动 1999 年的全国性反美活动
分支 研究	由康宁道分支出来的道路 研究人类行为 为了科学研究	运输工程学中的一个分支 作为医学研究
完成 规定	完成空中接力 查理大帝规定每座修道院都要 办一所学校	土地改革的完成 规定完成时间
测试	被测试的动物	要求通知父母的规定 不计其数的心理测试 夏志明被伯明翰邀请接受测试

管辖	由地方公安机关管辖 隶属昆明铁路局管辖	脱离其管辖
设置	全校课室均有设置电脑 千阳号也特别为她设置了花圃	新闻总署的机关设置
连接	由高速公路全程连接 弓毛库连接着弓尾	无障碍的便利连接
计划	南航计划以更高效率的空客 A321 取代它们	白素贞的计划
教学 合作	他重回桥梁道路学院教学 合作的形式 和欧洲及世界各地来沪的著名 音乐家合作演出	校园小记者计划 学院的教学工作 保持着合作关系 开展形式多样的合作交流
合并	1999 年合并原上海农学院	第四批全国重点文物保护单位 合并项目
办公	玉带镇与长芦街道合并为新的 长芦街道 在广州办公 1927 年 2 月 21 日武汉国民政 府正式办公	两幢高层办公塔楼
出版	1826 年出版的《最后的萨拉只 家族的传奇》 在 1978 年首次出版 该校中文学会每学年都会出版 文集	相当于出版秘书
加工	经过红曲菌发酵加工 加工农产品	加工企业 废旧物资回收加工
污染	会污染水源 污染莱茵河的事件	西湖水质污染
创造	创造了对犹太人比较宽松的 生活环境	引入了创造主题
提高 传播	提高烟草的产量 向海外地区传播 传播到埃及、中东和中亚地区	知识的创造 电影制作国际化水平的提高 传播距离有限
死亡	岛核电站一名抢险人员突然死 亡	造成许多无辜生命的死亡
成立 批评	翡翠 1 台于 1984 年成立 批评美国内阁成员出卖盟友	成立大会 批评亦有不少 就中国兔毛产业提出批评
游戏	游戏人间	游戏在北美地区发售

投资	跨国公司进入园区投资 嘉木扬·图布丹联络企业家陆 续投资 160 万元人民币	游戏大师 涉足的投资领域非常广泛
变化 联系	春天天气周期变化 将特菲尔和莫札特的角色联系 起来 记者随即联系了郭德纲的经纪 人	解释某些变化 彼此文化联系有着悠久的历史 黎波里与突尼西亚边境的联系
供应	供应中国境内家庭食用	合法的供应非常稀少 日后配件供应
建造 支持 统治	它在空间系统中心建造 一直支持工党 姚里氏统治辽东 由一个主持该城市的宗教仪式 的祭司或国王统治	参加建造工作的人 在他们的支持下 在他的统治时代 对思想的专制统治
运行	木星运行到太阳和天王星之间	伽利略号的运行轨迹 各项科学技术的运行
领导	大不列颠棉花行业领导的工业 革命	局领导认为 亚历山大三世的领导
创作	席勒开始创作剧本《强盗》 大量优秀的作品就是这一时期 创作的	从事传统漫画创作工作 埃米内斯库的创作生涯
反映	这些内容明显反映了司法界官 僚的意图 反映巴黎历史变迁的资料片	引起连锁反映
进口	多达一半进口自美国	汽车均以台湾、日本及韩国进 口二手车为主 这些进口使其他工业的投资萎 缩
保护	不需要保护补给线的安全 为了保护以色列	该物种的保护状况
管理	该组织由一理事会大会管理 管理销售额 100 亿美元的零食 部门	毕业于经济管理专业 做必要的管理
爆发	爆发九月政争 抗日战争爆发后	第二次世界大战的爆发
发生	首次经验发生在初中三年级时 没有发生特别重大的疫情 在中国大陆发生的文化大革命	西安事变的发生

入侵	亚述人和埃兰人入侵 外来食肉动物的入侵 他入侵乌拉圭	
命令	阿保机命令他支援	华盛顿发表的命令 服从任何命令
改革	改革了现有制度	推动国内的民主改革 这项改革
制造	用它们的皮来制造与修补船只 最初制造出来的磷酸	中国制造
象征	象征荣耀、诗歌和音乐 用五角星来象征它	象征式的国家元首 最富有特色的象征
开放	京都御苑对公众开放	开放时间为早上 10 时至晚上 8 时半
存在	开放光棍节系列活动 有珊瑚礁存在 实际存在的历史人物	审美距离的存在
批准	国家计委正式批准榆林为国家 能源重化工基地 波兰下院正式批准发行波兰的 货币	大法官威廉·巴克的批准
调查	文帝派使者去调查	参加了民族调查 民意调查
处理	交由华盛顿会议处理 处理家务	这样的处理 土壤处理
选举	议员每 2 年会选举一名议长	参加国会两院议员选举 选举伎俩
融合	雅语和吴语融合 它融合了动作游戏的一些特征	恒星核合成过程的融合
播放	可播放广播 唯一经常播放的就是这部电视剧	一个音乐分类、播放及管理的 软件
纪录	记录了种植、收获、播种和土壤 处理等农业技术	女子世界纪录 留下的记录 出生记录
经历	大韩民国经历了民主与独裁统 治的反复交替 印度也经历了一场空前的浩劫	巴黎的经历 吸毒经历
参谋	让他参谋参谋	第一航空舰队航空参谋 澳洲参谋总长

训练	它们很容易被训练	负责军事训练工作
报告	在迈阿密训练后 中国部分网站报告了相关新闻	接受步兵的训练 报告也考虑到了意外的成因 中国体育博物馆房屋安全检测报告
建设	建设该中心的目的 近期建设的伊丽莎白港区 建设时未使用设计方案规定的 墙体砂浆	港口建设资金 从事杭州轨道交通工程建设
公布	提名名单公布了	公布日期为 1983 年 10 月 12 日
同意	将之公布于众 吉西的父母同意进行实验性治疗 洪秀全同意了许多建议	依赖于各州政府的同意 在民主党领袖卡斯特伦的同意下
实验	被实验的人员	从实验上 对阿兹海默症的医学实验
治疗	治疗各类患者 常用于治疗过敏症状	该治疗方法 如果气胸得不到治疗
打击	坚决打击和依法严惩 打击改革派	作有效的打击
选择	选择婴儿性别 选择重回妻子身边	天然光源外的另一选择 霸占该区域是最有利的选择
注意	应当注意的是 请特别注意	没有受到多少注意 引起不少欧洲豪门的注意
描述	该小说故事主要描述爱情故事 于 1932 年首先描述	关于十八铜人的具体描述 宽频是一种相对的描述方式
记录	通过密码记忆系统记录进度 记录了种植、收获、播种和土壤 处理等农业技术	对记录虎视眈眈 以前的唐朝早期记录
生长	有珊瑚礁生长的条件 生长于海拔 1,500 米至 2,400 米 的地区 生长在印度南方	生长与破坏形成了一种平衡 适应其生长环境 原制作班底
制作	由俄罗斯与法国公司联合制作 自家手工制作的 制作了自己的相机	他也参与了制作 参与电影的配乐制作
报道	日本报章也报道中国发现了天 才神童 着重报道美国和其他西方侨民	有报道说 作了报道

命名	在中国的活动 学校以明朝大儒王阳明命名 由松坂的曾祖父母命名 命名为东京街西	车站命名与线路的等级相适应 命名和定义并不统一 他们的命名是取自于《神奇之书》
改造	不惜改造自己唯一的儿子 旧街坊改造后	众故宫改造方案之一 重音的改造 整体改造
来往	来往于洛布尼亚和萨维奥洛夫站的机场 严令儿子不得与李敏来往 来往蒲飞路及北角码头	本区成为来往通道 与朝鲜有贸易来往
出生	在他出生之前 出生于一个经济状况良好的天主教家庭 孩子便出生在这里	出生季节
教育	他这样教育孩子	亲自负责政治教育 英国教育制度
运作	运作得不好 由中国动物园学会组织运作 以相同速度运作	它的起源和运作 参与了许多管道的运作 维持正常运作
效力	他目前效力于法甲球队波尔多 期间效力多支西班牙低组别联赛球会	票据所具有的法律效力 其全部效力
讨论	集中讨论理论问题 讨论为什么东京的城市建设比不上沈阳	有足够的讨论时间与效率 必要性的长期讨论
转变	过冷液态转变为玻璃态 转变教学方式	这个转变 不纯的金属到金的转变
代理	由尖端出版代理发行 已正式委托律师代理这起名誉纠纷	一家广告代理公司 代理司令长官
规划	在规划基础设施和矿产探测中 规划时曾称为赛马场站 规划为七座主要客运站之一	规划未能实现 高速公路建设规划 在规划过程中
呼吸	畅快地呼吸	无氧呼吸
拍摄	张艺谋拍摄了电影 经常被拍摄成为电视剧 在美国路易斯安那州拍摄	使用拍摄作品展示了上海人穿着睡衣上街的情景 确定拍摄场景后
相信	甘地更相信尼赫鲁	促成人的“相信”

继承	他才相信自己能打得比老鹰队上的控球后卫好 胜赖无力继承 继承父亲的日向佐土原城	财富及继承问题 与亚历山大帝国其他的继承政权作战
教授	他继承了父亲的工艺和博学 他教授考古学	山手大学教授 李维祀教授 地质学教授
战斗	在战斗时	发生战斗 其战斗背景和战斗人物
移民	移民海外 移民到新加坡	有些移民 外国移民
主张	他主张进攻利潘迪特兰堡 主张“实业救国”	赞同外相东乡的主张 有“世界大同”的主张
威胁	威胁人类的安全 巴哈欧拉的声明威胁了米尔扎·叶海亚宗教领袖的地位	带给她本人的威胁 高卢带来的威胁 对波音的威胁
投票	投票通过 由人民投票选出	根据投票结果 违背法案的投票
研发	成功研发疫苗株 由他所研发	高速列车研发实验基地
轰炸	意外轰炸到苏黎世 北约轰炸中国使馆	美军的轰炸 遭到猛烈轰炸
证明	这进一步证明了和性别有关 证明其安全性没有问题	进步的证明
抗议	引起中国人民抗议 抗议北约轰炸中国使馆 抗议反动当局的暴行	第二次抗议 提出了抗议 要求抗议群众撤离
营运	JR 东海负责营运东海道新干线 此路网于 1855 至 1938 年在巴黎营运 在凡尔赛则一直营运至 1957 年	接驳车的营运业务 随着 A380 将投入营运
自治	得以高度自治	格陵兰自治议会
进步	古希腊艺术也随之进步	脑硬件技术的进步
射击	近距离开枪射击	国际射击中心
灭亡	清朝灭亡后 马其顿王国灭亡了波斯帝国	清朝的灭亡
贡献	贡献了主题	对天文学的贡献 有很大的贡献

居住	有人类居住 在大连、天津居住	在公共事务方面的贡献 健康的居住环境
----	-------------------	-----------------------

致谢

我总认为，我就好像进了一间有数不清屋子的大房间，一心迷迷茫茫，随手那么一推，正好就推开了一间屋，这屋里又正好是值得我热爱的东西。尽管现在的我仍是小孩子做游戏式的做研究，有时候笨拙得自己都觉得愚蠢，但我还是认为语言学研究对我是有重大意义的。

坦白说，完成这篇并不算出色的文章并不容易。开题时仓促不已，写作时困难重重，修改时眼花缭乱。为了完成这篇文章，我学到了很多，包括技术上的，也包括内容上的。这些东西并不简单，尤其还涉及到编程、数学等等，但幸好我学到的这一点浅薄知识勉强能帮得上我。

能完成这篇文章，我首先要感谢我的父母。他们是最朴实善良的人，是最慈爱无私的长辈，是我人生中永远的光芒。他们始终如一地支持我的各种决定，几乎无条件地包容我的任性，却总是觉得自己做得还不够。他们永远信任我，永远爱护我。我以有他们为父母为幸。

还要感谢我的导师蔡言胜老师，是他帮我走上数据驱动研究的路，在论文完成过程中和三年的生活中蔡老师对我有诸多帮助和关照，感谢蔡老师的教诲和指导。也要感谢郭利霞老师、温宝莹老师、夏全胜老师等诸位师长给予我的无私帮助、鼓励和认可，这对于我这个面对科研始终心里没底的笨学生来说太重要了。

感谢刘海涛教授对我冒昧叨扰的包容，并耐心指导我这个粗心大意的愣头青完成了一点小成绩。刘教授学问精深广博，品德高尚坦荡，为人幽默和蔼，是我脑袋里“大先生”的“理想化认知模型”。“高山安可仰，徒此挹清芬”，希望未来我能从刘教授身上学到万分之一。感谢郝瑜鑫教授慷慨提供了中介语树库资源，没有郝教授的帮助，这篇论文恐怕还要大费周章。

感谢同学杨苛鑫、徐利峥、韩菁雯三年的照顾和担待，与你们的交流使我获益颇丰，让我的生活也不至于那么枯燥。感谢师妹盛建霞、冯源给我的帮助，希望数次求助没让你们觉得困扰。与你们为友很愉快，愿你们能得偿所愿。

虽然多少有点迷迷糊糊，但总算走到这里了。希望能在下一个致谢里等到更好的自己。

个人简历

杨牧（1999.02.05-），男，河北邢台人。于2016年6月山东大学翻译学院英语专业取得文学学士学位，于2020年9月-2023年6月南开大学汉语言文化学院语言学及应用语言学专业攻读文学硕士学位。

硕士阶段研究成果如下：

[1]杨牧,蔡言胜.依存句法分析的回顾与发展[J].现代语文.2022(01):89-95.

[2] Yang Mu & Liu Haitao. The role of syntax in the formation of scale-free language networks
2022 EPL 139 61002.

获奖情况如下：

1.2022年获南开大学公能奖学金二等奖一次；

2.2022年获南开大学学术竞赛类专项奖学金一次；

3.2022年获南开大学汉语言文化学院五四论文比赛三等奖一次。