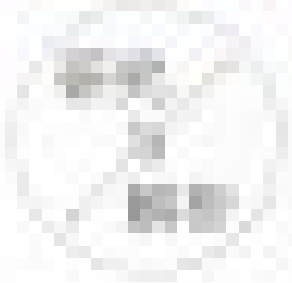
蒋平◎著

基于数据库的汉语零形回指

研究与解析

中国社会科学出版社

基于数据冲的
双循环形回指



蒋平◎著

基于数据库的汉语零形回指

研究与解析

中国社会科学出版社

图书在版编目(CIP)数据

基于数据库的汉语零形回指研究与解析 / 蒋平著. —北京：中国社会科学出版社，2017.9

ISBN 978-7-5203-0841-0

Ⅰ.①基… Ⅱ.①蒋… Ⅲ.①数理语言学-应用-汉语-研究 Ⅳ.①H1-0

中国版本图书馆 CIP 数据核字(2017)第 205633 号

出 版 人 赵剑英
责任编辑 熊 瑞
责任校对 季 静
责任印制 戴 宽

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲 158 号
邮 编 100720
网 址 http：//www.csspw.cn
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印 刷 北京明恒达印务有限公司
装 订 廊坊市广阳区广增装订厂
版 次 2017 年 9 月第 1 版
印 次 2017 年 9 月第 1 次印刷

开 本 710×1000 1/16
印 张 22.5
插 页 2
字 数 318 千字
定 价 99.00 元

凡购买中国社会科学出版社图书，如有质量问题请与本社营销中心联系调换
电话：010-84083683

国家社科基金项目：
基于数据库的零形回指解析方式实证研究，09BYY001

前　言

一

零形回指又称零回指、零形式或零形指代（zero/null anaphora），是指在语言表达中再次提到某个指称实体时，采用零形式进行指代，表面上没有具体的语言符号或语音形式的指称现象。当今，回指的先行语解析问题已经成为计算机自然语言处理与机器翻译中需要解决的几个难题之一，零形回指是其中的难点。汉语是零形回指使用频率极高的语言，其先行语的确认或计算机解析是汉语自然语言处理和机器翻译中的难题，引起了国内外语言学研究者及计算机科学研究者的关注，至今没有找到可靠的解决办法。本研究是基于数据库的零形回指研究与解析，以计算机自然语言处理及机器翻译中关于汉语零形回指特征的描写与运用、规则的建立和零形回指先行语的解析为目的，选择了汉语三种不同体裁的语篇语料，对其中的零形回指、零形回指的先行语及其相关的各种因素包括距离因素、位置及位置关系、主题性、生命性、层次性、平行性、关联性等进行分析，对这些因素在语料上进行标注之后，建立了相应的数据库记录这些信息。在获得了对汉语零形回指的重要特征描写的基础上，形成了零形回指先行语的解析规则和原则，并对本研究的数据库中的6401例零形回指的先行语进行了解析试验，获得了可靠的结果。同时，为了测试本研究所使用的特征和规则在智能化计算机运算中的可行性和可推广性，本研究将数据库中的零形回指关系整理为124类，将其所涉及的108例语句共1230项有关成分全部标注之后，在智

能化的 Logistic Regression 解析系统中进行尝试，获得了较为理想的结果以及对本研究所采用的特征和规则的检验。

二

本研究的理论框架是认知功能语言学。包括可及性理论、层次性理论、中心理论等。但在运用中，对其中的内涵和要素进行了扩展和调整。

基于 Ariel（1988）的可及性理论，本研究以语篇距离作为基本标尺，在以往研究的基础上，对零形回指的分布情况进行了再调查，同时，以凸显性为指导，对零形回指先行语和回指语所在的位置进行记录和分析，结合框架因素和竞争因素对零形回指的基本规律和特殊规律、先行语的特点、干扰成分的数量和强弱等进行分析。

借鉴 Keenan & Comrie（1977）的名词短语可及性等级体系，本研究对语料中零形回指先行语的句法位置规律进行了分析，得出汉语零形回指先行语的句法位置也有不同的可及程度。不仅如此，零代词的句法位置也同样具有可及性标示作用。先行语与回指语均在主语位置的可及性显然要高于先行语和回指语都在宾语位置或旁语位置的情况。为此，本研究将可及性概念在距离因素的基础上，进一步扩展到句法位置，是一种比 Ariel 的距离可及性分类体系更为具体而微观的做法。

运用 Fox（1987）和 Chen（1986）关于语篇结构的层次性和平行性的观点，本研究讨论了汉语零形回指和先行语所在结构的层次、层次的特点及其相互关系。不仅如此，我们还将这一概念拓展到研究回指环境中的其他语篇单位的结构层次，也用于对零形回指位置发生变化、反指和不同形式的长距离回指等问题的讨论。

采用中心理论和指称参照点等理论的核心概念，本研究进一步讨论了零形回指先行语的凸显性或主题性、回指的延续性、前指和后指、跨句回指等问题，并在中心理论的框架指导下，在已有研究与试验的基础上，基于本研究的数据分析，对回指解析的有关参数与特征进行了调整和修改，并对零形回指的先行语进行了解析试验。

三

在语料的选取上，本研究采取了比以往研究更为合理更为完善的考虑，选择了三种不同体裁的汉语语篇语料。之所以这样选择，是因为目前除了叙事语篇有较多的研究之外，新闻语篇和学术语篇的零形回指研究很少，很难知道其内部的零形回指面貌。为了使本研究的语料有更广泛的代表性，我们选择了三种体裁，同时，也希望借此机会能够研究不同语篇中的零形回指异同。之所以在叙事语篇之外添加新闻和学术语篇，也是考虑了这两种语篇的广泛性。本研究对语篇的年代跨度也有较全面的考虑，主要从20世纪30年代（新闻与学术语篇从50年代开始）到21世纪初，按照改革开放前的语篇和改革开放后的语篇分为早期和近期两个时段。三种语篇的语料各选取9万字，共27万字。

在对语料的标注和分析上，本研究不仅根据实际情况需要，对位置、距离、层次等因素进行了扩展，还对一些以往较为模糊的标准进行了界定，取得了创新。其中包括对零代词的标注标准的确定。这是一项较为复杂的工作，且在以往研究中没有统一的标准。本研究制定了零代词标注的10条细则，包括对以往未得到关注的零形回指的标注。例如，汉语中经常出现“的”字修饰语结构以动词开头，且其前未带名词词组（相当于英语的定语从句）。在本研究中，其动词前均计算为有一个零代词，定位为次层次结构（从句结构）中的零回指。

语料的基本标注单位也是棘手的问题。本研究根据语块理论以及当今计算机语言学正在建设大型语块库的新动向，结合对零回指语料的反复考察，规定语料的基本标注单位为词块。以切合自然语言表达的实际情况和语篇组织的心理过程，也与计算机模拟自然语言处理的最新做法取得一致。同时，本研究还对一些特殊的语块的划分进行了规定，使用了统一的标准。关于句子、小句单位的确定，本研究除了将句号、问号、感叹号、分号计为隔句符号，前后编号为不同的句（其中，将分号计为隔句符号是本研究的新做法）；将逗号、冒号、破折号计作小句分隔符，前后编号为不同的小句之外，还对标点符号过多的和无标点符

号情况下的小句单位进行了界定。此外，还对语句的结构关系、语句的结构层次与结构层次类型、反指、定语从句中的零形回指，以及语料标注所涉及的句法成分、句法成分的内部分类等各种情况均进行了确定。例如，层次标注分为主次两种。插入算作次层次，前置定语从句、状语从句算作次层次，后补小句、内嵌句算作次层次，宾语从句算作次层次，定语从句中包含的定语或状语从句以及其他从句中包含的从句算作二级次层次。次层次的小句，未必仅指语法上的从句，而是语义上的，而且，也可以是整句作为次层次的结构。在标注宾语及介宾类的成分上，本研究将“有”之后、“是”之后、其他存现动词之后、“把、将、连”之后的名词成分，以及直接宾语和间接宾语等均采取分小类标注，以便考察之用。

四

经过多年深入而反复的分析、研究与探索，本研究获得以下一些主要发现：

1. 关于零形回指在汉语三种不同体裁中的个性和共性。研究发现，叙事语篇的零形回指最多，学术语篇的最少。从回指的句法位置看，新闻语篇的最简单，其次是学术语篇，叙事语篇的零形回指最复杂。从位置关系上看也是如此。另外，三种体裁还在回指关系的各自特色、指称对象的生命性和非生命性、抽象回指、无指等方面呈现出差异性。然而，它们有着重要的共性，包括：三种体裁的零形回指位置都以主语为主，其他位置上的零形回指都很少。在先行语方面，三种体裁也都是主语占据绝对数量，共有 4643 例。其次是宾语位置，共有 837 例，其次是主题位置，共 828 例。其他位置上的先行语都很少。在回指的位置关系上，三种体裁都以 S—S（S 表示主语）位置关系的零形回指最多，而且，排在前三位的都是 S—S、T—S、O—S（T 表示主题，O 表示宾语），排在第四位的都是 O—O 位置关系的零形回指。其他位置关系的零形回指很少，约为 6%。最后，在三种体裁中，叙事语篇的零形回指最具有代表性，从而说明，以往研究选择叙事语篇作为研究的材料有很

好的代表性，也可以说明，零形回指的解析，就书面语而言，无须考虑分体裁解析。

2. 汉语零形回指的距离特征。在 Ariel（1988，1990）的研究体系中，先行语和回指语之间的距离是衡量先行语可及性的第一因素。我们发现，汉语零形回指也一样。距离是最重要的因素。在以往研究的基础上，本研究根据实际情况，对距离的统计做了调整，分为：隔段（先行语和回指语所在的段落之间还有别的段落）、前后段（先行语和回指语在相连的前后段落）、隔句（先行语和回指语所在的句子之间还有别的句子）、前后句（先行语和回指语在相连的前后句子）、隔小句（先行语和回指语在同一句子，但是之间还有别的小句）、前后小句（先行语和回指语在同一句子，并且在相连的前后小句）、小句内（先行语和回指语在同一小句内）。统计发现，汉语的零形回指绝大多数出现在先行语为前后小句的语境中。其次是先行语和零回指在同一小句内。第三个典型的距离是前后句，即先行语和回指语出现在前后相连的句子中，其间没有其他句子或者小句。加起来看，零形回指在这三种相邻语句中使用总量达到了 6037 例，占了总数的 94%。这为零形回指先行语的解析提供了极为重要的信息。

3. 句法位置上，本研究发现，如果说距离可以体现汉语零形回指使用的基本面貌，位置规律则更为突出。汉语零形回指的先行语最集中地出现在主语位置，占总数的 72.5%。其次是宾语位置和主题位置，分别为 13.07% 和 12.9%。主语的属格语位置有一些先行语（共 64 例）。旁语和宾语属格语位置有少量先行语，分别为 23 例和 6 例。没有发现其他位置的先行语。汉语的零代词更是集中在主语位置，共有 5916 例，占了总数的 92.4%。还有一个重要的发现是，零代词在其他位置出现的机会特别少，宾语位置和主题位置有一些零代词，分别为 274 例（4%）和 198 例（3%）。只有 10 例零代词出现在主语的属格语位置，3 例出现在旁语位置。其余位置没有零代词的使用。

就位置关系看，先行语和回指语都出现在主语位置的零形回指最多，共有 4476 例。其次是主题—主语位置的零形回指，共有 722 例。

这两种合计有5198例，占零形回指总数的81%以上。类似的位置关系还有主题—主题，共54例，主语—主题，共130例。与前两种加起来，共有5382例，占了回指总数的84%。其余位置关系的零形回指数量不多，却有几种是以往未曾讨论过的。例如，“主题—旁语”、“主语—旁语”、“宾语—旁语”位置关系的零形回指，本研究各有1例；“主语属格语—主题”位置的零形回指，本研究有2例；“宾语属格语—宾语”位置的零形回指，本研究有1例。此外还有“旁语—主题”和“旁语—主语”位置关系的零形回指。因此，本研究在语料上的扩充，获得了更多未知的零形回指位置关系。

在位置关系中，我们还发现了比较出格的两类。一类是先行语出现在宾语属格语或旁语位置的零形回指。这些位置在语篇表达中很不凸显，出现零形回指的先行语的可能性非常小。另一类是零代词出现在旁语位置，其先行语分别出现在主题、主语和“把”、“将”、“连”引入的名词成分的位置上。这些先行语的位置在语篇表达中是比较凸显的，可能说明，零形回指在旁语位置的使用，至少其先行语要足够突出。

4. 零形回指句法位置的排序。综合对零形回指位置的分析和比较，本研究参照Keenan & Comrie（1977）提出的关于不同的句法位置上的成分实现定语从句化的可及性等级体系，在以往研究的基础上，获得了对汉语零形回指先行语的句法位置排序的修正：

T > O1/O3 > SM > S > O > OBL/OM

以上等级体系的重要性不仅在于，它把Keenan & Comrie（1977）关于定语从句化的句法位置等级体系进一步扩展运用到零形回指的研究，更重要的是，在新的体系中，不仅主语的属格语排到了主语的前面（以往的研究或者把它笼统地作为属格语，列在旁语之后，或者作为其他类，列在序列的最后，最好的是将它排在主语的后面），宾语的属格语与旁语并列（以往的研究或者通通算作其他类，列在最后，或者将宾语属格语排在旁语前面），而且单独考虑了存现结构的宾语，并置于以上等级中仅次于主题的地位。

王德亮（2004）在基于中心理论的零形回指研究中，运用了“主

题 > 主语 > 宾语 > 其他”排序，结合其他规则的运用，取得了较好的回指确认结果。段嫚娟等（2009）对中心理论的排序进行了改进，吸收了将 SM 靠前排列的做法，取得了更好的效果。相信本研究基于大量数据分析获得的位置等级序列的进一步修正和优化更有利于中心理论的回指解析。

5. 零形回指的生命性与主题性。以往研究对零形回指先行语的生命性已有一些讨论，但未有过详细统计。本研究对此进行分体裁统计发现，除了学术语篇之外，零形回指的先行语最主要的是生命性的实体，其次是无生命的实体或事件，再次是抽象概念。场所、时间和天气很少成为零形回指的先行语。从数量最多的两种指称对象来看，生命性的实体和无生命的实体或事件无论在三种体裁中，还是在总体比例上，都占绝对数量。在本研究的 6401 例零形回指中，共有 3695 例先行语是生命性的实体，占总数的 57.725%，内部存在一定的体裁差异。新闻语篇中，生命性实体担任先行语的数量是 1025 例，占该体裁零形回指总数的 51.61%；学术语篇中，生命性实体担任零形回指先行语的数量只有 426 例，占该体裁零形回指总数的 30.47%，排第二位，比例最高的是非生命性的实体或事件担任先行语，达到了 49.57%；在叙事语篇中，生命性实体的数量最高，达 2244 例，占该体裁零形回指总数的 74.38%。

三种体裁的零形回指先行语在生命性上的差异与体裁的特点密切相关。新闻语篇的报道不仅要以人物为主，向读者详细说明事件的原委也是报道的主要任务。这就是为什么在新闻语篇中这两种类型的先行语比例都高。学术语篇不关注人物的进展，而以研究对象、问题或研究目标为主题，因此，生命性实体的零形回指数量和比例都不高。叙事语篇以人物描写和人物的活动为叙述的主体，因而，生命性的实体被关注的频率最高，这也是为什么其生命性实体的零形回指比例最高。然而，在总体数量上看，汉语零形回指先行语的生命性排序很清楚，可以形成以下等级体系：

A > I > C > L > T > W

（A 为有生命的实体；I 为无生命的实体或事件；C 为抽象概念；L 为场所；T 为时间；W 为天气）

生命性和主题性有内在联系，基于这两方面的关系，本研究成功地解释了为什么在句法位置的等级体系中，主语的属格语和存现结构的宾语位置可以靠前，以及“把”、“将”、“连”所引入的成分可以成为零代词的先行语。

本研究发现，主语的属格语能够优先于同等条件下的主语是因为它所表达的指称对象的生命性。生命性带来了主语属格语在认知框架上的凸显性。由于生命性实体的运动性和方向性，较之非生命的、静态的、抽象模糊的实体引起更多的注意，更容易成为表达的出发点及关注点而成为先行语。因而，主语属格语的优先排序得到了认知语言学的凸显观和图形—背景观的支持。

存现动词引入的宾语在过去多作为一般的宾语处理，直到近些年关于主题引入方式的研究对这一类宾语的主题性提出了新的看法。本研究不仅发现，当先行语环境中有存现句 S 和 O 时，零代词更多地回指 O，而且发现，这是由于存现宾语的认知地位决定的，并从认知语言学的图形—背景关系的角度进行了成功的解释。

汉语介词引导的 NP 在下文很少用零形回指，“把”、“将”、“连”则是例外。这也是由于这些成分的强调作用及所引导的 NP 的次主题地位决定的。

6. 平行性和层次性。本研究发现，零形回指的平行性有句法位置平行性和语篇结构平行性两方面。一般来说，语篇结构平行，句法位置则平行。句法位置变化，语篇结构一定出现了不平行。比较典型的位置变换是：1）先行语在宾语位置，零回指出现在后续的主题或主语位置（O—T/S 位置关系）；2）先行语在主语位置，零回指在后续的宾语或旁语位置（S—O/OBL 位置关系）。

关于层次性，本研究发现，并不是所有的层次变化都引起零形回指位置的变动。这与以往的看法不同。出现在主句结构之前的从属状语（FAD1）、出现在主句结构之后的从属状语（BAD）、插入的状语结构

（IAD）、插入句（INS）和出现在主句结构后面的内嵌结构（EBAD）基本不影响零形回指的位置。出现在主句结构之前的从属定语（FAD2）和出现在主句结构之前的包含着状语或包含着另外一个定语的从属定语（FAD3）绝大多数也不会引起零形回指位置变化，只有当该定语结构所修饰的中心成分是宾语的时候，才出现零形回指位置变化。综合各方面的分析，本研究发现，出现在主句结构之后的从属补语（BADC）是主要的引起零形回指位置变化的次层次结构，其中又分为存现结构的补语、动补结构和补充说明结构，零代词往往出现在这些结构的主题/主语位置，先行语出现在其前的上一层次的宾语位置。

虽然从总体上看，零代词发生位置变化是由于其先行语出现在更高的结构层次，也有反过来的，这种情况多出现在 S—O 类型的换位回指中。因此，与以往观点不同的是，本研究对零形回指的换位更客观的解释是结构层次发生变化而产生的，未必都是从高的结构层次到低的结构层次。

关于平行性和层次性，还有两个重要的现象是，反指多出现在平行的句法位置，但多不是平行的结构。有两类次层次结构中出现零形反指。一种是在 FAD1（出现在主句前的从属状语结构），一种是在 FAD2（出现在主句前的从属定语结构）。

长距离零形回指多与平行性和层次性有极大的关系。一般来说，后续主结构层次中的零代词往往与前面主结构中的指称对象相联系，而且多以平行句法位置为主。次层次结构中的零代词，或者与其前的次层次结构中的平行位置的指称对象相联系，或者回指就近的更高层次宾语位置上的指称成分。这是主要规律。但是，我们也发现少量其他形式的出现在主次结构层次之间的长距离零形回指。还有，虽然长距离零形回指主要在主语—主语位置，但是，也有其他位置，如旁语、宾语位置；虽然夹在长距离零形回指中间的结构主要是次层次的，但是，也有其他层次，如并列结构，而且，中间结构并非总是简单结构；另外，长距离零形回指并非总是在主结构，也有次层次结构及交叉出现的长距离零形回指。这都是以往研究未曾注意的现象。

最后，平行性和层次性不是零形回指位置变化和距离变化的唯一原因，语句的主题性和语篇主题都会带来一定量的回指位置变化和长距离零形回指。

基于层次性的主要表现，我们为零形回指先行语的解析建立了层次性规则。

7. 对零形回指模糊性的揭示，并以具体的实例从多方面论述其模糊性的主观原因和客观原因。

8. 其他方面。除了以上主要发现之外，我们还发现零形回指的两种动词控制关系，一种是主语控制，一种是宾语控制。可以进行较清楚的分类，并可以采用动词分类的方法分化其中的先行语。以往有研究者讨论汉语主从句间的回指问题，涉及某些动词。本研究对语料中的两类动词进行了列表，同时，还发现了少量兼有两种功能的动词。

本研究对零形回指语料的分时段考察发现，零形回指的使用在时段上无区别，但是，零形回指的使用数量、位置和复杂度存在个人差异，似与表达者的语言风格、思维习惯、思想的开放度和复杂性等有关。

五

归纳起来，本研究的创新之处主要有以下几个方面：

1. 以往研究多次提到回指的层次性，并进行了一定的分析与应用，但没有进行系统的语料分析与标注。本研究在语料中对此做了清楚的划分和标注，并在数据库中详细地记录了这一特征。而且，本研究所涉及的层次类别也比现有的研究更为丰富。

2. 本研究对中心理论关于“中心是语篇片段中作为语篇模型一部分的语义实体”概念进行扩展，认为中心是双向的，不仅可以按照语言的线性顺序向右延伸，也可以左向扩展和控制其领域范围内的其他成员。同时，本研究认为，作为语篇模型中的一部分语义实体，中心的延伸性在其所在的语篇片段中是延续的，不因为次中心或局域中心的出现而中断，以保持对主题的延续性和回指的延续性的有效解释。

3. 本研究还在指称参照点理论的框架下，对于汉语的零形回指，

增加了一个主语之上的参照点——主题，以及主语之下和宾语之下的若干参照点。另外，本研究认为参照点不仅形成领地，更重要的是形成不同层次的领地，这对于解释包括长距离回指和内包回指在内的零形回指现象有着具体积极的作用。

4. 本研究在 Duan（2006）、许余龙等（2008）和段嫚娟等（2009，2012）的研究基础上，在语句的划分上直接反映出句法层次。同时，针对段嫚娟等（2012：231）提出的问题，采取在语料标注时添加一些零形主题的办法，以有助于某些指代词的正确消解。并基于更为丰富的语料数据，对零形回指解析参数做了调整和修改，有助于获得更好的解析效果。

5. 基于以上各种处理，以及对数据库中的各种特征与关系的分析，本研究进行了零形回指先行语的解析试验，取得了对数据库中 88.3% 先行语的解析。运行的规则包括先行语的位置、零形回指的平行位置关系和层次结构等。位置、生命性、平行性、层次作为主要特征；距离和概率是基本原则。

此外，针对以上解析中的机械性，为了增加本研究数据解析的智能化程度和考虑今后在开放语料中运用本研究的回指解析特征与规则，我们选择了 Logistic Regression 系统尝试解析本研究的零形回指的先行语。用的是从本研究的数据库中归纳出来的 124 类零形回指关系及相关语篇信息（共包括 108 例语句 1230 项有关成分），写成 txt 文本，进行了特征标注，包括 NP 的类、位置、生命性、层次性四个部分。试验获得了对零形回指解析 83.2% 的正确率。该结果不仅表明了本研究思路和方法的可行性，也证明了特征选择的可靠性。试验还表明，如果增加语料的量，会取得更好的效果，并能够训练出一个模式，在开放的语料中进行尝试。

6. 当今，随着大型计算机语料库的出现，一些学者开始尝试机器学习型回指解析模式。其主要做法是首先在封闭语料库中对语料进行规则训练，称为训练语料库，然后推衍到大型的开放语料库中，由计算机自主学习算法和进行指代解析的公开测试。其最大的优点在于它的可扩

展性。但是，我国学者关于机器学习回指解析的研究多针对汉语或英语的名词短语或代词，尚未对汉语的零形回指进行尝试。本研究作为一种新的尝试，能够为机器自学习的零形回指解析提供有益的参考。这也是本研究进一步努力的方向。