学人论丛

许宁云 著

汉语篇章零形回指的解析与生成

——一项基于语料的向心研究

上海译文出版社

• 学人论丛 •

汉语篇章零形回指的解析与生成

—— 一项基于语料的向心研究

Resolving and Generating Zero Anaphors in Chinese Discourse: A Corpus-Based Centering Approach

许宁云 著

此书得到上海市教育委员会支出预算项目（教预07-45）
上海市教育委员会科研创新项目（09YS264）
以及上海海事大学学科带头人培养对象基金的资助

上海译文出版社

图书在版编目(CIP)数据

汉语篇章零形回指的解析与生成/许宁云著. —上海:上海译文出版社,2010.4
ISBN 978-7-5327-4978-2

Ⅰ.汉… Ⅱ.许… Ⅲ.数理语言学-应用-汉语-语言教学 Ⅳ.H19

中国版本图书馆 CIP 数据核字(2010)第 009496 号

汉语篇章零形回指的解析与生成
——一项基于语料的向心研究
许宁云 著

上海世纪出版股份有限公司
译文出版社出版、发行
网址:www.yiwen.com.cn
200001 上海福建中路 193 号 www.ewen.cc
全国新华书店经销
上海颛辉印刷厂印刷

开本 890×1240 1/32 印张 11.75 插页 2 字数 216,000
2010 年 4 月第 1 版 2010 年 4 月第 1 次印刷
印数:0,001—2,100 册
ISBN 978-7-5327-4978-2/H·931
定价:23.00 元

如有质量问题,请与承印厂质量科联系。T:021-35104888

序

许宁云是我迄今所带过的最省心的博士生之一：他学习努力刻苦，专业基础扎实，科研能力突出，读博期间发表了多篇高质量的学术论文，还获得过复旦大学研究生一等奖学金，是位很有潜力和前途的青年学者。不久前的某一天，他告诉我他的第一部专著即将出版，邀我为其作序，我欣然应允。对于这部著作，我想用四个字加以评价：实在、细致。

就像他这个人很实在一样，他的这本书也很实在。记得起初他跟我说打算搞回指的认知研究，这方面他有一定的积累，而且还有几篇论文作为支撑，但他最终却选择了鲜有涉足的回指的向心研究，而且是应用于汉语，这可是个硬骨头，而他却敢啃，也愿意啃，这是他的实在之处；另外，一般人搞回指研究就只搞解析或只搞生成，而他却宁愿多费力气也要两者兼顾，这也是他的实在之处。我估摸着他之所以选择这个课题，一方面是因为该课题应用性较强，实用价值较高，比较符合他务实的特点；另一方面可能跟他当初的“转型”还有点关系。记得他读硕士时钻研过符号学，硕士论文是“语言与思维”方面的，是比较宏观、思辨性的课题，读博后我建议他要

学会做细活，于是他一直很努力，而且转型得很成功。他选择的这个课题需要极大的耐心和超常的细致才能完成，这也正迎合了他挑战自我的愿望。

该书将国外计算语言学的主流理论“向心理论”用于探讨汉语篇章零形回指问题，这是件很有意义和价值的工作，也符合语言学界反对闭门造车，鼓励运用国外先进理论解决汉语问题这一潮流。可是向心理论毕竟是基于英语语篇发展起来的语篇结构理论，要将其有效应用于汉语语篇，首先必须要对该理论的参数加以适当修改，作者在这方面做了大量细致的工作。针对汉语的特点，作者分别对语段和语篇片段进行了独创性的定义，而且还修订了代词规则和下指中心集排序规则。在修订代词规则时，作者反复对照语料，对修订好的规则一次又一次地修改，直到该规则严谨到不出现反例为止。在制定排序规则时，为了提高排序的严密程度，作者还对复合名词的排序进行了仔细研究，真可谓细致入微。以上两个规则是向心理论中能产生跨语言差异的核心参数，因而一般的向心研究就只修改这两个参数便可进行运作了，而作者为了提高解析准确率，针对汉语语篇中跨语段指称这一显著特点，制定出一套包含六个操作规则的运作模式。在设计生成模型时，作者也是不断验证，反复修改，先后共设计了三种模型。而且为了进一步提高生成准确率，作者还专辟一章的篇幅对过渡类型的计算进行了深入细致的探讨。作者的细致不仅使其实现了“拟将向心理论全面而系统地应用于汉语零形回指

的解析和生成”这一目标，而且还催生了许多的创新之处，如语段的定义、语篇片段的界定、下指中心集的排序、复合名词短语的排序、代词规则的修订、过渡类型的计算、六种下指中心操作规则的制定，以及解析模型和生成模型的建构。这些创新之处分布在该著作分量最重的四个章节之中。一篇博士论文有如此多的创新之处尚不多见，且以上九个创新之处全都佐以语料验证，具有较强的说服力。有一次我看见作者从图书馆里复印的一本民间故事选集，上面每一页都布满了密密麻麻各种手工标注以及对各种参数变项所做的统计和计算，作者所付出的心血可见一斑。

是为序。

熊学亮

2010 年 2 月

前　言

本书是在我的博士论文(《汉语篇章零形回指的解析与生成——一项基于语料的向心研究》,复旦大学,2006)基础上修改完善而成。篇章零形回指的解析与生成是汉语篇章处理的关键环节,也是当前篇章处理研究的热点之一,而就目前的国内外研究现状来看,绝大部分研究都缺乏明晰的表征手段和形式化操作模式。本书的特色是将当前国外先进的计算语言学理论"向心理论"(Centering Theory)有效应用于汉语篇章零形回指的解析与生成,提出的模型和算法经过实际语料的验证可有效解析和生成汉语篇章零形回指,它对于对外汉语教学中回指的确认和理解具有一定的启发和辅助作用,尤其对于包括机器翻译(MT)、自动文本概要(Automatic Abstracting)和信息提取(Info Extraction)在内的自然语言处理(NLP)具有一定的应用价值。

就当前的零形回指研究来说,综合看来,主要有如下四种研究范式:

范式Ⅰ:零形回指的句法研究

主要探讨零形回指在小句内的句法约束和限制。该类观

点的主要代表是 Huang(1984, 1989)和 徐烈炯(1986),他们是在管辖与约束理论(GB)框架内根据语法关系和语法功能所做的句内解释,主要区分了代词(pronominal)和变项(variables)这两大类零形回指。对于处于主语位置的代词类零形回指来说,在限定性小句中零形回指是 pro,而在非限定性小句中零形回指是 PRO,它们可以依据广义控制规则(GCR)作出与最近 NP 共指的解释;而对于被空主题或零主题(empty/zero topic)约束的变项来说,零形回指与前置的空主题共指,而非最近 NP。

范式Ⅱ:零形回指的语篇功能研究

主要从功能语篇角度探讨零形回指的选择问题。Li and Thompson(1979, 1981)指出选择零形回指的关键因素是"结合度"(conjoinability),如果两个小句之间的结合度较高,那么一般倾向于在第二个小句中使用零形回指,结果就会出现主题链(topic chain)。陈平(1984)从主题连续性(topic continuity)和语义连续性两方面对结合度原则进行了细化,并提出了可预测性条件(predictability condition)和可忽视性条件(negligibility condition)。Zhou(1995)提出了整体共指(global coreference)原则,区别了局部共指和整体共指。徐赳赳(1990, 2003)指出动词词义及语境在零形回指解析中的重要作用,并拟定了相应的解释方案。陶亮(1993, 1997)提出显现指称(emergent reference)理论,认为零形回指可通过提示语识别、指称建构和信息整合这三个过程所实现的显现指称

加以确认。Cheng(1990)和 Lee(1990, 1995)提出了找回原则(recovery principle),主要通过最近原则(recency principle)和起始原则(opening principle)来预测主句中的零形回指。游毓玲(1998)强调语篇和词汇提示词对零形回指确认的作用,提出了找回规则(recovery rule),包括主要实体原则(primary principle)、最近原则和异指原则。许余龙(1995, 2004)根据指称对象的可及性等级及其所对应的副主题、期待主题和主题栈,提出了包含零形回指在内的回指确认原则(resolution principle)。

范式Ⅲ:零形回指的语用研究

主要从语用学角度对零形回指的确认作出解释。该类观点的主要代表是黄衍(1994)。他采用 Levinson(1987, 1991)的语用框架,指出零形回指在语篇中的分布可通过 Q-原则、I-原则和 M-原则之间系统性互动加以预测,互动过程还受到异指假设、信息凸显度以及关于会话涵义的一般连贯性条件的约束。

范式Ⅳ:零形回指的认知研究

主要从认知角度探讨零形回指的确认问题。该类观点的典型代表是 Tomlin and Pu(1991),他们认为指称距离模式尤其是 Givon(1983)的主题连续性的解释力不够,指出指称对象的激活状态对于指称形式的选择至关重要。据此,他们提出了心理表征中处于不同激活状态的指称对象分别对应于不同的指称形式,认为零形回指是在指称对象处于当前激活状

态(currently activated)下启用。

综合来看,前贤主要从句法、语篇功能、语用和认知这几个方面探讨了零形回指的分布规律及确认方法,构建了不少深邃的零形回指确认与生成理论,得出了许多具有深远影响的观点。然而这些研究在很大程度上还停留于"解释"阶段,因而尚不能视为严格意义上的"解析"和"生成",而且这些研究应用性较弱,不太适用于计算机处理。

向心理论(Centering Theory)(Grosz *et al.*, 1995; Walker *et al.*, 1998, inter alia)是当前计算语言学中用于篇章回指解析与生成的主要理论模式之一。为了验证该理论中规则和限制条件的跨语言适用性,许多学者将其应用于各种语言的回指解析与生成。其中有些学者将其应用于零形回指的解析与生成(Kameyama, 1985, 1986, 1988, 1998; Walker, Iida, and Cote, 1990, 1994; Mitsuko *et al.*, 2001; Turan, 1995, 1998; Di Eugenio, 1990; Rambow, 1993; Ryu, 2001; Prasad, 2003; Prince, 1994)。

然而在国内,很少有学者将向心理论应用于汉语篇章回指的解析与生成。从发表的文献中,笔者只发现少数几篇有关向心理论的文章。其中,袁毓林(2003)只是在探讨"焦点"问题时顺便提及了向心理论,认为心理焦点类似于向心理论中的回指中心(backward-looking center);苗兴伟(2003)和刘礼进(2005a, 2005b)分别对向心理论进行了评介,但没有实质性地将其应用于汉语篇章分析;王德亮(2004)将向心理论

应用于汉语篇章零形回指的解析，但他使用的算法是采用 Iida (1998)的宏观模型(Global Model)，而且对于许多细节问题，如代词规则的修订、过渡类型的设计，语段的切分标准等都没有进行处理，因而在很大程度上影响了解释力和解析成功率。

本研究正是在这种背景下，拟将向心理论全面而系统地应用于汉语零形回指的解析和生成。

本研究旨在推导出用于解析和生成汉语篇章零形回指的计算模型，有如下两个主要研究目标：

1）构建汉语零形回指解析模型并提出相应算法，在实际语料中对算法进行验证，解析正确率须达到95%以上。

2）构建汉语零形回指生成模型，在实际语料中对模型进行验证，生成正确率须达到95%以上。

本研究的内容包括以下五个方面：

1）由于向心理论具有跨语言的特征，因此必须对其参数进行适当的修订，以便对汉语篇章进行应用分析。这些参数包括语段定义、语篇片段切分和下指中心集排序。

① 语段(utterance)是篇章组织的基本单位。基于先前的定义方法(Li, 1956; Hu, 1981; Huang & Liao, 1981; Mann and Thompson, 1987; Crystal, 1991; Zhu, 1995; Poesio, 1995; Traum & Heeman, 1996; Bussmann, 1996; Chu, 1998; Kameyama, 1998; Aronoff & Rees-Miller, 2001; Song, 2001; Xu, 2003)，本研究推导出语段的定义方法，此方法适用于汉语篇章的向心分析，因为它既符合汉语的句子特征，又便于计算机处理。

② 语篇可切分为一个个语篇片段(discourse segment),但语篇片段的切分标准和方法却尚无定论。为避免出现无回指中心(Nil)和零过渡类型(NO Cb),并基于 Cheng(1990)的主题连续段(topic continuity),本研究推导出适用于汉语语篇向心分析的语篇片段切分方法。此切分方法具有如下四个优点:1)能避免因过度切分而导致的过多的无回指中心和零过渡类型,因此可使较多的过渡类型参与决定回指形式的分布;2)能有效解决向心与宏观语篇结构的互动,以及将向心应用于拓展语篇等问题;3)可使可推导实体(inferables)作为后续语段指称的潜在指称对象;4)它尤其适用于汉语语篇的向心分析,因为在汉语语篇中,跨语段指称和跨段落指称较为普遍,而且零形代词、代词以及全称名词短语有时可以互换使用。

③ 不同的语言拥有不同的下指中心集排序方法,且决定排序的因素在向心文献中还没有完全确定。基于 Chao(1968)所提出的主题(topic)和 Li & Thompson(1979)的主题显著性(topic-prominence)以及 Chen(1984)的可及性排序(Accessibility Hierarchy),本研究制定出汉语下指中心排序方法。为验证此方法的可行性,本研究进行了语料实证,结果证明此方法是有效可行的。此外,本研究还探讨了促使实体突显的其他因素,如存现结构以及高意图性(high intentionality)和控制(control)的介入。为进一步提高此排序方法的全面性,本研究还就如何对复合名词短语进行排序进行了探讨。基于 Tetreault(2001)的观点以及 Walker and Prince(1995)、

Gordon et al.（1999）和 Hobbs（1978）的方法，本研究提出汉语中复合名词短语的排序方法。此方法较为折中，因而较适用于汉语语篇中对复合名词短语的有效排序。

2）修订向心理论中的第一向心规则（或称代词规则），并制定其他六个下指中心操作规则，即下指中心排序规则（the *Cf* Ranking Rule）、下指中心提升规则一（the *Cf* Promotion Rule Ⅰ）、下指中心提升规则二（the *Cf* Promotion Rule Ⅱ）、下指中心迁移规则（the *Cf* Transfer Rule）、下指中心删除规则（the *Cf* Deletion Rule）以及下指中心移出规则（the *Cf* Displacement Rule）。

3）以向心理论为理论框架，并基于以上修订参数和系列规则，推导出第一个计算模型，即汉语零形回指解析模型，称为 RICM（Revised Integrated Cache Model）。此模型是对 Walker（1996）集成缓存模型（Integrated Cache Model）的改进，它吸取 Walker（1996）的“反堆栈”（anti-stack）思想，并利用 Cheng（1990）和 Lee（1990，1995）的找回原则（Recovery Principles），因为词汇语义可作为寻找指称对象的理想寻找提示语（retrieval cues）。与堆栈模型（Crosz and Sidner，1986）、宏观模型（Iida，1998）和缓存模型（Walker，1996）相比，此模型的优点是既可以不求助于宏观排序列表来解析跨语段零形回指，还可以解决排序较低实体充当回指中心的问题。为验证此解析算法的有效性，本研究将算法应用于实际语料，所采用的语料是选自《中国民间故事选粹》中的 18 篇短篇故事。实验

结果表明,在语料中出现的所有零形回指中,95% 都被本算法成功解析,即解析正确率为 95%。因此本算法是有效可行的。

4) 推导出第二个计算模型,即汉语零形回指生成模型。此模型将向心过渡类型(Centering Transitions)作为回指词分布的限定条件,因为过渡类型是生成回指形式的有效方法之一(Turan, 1995; Kim, 1999; Ryu, 2000)。本研究从语料中提取出所有相关的过渡类型,并基于这些过渡类型推导出零形回指生成算法。通过语料验证,此算法的生成准确率高达 96.75%,因此此算法是有效可行的。

5) 由于过渡类型的计算对于本研究,尤其是零形回指的生成至关重要,本研究对其进行较为深入的探讨。通过结合 Laurel Fais(2004)的定义和 Strube and Hahn(1999)的分类方法,本研究设定了 18 种过渡类型。这些过渡类型在分类上更为细致,且在推理努力上能保持高度的一致性,更为重要的是,它们可以有效处理为可推导下指中心设定过渡类型的问题。此外,这些过渡类型还可用于进一步提高本研究所提出的零形回指解析算法和生成算法的有效性。

本研究拟解决的关键问题主要有以下八个方面:

1) 语段的定义问题

2) 语篇片段的切分问题

3) 下指中心集的排序问题

4) 复合名词短语的排序问题

5) 代词规则的修订以及其他相关规则(下指中心操作规

则)的制定问题

6）解析模型和生成模型的推导问题

7）解析模型和生成模型的语料验证问题

8）过渡类型的计算问题

本研究拟采取的是定性分析和定量分析相结合的研究方法。就解析模型来说,思路如下:

1）将向心理论作为总体理论框架,并针对汉语的特点,对 Walker(1996)的集成缓存模型进行适当改进,同时吸取 Walker(1996)的“反堆栈”思想,并利用 Cheng(1990)和 Lee(1990,1995)的找回原则,构拟出汉语零形回指解析模型的总体架构。

2）修改向心参数。

3）修订代词规则并制定六个下指中心操作规则。

4）基于以上的(1)、(2)、(3)推导出汉语零形回指的解析模型和算法。

5）对所选语料进行手工标注(manual annotation)。

6）将解析模型和算法应用于标注后的语料,进行验证。

7）统计结果,并对未被成功解析的案例进行分析。

8）总结

就生成模型来说,思路如下:

1）将向心理论作为总体理论框架,根据 Turan(1995)、Kim (1999)和 Ryu (2000)观点,即过渡类型是生成回指形式的有效方法之一,提出将向心过渡类型(Centering Transi-

tions)作为零形回指分布的限定条件的理论构想。

2）从已标注的语料中提取出所有相关的过渡类型。

3）基于这些过渡类型的分布规律，推导出零形回指的生成算法。

4）将算法应用于标注后的语料，进行验证。

5）统计结果，进行改进。

6）对改进后的算法进行验证。

7）统计结果，再次改进，直至确定最为有效的算法。

本研究的创新之处主要是以下四个方面：

1）在对语段进行重新定义的基础上，制定了语篇片段的切分标准。

2）制定了下指中心集的排序规则、复合名词短语的排序方法以及过渡类型的计算方法。

3）修订了代词规则，并制定了六种下指中心操作规则。

4）推导出汉语零形回指的解析模型和算法，以及零形回指的生成模型和算法。

本书是应用向心理论探讨汉语零形回指问题的一次大胆尝试，在写作过程中，曾得到我的导师熊学亮教授的悉心指导，以及褚孝泉、曲卫国两位教授的启发和修正，在此我向他们致以衷心的感谢。另外我还要感谢家人，正是由于她（他）们的鼎力支持和默默付出，才使我能够定心钻研，收获成果。由于本人水平有限，书中定有疏漏和欠妥之处，敬请专家和同仁们不吝指正。

ABBREVIATIONS

BFP	the algorithm developed by Brennen et al. (1987)
Cb	the backward-looking center
CC	the transition type of cheap continue
CCCo	the transition type of cheap continue cohesive
CCFs	the current forward-looking centers
Cf	the forward-looking center
Cp	the preferred center
CR	the transition type of cheap retain
CRCo	the transition type of expensive retain cohesive
Cs	the transition type of complete shift
CSs	the transition type of cheap smooth-shift
CT	the centering theory
CW	cue words
DRP	disjoint reference presumption
EC	the transition type of expensive continue
EPC	the existential-presentative construction
ER	the transition type of expensive retain