HANYU JUNEI ZHAOYING GUANXI JIEXI YANJIU

汉语句内照应关系解析研究

中国語文内照応関係 解析に関する研究

◎ 于素秋 著
◎ 李东光 译

中央民族大学出版社

汉语句内照应关系解析研究

中国語文内照応関係　解析に関する研究

于素秋　著
李东光　译

中央民族大学出版社

图书在版编目(CIP)数据

汉语句内照应关系解析研究/于素秋著;李东光译.
北京:中央民族大学出版社.2008.7
ISBN 978-7-81108-519-8

Ⅰ.汉… Ⅱ.①于…②李… Ⅲ.汉语—句法—研究
Ⅳ.H146.3

中国版本图书馆 CIP 数据核字(2008)第 113335 号

汉语句内照应关系解析研究

作　　者　于素秋
译　　者　李东光
责任编辑　晓　默
封面设计　汤建军
出 版 者　中央民族大学出版社
　　北京市海淀区中关村南大街 27 号　邮编:100081
　　电话:68472815(发行部)　传真:68932751(发行部)
　　68932218(总编室)　68932447(办公室)
发 行 者　全国各地新华书店
印 刷 者　北京宏伟双华印刷有限公司
开　　本　787×960(毫米)　1/16　印张:10
字　　数　100 千字
版　　次　2008 年 7 月第 1 版　2008 年 7 月第 1 次印刷
书　　号　ISBN 978-7-81108-519-8
定　　价　20.00 元

前　言

现代科学技术的飞速发展，改变了世界，也改变了人类的生活。虽说中国的计算机普及晚于那些发达国家，但是随着现代化建设的突飞猛进，如今计算机已在中国人的生活中起着非常重要的作用。尤其是被称为信息社会、信息时代的21世纪，计算机所起的作用越来越显著。无论是各行各业还是个人，人们从网络获取大量的信息，以丰富和更新自己的知识结构和能力结构，不断地激发创新意识，更圆满地完成自己所承担的任务。信息的表现形式多种多样，而作为信息载体本身，语言仍然是更为有效且更为普遍的表现形式。

近二十年来，在中文信息处理领域，基础理论研究、技术开发、应用等各个方面都取得了显著的成果。计算机从数据处理、信息处理向知识处理发展，在伴随处理能力提高的同时，对语言文字处理的要求也变得更深、更广、更严格。汉语中普遍存在由于省略以及代词的使用而产生的照应现象，如何能够准确恰当地处理汉语中的照

应关系，对于人机界面、机械翻译、文献检索等的应用至关重要。

于素秋教授的《汉语句内照应关系解析研究》，是有关机械处理汉语中照应问题的专著。作者在前人研究的基础上，经过进一步深入探讨，对汉语中的照应问题研究有所突破。关于汉语的照应问题，在此之前已经有学者从词义角度、从文章构成角度、从结构功能角度进行考察，分析了照应现象的外在的表现形式和内在结构形式。但这些研究成果都仅仅对照应现象进行了探讨，没有提出具体的解决方案。本著作在参考日语照应解析方法的同时，尝试解决汉语中的照应问题，提出了具体的解决方案。

作者基于人们对照应的理解顺序，提出了照应解析的具体方法。首先，利用代词和谓语的意义属性进行解析处理。接着按照汉语中的句子构造关系、关联词语的制约关系、句子成分间的语义关系等制定照应规则。对于代词和谓词无法解析的现象，设定其他的照应规则进行解析处理。作者将汉语中的照应现象分为四类，以此作为解析的对象。在这四类当中，又可分为由代词引发的照应和由于省略照应词引发的零照应两个大的类别。通过作者的研究，汉语中代词引发的照应

现象内在的规律更加明晰化。人称代词和指示代词只能代替与自身意义属性相同的词。即代词的使用必须依存于和自己意义属性相同的词。另一方面，由于省略而引发的零照应也必须紧紧依存于句子构成、话题、焦点等。

在我国，使用汉语的人口在90%以上，汉语也是各少数民族的通用语言。汉语具有悠久的历史，是拥有世界上使用人口最多的语言之一。因此本研究成果，无疑对汉语信息处理技术，乃至信息化建设有着重要的意义。

本书初版是用日文撰写的，现在由李东光先生将其翻译成汉文出版，能够与更多的读者见面，这不能不说是一件令人欣慰的事。

由于汉语的句子构造十分复杂，仅用代词和谓语进行照应解析还是有无法确认的时候，对于这部分照应规则的解决，作者提出了自己独创性的见解。这对读者了解当前解析汉语中句内存在的照应关系方面研究新进展，以及了解今后有待解决的新问题都有裨益，相信读者一定能够从中得到自己所需要的东西。

姚丽娟

2008年6月

目　录

图表

表：

图：

第1章 序 论

1.1 绪言

人类社会已经步入21世纪。21世纪被称为信息社会、信息时代。随着计算机信息处理能力的提高，工业、农业、商业、服务业等各个领域对信息的需要也在与日俱增。世界上的所有信息都是通过各种媒介传向各处。在任何领域中，都不能缺少信息的存在。

信息的表现形式是多种多样的。现代通信技术的信息载体是全面的、多元化的。比如数据、文字、声音、图表、影像（静止影像和活动影像）等。而作为信息载体本身，语言仍然是更为有效、且更为普遍的表现形式。我们以文字为媒介，记录信息，进行交换并加以利用。一般来说，作为传递信息的媒体，如报纸、杂志、书籍等文字媒体；电视、广播等声音媒体；还有电话、传真，以及今日极具人气的网络都是典型代表。无论是哪种媒体，最基本的形态都是我们日常所使用的语言。只有用语言，我们人类才能自如地操纵和利用各种媒体。人和人之间，用自然语言传递信息，而将人和计算机之间连接起来的则是程序语言。以往的观点认为，自然语言作为语言学，属于人文学科领域。伴随计算机的发展，产生了把自然语言作为一种自然现象来对待的这种新看法。另外，由于语言学是各门科学的基础，它不仅是哲学或

人文科学发展的突破口，也是将自然科学和思维科学结合起来的关键，因此，它也被认为是引领了各门学科发展的学科[①]。20 世纪后期的语言学，吸取了其他相关学科的方法论，逐渐发展成为一种跨学科的专业。

对自然语言理解是一项十分复杂的课题。它与认知科学这一自然科学和社会科学的跨科学研究领域有着密切关联，同时也涉及计算机科学、人工智能、数学、语言学、心理学、哲学等各个领域。

迄今为止展开的各种研究证明，计算机的自然语言处理能力已经非常成熟。随着目前的语音识别、声音合成技术的发展，以及根据固定语法利用行文解析技术而进行的机器翻译系统开发等，来自各个角度的研究都不断显示，自然语言可以成为计算机处理的对象。而且这些研究成果在信息处理技术的应用领域中将得到进一步推广，且不断向前发展。

1.2 本研究的背景

当今的世界已经被信息的浪潮包围。全世界国际互联网的使用者数量在逐年增加。面向 21 世纪国际互联网的通讯时代，语言处理以信息智能处理为对象，备受国际瞩目。1996 年，联合国大学高等研究所（UNU/IAS）提出实施通用网络语言工程（UNL）的方案，为了在国际互联网上实现自然语言处理的应用，该提案被广泛地

① 姚天顺：《自然语言理解：一种让机器懂得人类语言的研究》，清华大学出版社，第 8－9 页，1995 年版。

开始实施[①]。当时的国际互联网是一个以英语为主导语言的网络。因此，如何用汉语以外的语言，利用网络交换信息，共享资源，成为当务之急。

汉语是占中国人口90%以上的汉族等英语以外的语言，也是各少数民族的通用语言。汉语具有悠久的历史，是拥有世界上使用人口最多的语言之一。因此以汉语为目标的自然语言处理研究，尤其是应用研究也就成为最重要的课题。基于自然语言处理技术的汉语文本的自动分类、信息检索、信息提取与生成、机器翻译等智能化技术，是中国“九五”期间就开始重点研究、重点开发的项目之一。

汉语信息处理技术为中国的现代化以及信息化建设发挥越来越重要的作用。在国务院颁布的《国家中长期科学技术发展纲领》中，汉语信息处理技术被指定为高端科技之一。近十多年来，在汉语信息处理领域，基础理论研究、技术开发、应用等各个方面都取得了显著的成果。计算机从数据处理、信息处理向知识处理发展，在伴随着处理能力提高的同时，对语言文字处理的要求也变得更深、更广、更严格。自然语言处理中的汉语分词解析和语义解析、语料库建构与管理技术，以语料库为基础语言解析法、文章结构解析与生成、机器翻译、文本检索、自动摘要、人机界面等相关的研究成果相继发表[②]。

---

① 陈力为、袁琦：《语言工程》，清华大学出版社，1998年版。

② 俞士汶、朱学锋：《计算语言学文集》，北京大学出版社，1996年版。

汉语中，普遍存在由于省略代词而产生的照应现象。然而，在近几年的汉语自然语言处理研究中，有关机器处理代词的照应、省略现象这一研究还没有显著成果。如果这个问题能够解决，将会对人机界面、机器翻译、文献检索等的应用起到巨大的作用。

本研究是在参考日语的照应解析方法的同时，尝试解决汉语中的照应问题。

1.3 本研究的目的及方法

所谓照应，通常是指用代词来替代本句中出现过的、或其他句子中出现过的名词或短语现象。相当于日语中的“コソアド”系列指代的事物。另外，“コソアド”系列也有不明确指代而进行省略的现象。这也将作为零照应在同一框架中进行处理[①]。

本研究以解析汉语中的句内照应现象为目的。包括第三人称代词的照应词、名词性指示代词的照应词、指示代词“这（this）/那（that）”＋数量词＋名词的照应词以及被省略的照应词。无论哪一个，只要与前文照应，都以先行词的存在为前提条件。

照应解析的重要问题在于，如何将照应对象与指示对象做同定。由于汉语的文章结构极其复杂，解析照应关系也就变得非常困难。关于汉语的照应问题，廖秋

① 横山晶一：《自然语言处理》，アスキー出版社，第163－185页，1988年版。

忠[①]从词义角度，陈平[②]从文章构成角度，沈阳[③]从结构功能角度都进行了考察，分析其形式。这些论文都是仅仅对照应现象进行探讨，而没有提出具体的解决方案。与此相对的，日语中处理照应现象有很多的方法。最简单的方法是，将文中出现的名词进行保存，当出现代词的时候，将其处理为离代词最近的名词。仅仅如此的话，进展还是比较顺利的。但句子一旦变长或变得复杂，马上就会漏出破绽。利用语义属性匹配名词的方法，使用话题、焦点等方法进行的研究也在开展[④]。然而，经常会有下列的情况，比如用代词指代前文的全部内容，或者不仅是前面出现的名词，而是包括其在内的（上位概念）名词；反之，只是指代某部分特定名词（下位概念），这种情况下处理起来就比较困难。也有从使用范例或表层表现[⑤]、名词的指示性[⑥]、语用学及语义学的制约[⑦]，以及中心理论[⑧]等来自各个观点的研究。这些研究

---

① 廖秋忠：《现代汉语篇章中指同的表达》，载《中国语文》1986年第2期，第88－96页。

② 陈平：《汉语零形回指的话语分析》，1987年第5期，第363－378页。

③ 沈阳：《名词空位的控制性同指，照应性同指与词汇性同指》，《语言工程》，清华大学出版社，第25－30页，1997年版。

④ 村田真樹、長尾真：《用例や表層表現を用いた日本語文章の指示詞？代名詞の指示対象の推定》自然言語処理 Vol. 4 No. 1，pp. 87－109，Jan. 1997。

⑤ 村田真樹、長尾真：《用例や表層表現を用いた日本語文章の指示詞？代名詞の指示対象の推定》自然言語処理 Vol. 4 No. 1，pp. 87－109，Jan. 1997。

⑥ 村田真樹、長尾真：《名詞の指示性を利用した日本語文章における名詞の指示対象の推定》自然言語処理 Vol. 3 No. 1，pp. 67－80，Jan. 1996。

⑦ 中岩浩巳、池原悟：《語用論的？意味論的制約を用いた日本語ゼロ代名詞の文内照応解析》、自然言語処理 Vol. 3 No4. pp. 49－65，Oct. 1996。

⑧ 田村浩二、奥村学：《センター理論による日本語談話の省略解析》，情報処理学会自然言語処理研報，107－12，pp. 91－96，May. 1995。

对于汉语的照应现象而言，在某种程度上可以作为参考，但并不能直接使用。

照应现象深深依存于文章脉络，不会脱离于话题的连续性。文章中，指代对象第一次出现之后，话题就会在这种前提下继续展开。照应对象的表现形式也依存于指代对象的限定。因此，解析照应的时候，可以将照应现象的依存特征作为同定的线索和解析的参照点来加以利用。

本研究设定如下照应解析的步骤：

解析方法

基于代词对名词依存的语义属性对代词打分，在打分的基础上，与可能是先行词的候选名词进行匹配。通过将分值高的候选名词定为先行词的方法解析由代词表现的照应句。基于动词的结合价语法，利用谓词的语义属性与候选名词进行匹配，用同定先行词的方法解析由省略表现的照应句。基于设计符合汉语特点的照应规则，通过同定指示对象的方法解决用代词和谓词解决不了的照应句。

评价尺度

为了验证解析方法的有效性，使用实际生活中的汉语例句来进行模拟试验。由于没有标准的汉语语料库，因此由笔者选定受评句，主要从小说、面向小学生的百科文库、杂志、报纸、科技文章等范围中选取，做出包含句内照应现象句子的例句集。选择范围是能成为本研究的照应解析研究对象的句子、满足第 2 章中所分成的四类照应形式的句子和先行名词出现在句子前部的句子。

从上述例句集中随机抽取例句作为受评句。对随机抽取出来的句子作为照应解析对象逐一实施模拟试验。

本研究提出的方法，是从语法解析结束的阶段开始，设定照应解析所需要的信息在这一阶段已经全部被提取出来。由于汉语的语法解析目前还没有可行的系统，因此本研究的模拟试验部分将由人工完成。

1.4 研究概要

本研究由7章构成。

第1章，简述研究背景、研究目的及研究方法。

第2章，首先阐述汉语的代词功能、语义属性。基于对汉语代词所体现的照应现象的调查结果，将汉语中存在的包括照应词、省略的零照应在内的照应现象分为四类。限定解析照应关系的范围。

第3章，讨论代词产生的照应现象是如何依托于名词而存在的。利用代词的特点，提出解决代词产生的照应对象的方法。

第4章，由于谓词有控制所有成分这一特性，以及附加成分会通过谓词对所有成分产生影响这一特点，在对谓词进行探讨的基础上，提出补齐由省略照应词产生的零照应的方法。

第5章，制定照应规则库。照应规则的制定将依据汉语的句子特点、句子构成上的限制、句子间的关系等原则。用代词和谓词都解决不了的照应现象句，应用这一规则即可得到解决。能用代词或谓词解决的方法，除了汉语以外，也适用于其他语言，而照应规则是基于汉