

浅析“体词”的“动词”兼类现象*

韦向峰

中国科学院声学研究所 北京 100080

E-mail: weixiangfeng@sina.com

摘要: 本文介绍了 HNC 理论所定义的体词、动词及体词的动词兼类现象, 分别从特征语义块的复合构成、对仗性组合、“的”字否定和某些特殊类概念等方面给出了消除体词的动词兼类的一些规则, 最后给出使用这些规则的总策略和实验所得的数据。

关键词: 体词; 动词; 兼类; 对仗; 特征语义块。

A Tentative Study of Substantives Used as Verbs

XiangFeng Wei

Institute of Acoustics, Chinese Academic Science, Beijing 100080, China

E-mail: weixiangfeng@sina.com

Abstract: This paper introduces the definition of substantive, verb and the phenomenon of substantives used as verbs based on the theory of HNC. The paper puts forward some rules of disambiguation for the phenomenon respectively based on the complex structure of Eigen Chunk, the juxtapositional combination, Chinese “的” and some special concepts. Finally a general strategy and some experimental data about the rules are provided.

Keywords: Substantive; Verb; Trans-class; Juxtaposition; Eigen Chunk.

1 引言

对于已经定义好的词语分类集合 $\{A_1, A_2, \dots, A_n\}$, 会出现这样的情况: 一个词语既可以属于 A_i , 也可以属于 A_j ($1 \leq i, j \leq n$), 甚至可能同时属于多个子集合, 这种现象称为词的兼类现象。但在具体的语言环境中, 如短语、句子或篇章中, 词语所属的词类却往往是唯一的, 因此在自然语言处理中一般都需要消除兼类现象。

根据传统语言学理论, 汉语的体词包括名词、处所词、时间词、方位词、数词、量词和一部分指代词; 汉语的谓词包括动词、形容词和一部分指代词^[1]。汉语中常见的兼类现象有:

* 本文承 973 项目(G1998030506)、国家语言文字应用“十五”重点项目“语句语义标注规范研究”(ZD1105-43C)和中国科学院声学研究所创新项目的资助。

“动-名”、“动-形”、“名-动-形”等等^[1]。消除汉语兼类现象的方法有：一是基于统计概率的方法，如 CLAWS 算法和 VOLSUNGA 算法；二是基于规则的方法，根据当前词及前后词的词法、句法及语义特征来判定当前词的词性。三是统计与规则相结合的方法，把规则方法作为统计方法的补充。

根据 HNC 理论，一个概念特别是抽象概念可以从五个基本侧面来描述^{[1] P.20~22}。这五个基本侧面称为五元组 {v,g,u,z,r}，v、g、u、z 和 r 分别表示动态、静态、属性、值和效应。HNC 理论对“动词”的定义是^[2]：词语知识库的概念类别栏目中带五元组符号并以 v 为第一属性的词。“体词”的定义是：词语知识库的概念类别栏目中带五元组符号但不以 v 为第一属性的词。注意：无特别说明，下文提到的体词和动词都是指 HNC 所定义的体词和动词。

体词的动词兼类现象是指：按定义属于体词的词，在具体的语言环境（短语、句子和篇章等）中，既可以用作 v 概念的动词，又可用作非 v 概念的纯体词。如“领导”一词既可以作为纯体词使用，也可以作为 v 概念动词使用。

基于 HNC 理论的句类分析系统^[1]，它的第一个步骤就是对 l（逻辑）和 v（动态）概念进行感知识别。因此体词的动词兼类现象将会影响系统对特征语义块（v 概念）的辨认、句蜕的处理甚至句类的假设和检验。体词的动词兼类现象属于句类分析 20 项难点中的 14 号难点^[2]，本文将在文献【2】的基础上给出一些消除体词的动词兼类现象的策略和方法，同时给出具体的例句和实验数据。

2 基本的消除兼类策略

2.1 根据特征语义块的构成

文献【2】指出，消除体词的动词兼类现象，“关键举措是查看该词前面有无上装 QE 或 Eu，或后面有无下衣 hv。如果上装前或下衣后无‘的’，一律先按动词处理。否则一律先按体词处理。”

特征语义块表示式为 EK=上装+Ek+下装，其中“上装”包括 QE（基本判断逻辑修饰，如“确实、可能、应该”）、时态修饰（l6、l7）、语言逻辑修饰 la、属性修饰 Eu 和 qv（动词前加成分，如“来、去”）等，“下装”包括 hv（动词后属性修饰成分，如“着、了、过”）和 HE（对 EK 基本特性的说明）。“上装”和“下装”的出现是判定 EK 的核心 Ek（v 概念）出现的强有力的明确信息。

规则1 如果句子中在可兼动词的体词前后出现了上装或者下装，并且上装前或下装后没有“的”，那么该兼动体词一律先按动词处理。

例如例 1 中的兼动体词“代表”，其前面出现了上装 QE“始终”和“要”，并且 QE 前没有“的”，因此判定为动词，充当句子特征语义块的核心 Ek。

例 1 我们党 || 要 始终 代表 || \< 中国先进生产力 | 的发展 > 要求/。(R0104T4a1*21J)

RB1 QE QE Ek T4C

兼动体词前后有“的”字的情况将在本文的 2.1 节讨论。

2.2 根据特征语义块核心部分的复合构成

特征语义块核心的一般表示式为 $E_k = EQ + E + EH$ 。实际情况中特征语义块核心可以形成高低搭配 ($EQ+E$)、动静搭配 ($E+EH$)。在高低搭配中, EQ 是高层概念动词, 必须补充低层概念动词 E 才能构成特征语义块的核心; 在动静搭配中, E 是动词 (动态), EH 是非动词 (静态)。据此可得到规则 2 如下:

规则 2 如果句子中出现了高低搭配的 EQ (如纯 vv 类动词), 那么其后的兼动体词一律先按动词处理; 如果句子中出现了动静搭配的 E , 那么其后的兼动体词一律先按纯体词处理。

例 2 美国 || 企图 领导 || 西方国家 || [# 建立 || 所谓的世界新秩序 #]。(R4110J)

RB1 vv E RB2 $RC=\{31XY10*21J\}$

例 3 日本政府 || 开始 大 搞 教育。(T3P1*11J)

TA EQ Eu E EH

在例 2 中, 兼动体词“领导”之前有 vv 类动词“企图”, 因此“领导”为动词。在例 3 中, 兼动体词“教育”前有动静搭配的动词“搞”, 所以“教育”为纯体词。

2.3 对仗性组合准则

严格意义下的对仗性组合要求各并列成分是地位等同的, 并且在某些特性上必须相同。在汉语中, 对仗性组合标记主要有四类: 第一类是顿号“、”; 第二类是前后搭配的逻辑概念 ($lq5...h5$), 如“自...向...”、“从...到...”等; 第三类是语言逻辑概念 $l43$, 如“和”、“并”、“及”等; 第四类是语言逻辑概念 $l44$, 如“或”、“或者”等, 多数情况下也形成对仗性组合。

规则 3 如果句子中出现了仗性组合 (依据组合标记), 并且组合成分之一为纯体词, 那么组合成分中的兼动体词判定为纯体词; 组合成分之一为动词, 那么兼动体词判定为动词。

例 4 邓小平 || 十分 重视 || {发展 |[教育和科学、文化] 事业}。(D01T19*21J)

DA QE Ek $TBC=\{31XY40*21J\}$

例 5 我们 || 必须 [团结、教育和引导] || 这部分群众。(R111AXT3*22AR411J)

JK1 QE Ek JK2

在例 4 中, 出现了仗性组合标记“和”与“、”, 并且“科学”和“文化”都是纯体词, 因此判定兼动体词“教育”为纯体词。在例 5 中, 仗性组合成分“团结”根据规则 1 判定为动词, 所以根据上述规则 3 判定兼动体词“教育”为动词, “引导”也是动词。

3 其他的消除兼类策略

3.1 “的”字否定准则

规则 4 如果“的”字在兼动体词之前紧邻出现, 那么兼动体词一律先按纯体词处理。

规则5 如果兼动体词之前出现了上装,并且“的”字在上装之前紧邻出现,那么兼动体词一律先按纯体词处理。

规则6 如果兼动体词之后出现了下装,并且“的”字在下装之后紧邻出现,那么兼动体词一律按动词处理。

例6 中国人民 || 在[中国共产党的领导]下 || 空前 [团结^和组织] 起来。(R1J)(Cn)

RB Cn R

(注: Cn表示条件辅语义块)

例7 中国的民族区域自治 || 是 || 在国家的统一领导下^,

DB jD DC=Cn^{|31ElJm=3}

[{各少数民族|聚居}的地方/|{实行区域自治}^, {设立|自治机关}^, {行使|自治权}]. (jDJ)

@EIJ1=XX10*211J

@EIJ2=XY10*21J

@EIJ3=XJ

例8 各民族自治地方的自治机关 || 都是 || <中央政府|领导下|的一级地方政权>。(jDJ)

DB jD DC=<R411J>

需要注意的是,使用规则4和规则5时,如果在“的”字之前有I02概念(对象语义块指示标记),应该把兼动体词处理成动词。如例9中的“对”字,就是I02概念,它指示“军队”是“领导”的对象语义块,兼动体词“领导”充当局部的特征语义块。

例9 我们 || 必须 坚持 || <党|对军队|的绝对领导>。(X20Y80*21J)

X2B

QE X20Y80

Y8C=<|11R411J>

3.2 利用形容词的特点

一般情况下,形容词不能修饰动词,但可以修饰体词。在HNC理论中,五元组中的u概念大体上对应于形容词,五元组的概念组合gu表示兼有名词和形容词的词性,ug相当于语法学定义的“区别词”。由此可得到规则7如下:

规则7 如果兼动体词之前有u、ug、gu类概念修饰,那么该兼动体词按纯体词处理。

例如例10中的“卓越”是形容词,概念类别符号为gu,因此兼动体词“领导”可判定为纯体词。

例10 邓小平同志 || 是 || <全党全军全国各族人民|公认|的<享有|崇高威望|的卓越领导人>>。(jDJ)

DB

jD

DC=<D01J>

3.3 利用h\$g类概念

h\$g类概念使其前面紧邻的v(动态)概念g(静态)化,相当于名词的“后缀”。例如汉语中的“性”、“方式”、“事项”、“问题”等。由此可得到规则8如下:

规则8 如果兼动体词之后紧邻出现h\$g类概念,那么该兼动体词为体词按纯体词处理。

例11 教育方式 || 要符合 || 时代的要求。(jD02J)

DBC

jD0

DB02

例11中兼动体词“教育”后出现了h\$g类概念“方式”,因此按纯体词处理。

3.4 利用I9类指代概念

I9类指代概念如“这、那、某、人、这些、那个”等修饰的词不能是动词，只能是体词。因此可得到规则9如下：

规则9 如果兼动体词之前有I9类概念修饰，那么该兼动体词为体词按纯体词处理。

需要注意的是这里使用的是修饰关系，而不是相邻关系。因此有的时候需要考查两个词或短语的要素之间是否具有修饰关系。

例12 这位代表 || 匆匆 走出 || 大厅。(T2b1J)

TA Eu T2b TB1

3.5 利用句间逻辑概念

句子间的逻辑组合关系有的也具有一定的对仗性，例如“又”、“也”、“不仅...而且...”等等，利用这种对仗性要求，可以确定前后句子中兼动体词的词性。由此可得到规则10如下：

规则10 如果句群中出现了句间对仗性逻辑概念，并且某一句子中的兼动体词词类已定，则句群中处于对仗位置的兼动体词的词类与该句子相同。

例13 在中国共产党历史上|| 党||领导||中国人民||#进行了一场{把半殖民地半封建的旧中国|变成|社会主义新中国}的伟大革命#|，+十一届三中全会以来||又领导||人民||#开始了一场新的革命，{|要}把中国|由不发达的社会主义国家|变成|富强民主文明的社会主义现代化国家|#|。

(Cn)R4110J + (CnH)R4110J

例14 亚洲的发展||，[不仅]关系||亚洲各国各地区的发展和繁荣，+[也]关系||世界的发展和繁荣。

R010X*21J + R010X*21J

例13和例14中的两个兼动体词“领导”和“关系”都是动词。汉语中最基本的句型是三块句SVO(HNC表示为JK1+E+JK2)，在句群中当两个句子的JK1相同时，往往把后面句子的JK1省略，因此第二个句子在第一个语义块JK1省略的情况下，句间逻辑概念修饰的是第二个语义块E，即特征语义块(动态概念)。

4 规则的使用和效果

体词的动词兼类现象的本质就是：在大多数情况下属于体词的词，在某些特定情况下用作动词。因此，结合句类分析技术和前文给出的10条规则，这里给出处理一般的体词的动词兼类现象的处理策略：对于可兼动词的体词，先默认为体词；如果根据规则1、2、3、6判定为动词，则按动词处理；如果根据规则2、3、7、8、9可判定为纯体词，则按体词处理；如果符合规则10的条件，可按规则10处理；最后，如果句子中无其他动词，那么按动词处理。

本文对195句真实文本语料中的兼动体词“领导”、“教育”、“代表”和“关系”进行了研究，得到的结果如表1所示：

表 1 兼动体词的规则使用效果

词语	句子总数	规则覆盖率	体词句子数	动词句子数
领导	77	85.7% (66 句)	49	30
教育	35	97.1% (34 句)	32	4
代表	36	86.1% (31 句)	30	8
关系	47	97.9% (46 句)	46	1

从表 1 中可以看出,在总的处理策略指导下规则的覆盖率达 85% 以上。各个具体体词的动词兼类现象程度有所不同,“领导”一词的兼类现象最为严重,而“关系”一词在语料中的动词兼类现象很少。还有一点需要说明的是:在一个句子中有可能出现多个兼动体词,例如第一个“领导”在句子中充当体词,而第二个“领导”在句子中充当动词,所以体词句子数加上动词句子数会大于句子总数。

5 结语

本文讨论了 HNC 理论框架下体词的动词兼类现象,并给出了消除该兼类现象的策略和规则。消除体词的动词兼类现象的关键在于如何判定具体语言环境中体词是否用作了动词,即具体语言环境中对动词的辨认。这与特征语义块的感知有相似的地方,它们都是与 v 概念相关的辨认和感知。本文提出的策略和规则很大程度上都是依据某些概念的出现来判定的,并没有深入地探讨概念之间的关联性问题。只有对句子语义块之间及语义块内部的语义关系进行分析和判断,并结合相关的语法知识和世界知识,才能真正彻底地消除兼类现象。

参 考 文 献

- [1] 黄曾阳:《HNC(概念层次网络)理论》,清华大学出版社,1998年11月。
- [2] 黄曾阳:“句类分析的20项难点”,内部资料,1999年10月。
- [3] 房玉清:《实用汉语语法》,北京语言学院出版社,1992年1月。
- [4] 俞士汶等:《现代汉语语法信息词典》,清华大学出版社,1998年4月。
- [5] 冯志伟:“从汉英机器翻译看汉语句法语义分析的特点和难点”,“汉语计算与计量研讨会”,香港城市大学,1998年。
- [6] 晋耀红:“基于HNC理论的句类分析系统的设计与实现”,硕士学位论文,中国科学院声学研究所,1998年。
- [7] 苗传江:“HNC句类知识研究”,博士学位论文,中国科学院声学研究所,2001年。
- [8] 唐兴全:“HNC理论的五元组与词性”,《自然语言理解与机器翻译》P.253~258,黄昌宁、张普主编,清华大学出版社,2001年7月。