

信息处理用现代汉语词类划分研究

李竹

国家语委语用所计算语言学研究室

摘要：词类标记问题是自然语言信息处理的基本任务，为满足信息处理的实际需要，本文提出了一个用于中文信息处理的现代汉语词类划分方案，并给出了词类的分布特征，对名词、数词、动词、连词、前后接成分、语素、俗语等类别的细化问题进行了尝试。

Research on Parts-of-Speech in Chinese Information Processing

Li Zhu

Section of Computational Linguistics, Institute of Applied linguistics, State Language Commission

Abstract The Parts-of-Speech Tagging is a basic task in NLP. In this paper we provide an idea of Parts-of-Speech in Chinese Information Processing, and we also describe the Distribution features of Parts-of-Speech. In addition subdivisions of Nouns, Numerals, Verbs Conjunctions, Prefix, Suffix, Morpheme, idiom and etc. are discussed in the paper.

引言

用于现代汉语信息处理系统中的汉语词类和词类标记集有很多种设计方案。经过多年的研究，人们对信息处理系统中的词类问题，已经有了一个基本统一的认识。现行的语言信息处理系统中的词类体系，从本质上说，没有实质性的差别，但在划分词类时的具体做法又不完全一致，词类标记集的大小和使用的符号也相差很多，这给语言信息处理系统的信息交换带来了困难，现在越来越需要有一套面向信息处理的、统一的现代汉语词类和标记集。国家语委语用所计算语言学研究室承担了《信息处理用现代汉语词类及标记集规范》（国家社科“九五”重大项目《信息处理用现代汉语词汇研究》的子课题，项目编号97@yy001-4）这一研究课题。有关词类问题，在语言学界已取得了很多成果。虽然各家词类体系中，大类、小类有多有少、词类间的层级关系不尽相同，但人们划分词类的标准

基本统一，即句法功能是划分词类的主要依据。在课题的研究过程中，我们通过各种方式对国内有影响的词类及标记集做了调查，标注了一定量的语料，并在一定词集上做了归类试验。实验结果证明我们的方案是可行的。目前的词类标记集一般都区分成语、习用语和简称略语，我们认为成语和习用语都是语义内容丰富、形式稳定的固定用法，标注语料时区分成语和习用语意义不大，不如把成语和习用语合为一类（俗语），在其下，根据句法功能，分为名俗语、动俗语、形俗语、句子等细类；简称略语在语料中大量出现，我们可以认为简称略语就是词，因此把简称略语分解成名词、动词、区别词等，即不设简称略语一类，而把简称略语作为细类设在名词、动词、区别词等基本词类之下。总之我们从汉语信息处理的实际要求出发，本着好用的原则对现代汉语书面语的词类问题做了一些研究。希望各位专家给我们提出宝贵意见。在我们的研究过程中山西大学计算机系刘开瑛教授、郑家恒教授、北京大学计算语言学研究所俞士汶教授和中国人民大学胡明扬教授给了我们很大的帮助，在此表示感谢。

1. 对几个相关术语的解释

1. 1 词

为满足计算机处理真实文本的需要，文中的词包括以下几项：

(1)语言学词典中的词；(2) 俗语、简称略语和一些结构较为紧密的成分，如“总而言之”、“齐抓共管”、“勤学苦练”、“千千万万”、“说三道四”、“三头六臂”、“为了”、“除了”、“本着”、“贱卖”、“分之”等；(3)前后接成分；(4) 语素字、非语素字等等；(5) 标点符号及非汉字符号。

1. 2 基本词类

基本词类指《信息处理用现代汉语词类及标记集规范》中名词、动词、形容词等二十五个类。有关名词动词形容词等的说明见第3部分。

1. 3 细类

细类是隶属于某个词类之下具有某些特殊分布的词的类，因此细类不同于语言学词类划分中的小类，任意一个词类之下的细类之和可以小于这个词类。

2. 有关兼类问题的处理策略

当一个同形词，具有两类或两类以上词的主要句法功能时，这个词就成为一个兼类词。在考虑具体词的兼类问题时可能遇到如下的情况：

文本的领域特征影响某些词的兼类信息。

在某个特定的领域，一些词可能已经具备了另一类或几类词的主要句法特征，但在其他领域，还不具备这些特征。

我们认为词的兼类信息与系统所面对的语言文本的领域特征是相关联的。即面对不同应用领域的汉语信息处理系统中词的兼类信息可以不同。

3. 信息处理用现代汉语词类划分及相关标记

3.1 名词 (n) :

(1)名词主要用来充当主、宾语。(2)大部分名词可以作定语。(3)多数名词能直接作“有”的宾语。(4)除无量名词之外，名词能用数量词(组)修饰。极少数名词可以直接受数词修饰。

名词可设如下几个细类：

- (1) 专有名词(n1)：指称人和事物名称的名词，包括人名、地名、机构名等，专有名词一般不能用量词修饰。
- (2) 专用姓氏(n11)：指只能用作姓氏的汉字串。如“冯”、“邓”、“赵”、“吴”。
- (3) 人名(n12)。
- (4) 地名(n13)。
- (5) 机构名(n14)。
- (6) 无量名词(n2)：除专有名词之外，不受任何量词修饰的名词。如“词汇”、“女方”。

3.2 代词 (r) :

(1)代词是构成篇章结构的语法手段，在篇章中，代词具有连接篇章中句子、段落的功能。语境对理解代词起着重要的作用。(2)代词几乎能替代所有的词和词组，它所替代的成分的语法功能，就是它的语法功能。(3)代词一般不受其他词类的修饰。(4)代词不能重叠。

代词可设如下几个细类：

(1)体词性代词 (r1) : 我、你、这、那、这里、那里、这会儿、那会儿、谁、什么、哪、哪儿、哪里、这些等。

(2)谓词性代词 (r2) : 这样、这么样、那样、那么样、怎样、怎么、怎么样等。

(3)副词性代词 (r3) : 这么、那么、多、多么等。

3.3 动词 (v) :

(1) 动词常用作谓语, 不及物动词不受“很”修饰, 及物动词带宾语。(2) 绝大多数动词能带“着、了、过”。(3) 动词具有丰富的变化形式, 不少单音节动词有 VV, V 了一 V, V - V 等变化形式; 不少双音节动词有 ABAB 的变化形式。动词的各种变化形式仍然是动词。

动词可设如下几个细类:

(1) 助动词 (v1) : 助动词常用来修饰动词或形容词, 如“应该、能、可以、愿意”等, 有时能单独充当谓语, 可以放在“V 不 V”的格式中。

(2) 趋向动词 (v2) : 趋向动词常用在别的动词或形容词后边, 作动词或形容词的补语, 如: “扔过去”、“热起来”中的“过去”、“起来”。

(3) 系动词 (v3) : 表示主语和宾语间关系的动词, 如“系”、“为”、“乃”、“是”。

(4) 不及物动词 (v4) : 不能带受事宾语的一类动词, 如“躺”、“咳嗽”等。

(5) 及物动词 (v5) : 能带受事宾语的一类动词。

(6) 体宾动词 (v51) : 只能带体词性宾语的一类动词, 如“姓”、“写”、“骑”、“买”、“捆”、“驾驶”等。

(7) 存现动词 (v511) : 表示存在出现消失等现象的动词, 如“前面来了一个人”中的“来了”。

(8) 小句宾动词 (v52) : 带主谓结构作宾语的一类动词, 如“希望”、“认为”等。

(9) 双宾动词 (v53) : 可以带两个宾语的动词, 如“给”、“问”、“送”、“还”。

(10) 兼语动词 (v54) : 动词所带的受事宾语同时又是第二个动词的施事主语, 如“选举”、“让”、“派”等。

(11) 形式动词 (v55) : 如“进行”、“加以”等。这种动词后面只能跟某些双音节动词或用体词、形容词作修饰语的偏正结构。

3.4 形容词 (a) :

(1) 作定语。少数形容词不能作定语, 例: “跋扈”。(2) 作谓语。形容词作谓语时, 常有附加成分, 如程度副词、否定词“不”等, 如: “他很快活”、“他不快乐”。少数形容词不能作谓语, 例: “深切”、“酣畅”。(3) 能受否定副词“不”和程度副词如:

“很、太、非常”等修饰。(4)有些单音节形容词有AA重叠形式。有的重叠后成了副词,如:“暗暗下决心”中的“暗暗”(暗);双音节形容词有AABB, A里AB, ABB的重叠形式,重叠后都是状态词,如:“和和气气”(和气)、“马马虎虎”(马虎)、“亮堂堂”(亮堂)等。(5)形容词不受副词“正在”、“在”、“直”的修饰。

形容词可设如下几个细类:

(1)唯谓形容词(a1):只能作谓语或能作谓语并且只能用“程度副词+形容词+的”的格式作定语的形容词。

3.5 副词(d):

(1)一般只能作状语。(2)一般不能修饰名词。个别副词可以修饰名词,如“仅劳务收入就达……元”中的“仅”。

副词可设如下几个细类:

(1)关联副词(d1):在动词形容词间起连接作用的副词,如“只有……才、即使……也”中的“才”和“也”。

(2)可修饰名词性成分的副词(d2):如“就”、“是”、“仅”等。

3.6 连词(c):

(1)用于连接词、词组或句子,以表达它们之间的相互关系。(2)连词常成套出现,如“因为……所以”,“不但……而且”。(3)连词与副词可以配合使用,如“由于……就”。连词可设如下几个细类:

(1)并列连词(c1):表示并列关系的连词,其中一部分并列连词是连接词或短语的,如“和”、“及”、“与”等;另一部分并列连词连接的是句子(分句)或段落,如“则”、“并且”、“与其”等。

(2)从属连词(c2):表示从属关系的连词,这类连词连接的都是句子(分句),如“因为”、“虽然”、“只要”、“如果”、“即使”、“以便”等。

3.7 助词(u):

(1)助词常附在词或短语或句子之后。(2)助词基本上都是后置的。如:“伟大的、仔仔细细地、穿着”中的“的、地、着”。个别助词“所”是前置的,如:“所看到的、所理解的”。

助词可设如下几个细类:

(1)结构助词(u1):的、地、得。

(2)动态助词 (u2)：着、了、过。

3.8 俗语 (i)：

俗语是汉语中的固定用法，包括成语、惯用语、谚语、格言等。它们在汉语中语义内容丰富，稳定性强。

俗语可设如下几个细类：

- (1) 名俗语 (in)：如“海市蜃楼”、“井底之蛙”等。
- (2) 动俗语 (iv)：如“众口难调”、“吃老本”、“碰钉子”等。
- (3) 形俗语 (ia)：如“通情达理”。
- (4) 句子 (ij)：如“三个臭皮匠，合成一个诸葛亮”。

3.9 关联词语 (l)：

关联词语是起句段关联作用并且习惯上常在一起搭配使用的词组。如“总而言之”、“由此可见”、“一方面”等。

3.10 其他 (w)：

- (1) 标点符号 (w1)
- (2) 公式符号(w2)
- (3)其他非汉字串 (w3)，如外文字母、阿拉伯数字等。

由于篇幅的限制有关时间词(t)、处所词(s)、方位词(f)、数词(m)、量词(q)、区别词(b)、状态词(z)、介词(p)、语气词(y)、叹词(e)、拟声词(o)、前接成分(h)、后接成分(k)、语素字(g)、非语素字(x)的说明从略。

参考文献：

- [1] 俞士汶，现代汉语语法信息词典详解，清华大学出版社，1998
- [2] 胡明扬，词类问题考察，北京语言学院出版社，1996
- [3] 冯志伟，自然语言的计算机处理，上海外语教育出版社，1996
- [4] 姚天顺，自然语言理解，清华大学出版社，1995
- [5] 黄昌宁、夏莹，语言信息处理专论，1996
- [6] 朱德熙，语法讲义，商务印书馆，1984
- [7] 胡裕树，现代汉语，上海教育出版社，1984
- [8] 马庆株，汉语语义语法范畴问题，北京语言文化大学出版社，1998