“词类”三问：一个汉语词类知识学习者和使用者的反思

詹卫东

摘要：本文从三个方面对汉语的词类问题进行了扼要的反思。主要看法包括：（1）词类并不总是被看作是“集合”，有时也被看作是一个“属性”；（2）词类并不一定是针对全部词汇对象的“分类”，可以是只针对部分词语的“聚类”；（3）人们对一个词语属于某个类别的主观感觉主要是基于对比较常用的词的“意义”（及其常见用法）的把握，之所以需要进一步用“分布”来作为词类定义的标准，是试图使这种主观感觉显得更客观一些。

关键词： 词类 集合 属性 分类 聚类 分布

一 “词类”是看作“类”还是“特征”？

汉语词类问题的一个直接表现是现实中的词数量很大，而语言学理论系统假定的词类太少。现有的词类装不下所有的词。这又有两种具体情况。下面不妨以北京大学《现代汉语》教材的词类系统为例略加说明：

（1）如果严抠定义的话，“自动”没有一个类可以归进去。因为它不符合其中任何一类的定义（宋柔，2003，48-55）。

（2）“研究”既可以归入动词，也可以归入名词，因为它符合动词的定义：可以带宾语并且不受“很”修饰，如“研究学问”，也符合名词的定义：可以出现在数量成分的后面，作定中结构的中心语，如“这一项研究”。

用逻辑学上关于分类的说法来表述，第一种情况是分类系统没有达到**完备性**要求造成的。如下面图 I 所示，x 不在任何一个类中；第二种情况是分类系统没有达到**排他性**要求造成的。如下面图 II 所示，y 既在丙类中，又在丁类中。不过，在具体的词语归类操作层面，人们往往用同一种方式—— 兼类 ——去处理上述两种情况带来的困境。在图 I 中，让 x 兼属甲，乙两类，在图 II 中，让 y 兼属丙，丁两类。从逻辑的角度看，前一种兼类实际上是不存在的，不妨称之为“虚兼类”；后一种兼类是逻辑上允许[(1)]的，不妨称之为“实兼类”。

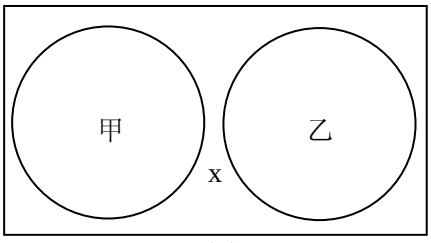

图 I

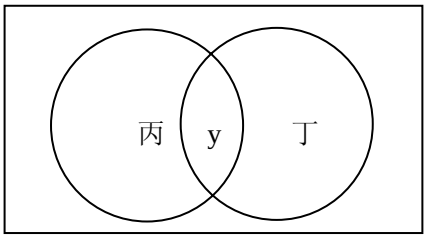

图 II

无论是“虚兼类”还是“实兼类”的情况，从逻辑上讲，解决办法都应该是增加新的类，从而达到整个分类系统既满足完备性又满足排他性的要求。但是，在语言学理论研究中，以及在面向信息处理的语言工程实践中，事实上人们往往不用增加类别，而统统用兼类的方式来解决上述问题。为什么呢？

在我看来，人们或许是“有意地”把词的“类”跟词的“功能特征”两个概念混同了。当人们说“自动、长期、临时”这些词兼属副词和区别词时，其实是把“副词”和“区别词”

当作一个“特征”概念在使用，而不是当作一个“类”的概念在使用。相当于说，“自动”具有副词性的特征（即有作状语的功能），也具有区别词性的特征（即有作定语的功能）[(2)]。“研究”的情况也是如此。当我们把“类”当作“特征”来看待时，逻辑上的所谓“完备性”和“排他性”等要求，就都被“合理地”放弃了。也就是说，原先一些严格的“类”的定义，在这时候，都降级为一个“特征”的取值。这样做，对于我们认识“自动”和“研究”等“兼类词”的用法性质（即分布特点）来说，是完全够用的。如果按照严格的逻辑要求去增加新的“类”，其实并没有增加新的“特征”，也就是并没有告诉人们关于这些“兼类词”更多的信息，除了体现表述上逻辑严密性的价值外，并无其他的实用价值。

在面对词类问题时，人们常常是想着“如何彻底根除兼类”，或者“如何尽量减少兼类词的数量”。但或许可以反思一下，目前词类系统中的“兼类”，到底是怎么产生出来的？

我想答案应该是：当人们揣着力求逻辑严密的词类划分系统去跟无比复杂的千万具体词语相见时，所谓的“词类”，有时候并没有被当作是**严格定义的“类”**，而是被看作为一些**松散（但有实用价值）的“特征”**。从而造成了并不少见的各种情形的“兼类”。从逻辑基础上讲，这样做有“逻辑漏洞”；从应用目的上讲，这样做“合理实用”。

二 “词类”应该是“分类”，还是“聚类”？

人们常常说“划分词类”。这种说法，意味着要把全体词汇中的成员当作考察和分析的对象（集合总体），人们要借助一些严格而清晰的标准，把这些成员，分别归入不同的子集，即具体的词类中。

我个人在学习汉语词类系统的过程中，越来越强烈地感觉到，事实上并没有人真的是把全部词汇当作考察和分析对象，来进行“分类”的工作。一般我们所熟悉的“词类系统”（即词类集合划分的结果），应该是以“自顶向下”（top-down）的工作方式，提出一种分类假设。汉语学界也已经给出过若干非常清晰的层级展开的词的分类树（陆俭明，2003，郭锐，2002）。从形式上看，词类的划分跟其他分类系统一样，没有什么两样：人们采用一系列标准，将一个集合划分成若干个子集。

但为何这种表达形式上看上去很清晰的分类系统，在实践中总是碰到“兼类”或者“无类可归”的问题呢？

我想原因应该涉及到两个方面：

（一）分类的实质在于分类的标准，如果标准不清楚，类就分不清楚。举简单的例子来说，就算只把词分为“实词”和“虚词”两类，也不容易分清楚。因为即使是仅仅要求给出“实词”和“虚词”的清晰的严格的标准，也是不容易做到的（参见下文第三部分中的讨论）。

（二）分类必然要考虑全体对象。如果在设计分类标准的时候其实有遗漏的对象，那当然就可能产生词类覆盖不到的情况。比如像“一起”（他们在一起玩，他们俩是一起的）这样的词，按照一般的词类定义，无法将它归入任何一个类（陆俭明，2003）。之所以存在这样“难归类的例子”，只是因为当初在设计分类系统的时候，并没有把这样的词（或者这些词的某些用法）纳入“分类的视野”。人们根据自己对一部分比较熟悉（常用）的词的语感，假设了一个分类系统，然后拿这个系统去套全体的词，出现一些套不进去的词，当属正常情况。

在汉语本体研究和一些面向计算的研究工作中，也常常有这种情况，就是人们对一类具有某种共性的词感兴趣（比如“谓宾动词”），把这些词搜集到一起，观察它们的共性。这种做法，是从词的全体集合中，抽取了某些对象，属于“聚类”的做法。在中文信息处理领域，现在也已经有一些基于语料库和机器学习方法，探讨词的自动聚类问题（陈浪舟、黄泰翼，1999）。这种“自底向上”（bottom-up），或者叫做“数据驱动”（data-driven）的工作模式，

其“聚出来的词类”远远多于人通过内省加观察部分语料的方式“分出来的词类”。

从逻辑角度讲，**“分类”的要求比较高，比较严格，是针对全体对象的操作；“聚类”的要求则相对较低，通常是针对部分对象的操作**（当然也可以针对全体对象）。

在分析一些难归类词的词性时，有人提出可以采取“排他法”和“类比法”来帮助判断（邢福义，2003）。这两种方式，可以看作是人在“聚类”时的具体办法，即观察一个词跟它相近（类似）词的距离有多近，以及跟它不相近词的距离有多远（排他），把它跟类似的词归到一类里面。而这样的工作，现在可以交给计算机程序去完成。

三 “词类”划分依据的是“意义”，还是“分布”？

现在一般现代汉语语法教科书都认为划分词类的依据是词的分布（功能）。而且有人从数学上证明了这种说法的正确性（白硕，1994）。但是，这种说法应该还有进一步反思和澄清的必要。

第一，这里面的“依据”是什么意思？

第二，这里面的“分布”是什么意思？

“依据”可以在两个背景下理解。一种背景是：词为什么可以分类，为什么分属不同的类。这在本质上是因为不同的词意义不同。正是词的意义不同，决定了词可以而且应该分类。在这种背景下，“依据”指的是原因（原由）。另一种背景是说：怎样给词分类。应该是根据可以观察到的形式特征来给词分类。在这种背景下，“依据”指的是操作标准。

“分布”是指语言成分所处的环境（位置）的总和。这里面涉及到的理论问题很复杂。其实并不是一个非常清楚的概念。比如，至少下面的问题还很难说有确定的答案：

（1）如何定义环境？也就是如何知道“环境”是相同的，还是不同的。比如：“好得很”跟“很好”中的“很”，是处在“相同的环境”还是“不同的环境”？也可以这么问，到底有多少“分布”（即多少个“不同的环境”），可以给出一张清单吗？

（2）环境的范围（尺度大小）如何确定？比如“醒”跟“困”可以出现在相同的环境中，“他醒了 — 他困了”，“他有点儿醒了 — 他有点儿困了”。但也有所处环境不同的情况，“* 他很醒 — 他很困”。那么，“醒”跟“困”的分布应该算相同呢，还是不同呢？

从现在常见的词类定义方式不难看出，人们先给出了一些“环境”—— 也就是句法结构以及相应的结构位置（比如主谓结构，主语，谓语等等），然后再根据词语是否能出现在某些环境中来划分词类。在这种工作模式中，表面上，词类可以给出清晰的定义。但实际上，困难转移到如何定义“环境”（句法结构）上了。语法书上常常列出“主谓结构、述宾结构、述补结构、定中结构、状中结构”等等所谓汉语的基本句法结构，作为确定词类分布的环境框架。可问题是，结构真的就是这么一些吗？“张三吃三个苹果”和“一天吃三个苹果”真的都是相同的所谓的“主谓结构”吗？如果**结构的完整清单**不容易定下来，那么，又如何去根据结构（分布）来定义一个完整的词类清单（满足完备性和排他性要求）呢？

对“依据”和“分布”做了上述思考后，或许对词类会有一些新的认识。

如果词类定义的起点（基础）并不像人们想像的那样坚实，那么，不妨把对词分类看作是像对万物分类一样，是基于一种先验假设（ontology），主要是对词义的语感的一种体现。这种语感因为其抽象性，有时候可能显得缺乏说服力。这时候，需要把词的“分布”（形式特征）拉来做一下证明，证明我们的分类语感是正确的。所谓考察词的“分布”来确定该词的词类归属，或许可以换一个说法：**用数量更多的语感来证明一个单一语感的正确性**。也就是说，划分词类，首先依据的是意义，然后才是分布。分布可以看作是一种证明手段。有的时候，这种证明因为其对人们某些模糊语感的颠覆而具有震撼性，比如我自己第一次学习到“战争”跟“打仗”，“突然”跟“忽然”属于不同的词类时，就产生了很强烈的震撼感。

但是，细细想来，如果一个人的语感是觉得“战争”跟“打仗”应该是一类的，他也有可能找一些“分布”依据，把它们“证明”到同一个类里去。

当你问我“桌子”的词性时，我不会去考虑它的“分布”，而是直接把它归入“名词”；当你问我“科普”的词性时，我就会犯嘀咕，我就开始在脑子里想“科普”的用例（它的分布），看看它的分布符合现有的词类“分布式定义”中的哪些定义，然后归入一类中。表面上看，对我熟悉的词，我依据“意义”归类；对我不熟悉的词，我得依据“分布”归类。不过，也还有另一种可能性，就是我脑子中原本并就没有给“科普”一个合适的类属。

我自己以往在认识词类时可能有这样的**“先入为主”的判断：每个词都该有个它所属的“类”**。这应该是从“分类”角度认识“词类”的自然产物。如果从聚类的角度来认识词类，或许会感觉到，其实有的词因为不在我们常接触的范围中，我们并没有把它跟其他一些词“聚”在一起，对它缺乏“类”的归属感。在给词作语义分类时，这种感觉尤其强烈。一般语义分类在处理那些人们熟悉的事物（比如动物、植物名称等等）时，看上去很漂亮，但如果要考虑“气泡、窟窿”这样的词该属于什么语义类，就很犯难。我的体会是，在有的场合，人们认识事物的方式是“分类”（即建立上下位层级概念体系）——因为它们在概念（意义）体系中的类属性质明确。但也存在一些对象，我们在认识它们的时候，可能并不是采用的树状分类模式——因为它们在概念（意义）体系中的类属性质不明确。这些对象，就是词语群体中的“特立独行”者。或许，它们的身上，有某种“拒绝被归类的基因”。

结语

无论是在语言教学中，还是在信息处理中，给每一个词定一个“类”，似乎都是无法回避的一件工作，于是，就出现了把一些“不太熟悉”的词硬塞入某个或多个“类”中的情形。

对此，我的看法是：

（1）在还没有其他更好办法的情况下，这样做是合理的，因为“分类”本身并不是目的，了解一个词的“用法”，才是目的。把一个词硬塞入某个“类”，其实也就是说明了这个词的 “局部特征（用法特点）”。当然，可能的代价是我们假装看不见该词的另一些局部特征或者把一些本不属于这个词的特征硬加在这个词身上[(3)]。

（2）“词类”的理想是“分类”[(4)]；“词类”的现实是“聚类”。依据“分布”（即一个词所能占据的语法位置的总和）来给词分类，是一个漂亮的理论假设，但在实践操作中，我们所知道，所能用到的，其实都只是“**一部分分布**”，而并非“**全部的分布**”。

（3）认识到上述局限性，并不意味着沮丧和放弃。相反，我认为，现行的依据词的分布来确定词的分类，是认识词的用法的可行途径。在具体操作中碰到词有兼类（其实应看作是兼有多个属性特征），是很正常的，而且兼类也是词有定类，只不过不是定在一类中，而是定在多类中罢了。

（4）人们对词类系统的认识不断加深的过程，应该是自顶向下和自底向上两个方向的思考同时进行的一个过程。前一个方向可能更多的是分类理论的思辨；而后一个方向则是一个小类一个小类地去了解词语的共性特征。等越来越多的小类词被人们发现（聚出来），汉语的词类系统应该会更好用，无论是对语言教学，还是信息处理。

附注：

(1) 如果不从分类的背景看的话，两个集合当然可以有交集。但如果是对一个集合进行划分，分出来的各个子类（子集合）之间，严格的逻辑要求也是不允许有交集。

(2) 当然，也可以把“自动”看成是两个词，其中一个的分布刚好符合区别词的要求，另一

个刚好符合副词的要求。如果持这种看法的话，从逻辑上说，也不是“兼类”，而是两个“自动”各属各类，互不相干。就好像“制服”在“他的制服很漂亮”中属名词，在“他制服了小偷”中属动词一样。

(3) 比如把“曾经”归入副词，就是假装看不见“曾经”还可以分布在像“曾经的爱情”这样的结构中；把“伟大”归入形容词，就有可能把“作状语”这个功能特征硬加给它。

(4) 数学上，对一个包含 n 个元素的集合划分非空子集，可能存在的划分个数叫 Bell 数（详见 http://en.wikipedia.org/wiki/Bell_number）。该数随 n 的增加而成组合爆炸性增长。北京大学信息科学技术学院的于江生博士因此把全部词汇集合划分子集的难度比喻为“寻找宇宙中的一粒沙”，我非常赞同。

下表列举了一些 n 值与 Bell 数，读者可由此体会子集划分的可能性之多。

n	1	2	3	4	5	6	7	8	9	10	……
Bell 数	1	2	5	15	52	203	877	4140	21147	115975	……

参考文献

白硕（1994） 词类划分的数学理论，《软件学报》第 6 期，科学出版社，北京，25-31 页。

北京大学现代汉语教研室编（1993/2004）《现代汉语》，商务印书馆，北京。

陈浪舟、黄泰翼（1999）一种新颖的词聚类算法和可变长统计语言模型，《计算机学报》第 9 期，科学出版社，北京，942-948 页。

郭锐（2002）《现代汉语词类研究》，商务印书馆，北京。

陆俭明（2003/2005）《现代汉语语法研究教程》（第三版），北京大学出版社，北京。

宋柔（2003）统计和规范中的误区，载孙茂松等编《中文信息处理的若干重要问题》，科学出版社，北京，48-55 页。

邢福义（2003）《词类辨难》，商务印书馆，北京。

（北京大学汉语语言学研究中心）