现代汉语语料库分词中的若干问题*

孙宏林
（北京语言学院语言信息处理研究所）

摘　要：本文讨论了现代汉语语料库分词中的若干理论和方法问题。文章提出了指导分词的三个原则：词是一个句法·语义范畴；词的划分是相对的；应该区分语料中的不同层次。文章从功能角度重新定义了“单用”的概念，并讨论了扩展法的局限性及其解决办法。

On Word Segmentation in Contemporary Chinese Corpus

Sun Honglin
(Language Information Processing Centre, Beijing Language Institute)

ABSTRACT: This paper poses the three principles regarding the word segmentation in contempory Chinese corpus: word is a syntactic—semantic category; word is not absolute; the differnt types of the materials in the corpus should be differentiated. The paper also dicusses the two main criteria for the word segmentation——"independent use" for simple words and the extention method for compound words.

§1 引言

最近几年，随着语料库语言学的发展，各种规模的汉语语料库正在建设之中。现在的许多语料库已不再是原始文本的简单汇集，而是经过分词、词性标注、句法标注、语义标注等各种标注从而成为各个层次上的语言知识库。在各种标注之中，分词是最基本的一步，词性标注也好，语义标注也好，句法标注也好，都是建立在分词基础之上的。从这个意义上说，分词工作是整个语料加工的基础，是整个语言信息处理的基础。因而分词的合理与否，将直接影响到以后各部分的处理，从而影响到语言信息处理的全局。

“词”作为一个语言学的术语进入汉语是本世纪初的事。几十年来，语言学家们曾经试图用各种方法来界定词，但直到今天，词和语素、短语的划界问题仍是汉语语法中的一大难题。虽然92年颁布了信息处理界分词的国家标准——《信息处理用现代汉语分词规范》（以下简称《规范》），但在采用这一规范对语料进行分词的实践中，我们发现有许多问题还得不到很好的解决。因为这一规范是面向信息处理各个领域的，而各个领域对分词的要求则不尽相同。例如，一个汉字输入系统对分词精度的要求就要比一个自然语言理解系统低得多。因此《规范》指出，“各领域可以根据其专门需求，进一步补充和细化规范的规定”。各个语料库由于其目的不同，对分词的要求也可能不同。我们所要做的“现代汉语研究语库”是一个多级加工的语料库，其目

* 本课题先后得到国家教委人文社科规划项目和国家自然科学基金重点项目（No. 69433010）的资助。文中的一些看法得益于和学友孙德金先生的多次讨论，谨致谢忱。

的是为自然语言理解和语言学研究服务的。从这一角度来看,《规范》显得过于笼统,有些规定可操作性不强。比如,对语言中占大多数的名词,《规范》的规定只有15条(其中普通名词10条,专有名词5条),这对于纷繁复杂的语言现象来说显然是少了点。因此针对我们的特定目的,我们觉得有必要从指导原则和具体方法上都需要对分词进行进一步的研究,以制定出可操作性更强、更加细化的语料库分词规范。

§2 分词的理论原则

2.1 词是一个句法·语义范畴

严格地说,"词"可以指称三个不同的概念:词汇词、语法词和书写词。书写词是指拼音文字中连在一起写的一串字母(中间可以有连字符)。由于汉语采用非拼音的汉字来记写,所以没有书写词的问题(当然如果用汉语拼音就会遇到这个问题)。语法词是从语法角度定义的语言的基本单位。现在汉语语法学界对语法词比较一致的看法是:单纯词应该能独立运用,合成词应不能扩展。词汇词是从语义角度定义的基本单位,其基本特征是具有专门的意义。如果是一个结合体,那么其整体的意义不等于部分意义的简单相加。语法词属于句法范畴,词汇词属于语义范畴。在讨论词的划界问题时,有的把这两个概念对立起来(如吕叔湘1979),有的则只考虑如何从形式上定义词,根本不考虑意义(如陆志韦1956,1957)。我们认为不应该把语法词和词汇词割裂开来。语法词的描述着重于形式和功能,词汇词的描述着重于内容和意义,它们是一个事物中互为表里的两个方面,是密不可分的。在划分词与非词的时候,必须同时考虑到这两个方面,否则就会造成不应有的困难和混乱。

从形式语法的角度看,一个语法模型包含着相互依存的两个部分:词典和规则。词典里的一个个条目就是词,它是句法分析的终极单位或最小单位。现在人们逐渐认识到,句法分析不是语言分析中一个独立的过程,它应该和语义分析同步进行,即使不同步,句法分析也不能不利用语义信息。所以词不光是句法分析的终极单位,同时它也是语义分析(语义解释)的终极单位。

由此可见,词既不是一个纯句法的概念,也不是一个纯语义的概念,而是一个句法·语义范畴上的概念。在确定词的时候,既要考虑到词的形式特征,也要考虑到词的语义特征。

2.2 词的划分不是绝对的

如上所述,词是语法模型的一部分,因此词的划分不是绝对的,它依赖于特定的语法模型。陆俭明(1988)和王洪君(1994)都指出,应该在整个句法的框架内认识分词。王洪君(1994)更明确地提出汉语词的确定只能依靠排除法,即对于一个语素组合体,我们要想确定它是不是词,首先要看它的组合符合不符合短语规则,如果符合短语规则那么它是短语,否则就是词。我们知道,语法模型都是语言学家对语言现象和规律主观认识的产物。这种认识既有客观的一面,也有主观的一面。从某种意义上说,所谓分词只是在词典和规则、词法和句法之间人为地划一条界限。这也正是在分词上许多分歧产生的原因。例如由量词重叠构成的AA结构,如"个个、条条、棵棵、回回、顿顿"等,《规范》规定一律不切分(文献4)。也就是说,它把量词重叠看成构

词现象而不是句法现象。但我们倾向于把这类结构处理为短语，因为这些量词重叠式不仅结构形式相同，而且具有相同的语法意义，“AA”都可以解释为“每一A”，如“个个”表示“每一个”，“顿顿”表示“每一顿”。既然如此，我们只要有一条句法规则就可以解释所有这些结构。当然这并不意味着把它们处理成词就一定不对，如果把这些结构都收入词典，其数量也不是无限的。只不过照这样的话，句法中就减少了一条量词重叠构成NP的规则，而在构词法中增加了一条量词重叠构成名词的规则。

如上所述，词和短语的划分在某些场合并非是绝对的，不同的人可能有不同的倾向性。由于语料库是一个公共资源，它应该尽可能适应不同用户的需求，这就要求分词的结果易于将来的变通处理。例如，对于“双音节词＋准后缀”的结构（如：大型化、简单化、制度化），有的倾向于合，有的倾向于分。如果语料库都把这类结构合起来，这样虽然满足了前一类用户的需求，但是不能满足后一类用户的需求。虽然可以自动把准后缀和它前面的成分切开，但切开后不能自动得到前一成分的词性，如“大型、简单、制度”分别属于不同的类。相反，如果把这类结构都分开则两类用户的需求都能满足，对于从合的用户只要把这类结构（词性标为“准后缀”的单位和它前面的那个单位）都合起来就可以了。

2.3 应该区分语料中的不同层次

从共时的角度看，任何语言都是一个系统，但这又是一个非常复杂的不均质的系统，对于象汉语这样历史悠久的语言来说就更是如此。我们现在处理的语料大部分是书面语的，现代汉语书面语中的成分相当驳杂，里面有口语的成分，有文言遗留的成分、有外来的成分、有方言的成分等等。不管采用什么样的分析方法，都无法使这些性质不同的成分产生一致的分析结果。所以，我们认为应该首先区分出语料中的不同层次，然后对不同层次的现象分别进行分析，这样才能得出比较可靠的结论。至于机器如何识别这些不同的层次，那是下一步要考虑的问题。

现代汉语的核心部分还是汉语口语，即一般老百姓的口头语言。从“五四”时代的“白话文运动”以来，汉语书面语逐步趋于和口语一致。但由于文言文传统的影响，一般知识分子所写的东西中或多或少地存有一些文言遗留的成分，如：

有……之嫌　值此新春佳节之际　塔身状如火箭　初获世界冠军

穿过一条小径　长达一百多年之久　啤酒素有“液体面包”之称

对于这些文言遗留成分，就必须按文言语法来对待。

书面语中还有大量的科技术语，其中相当一部分是从外语中翻译过来的，其中绝大多数是表示新概念的名词，如：

微分方程　傅立叶变换　光导纤维　乳酸杆菌　波粒二象性

对于这些科技术语，应该考虑术语意义的完整性，而不能因为音节的关系而把一个完整的术语拆开。

普通词汇的长度一般不超过四个音节（除了极少数成语和惯用语），但这些科技术语应该例外。我们不应该把这些现象跟普通词汇相提并论。英语中这些专名和术语往往由若干书写词组成，但在做句法和语义分析的时候不得不煞费苦心把它们合起来，我们何必又要先把它们切开，然后也煞费苦心地把它们合起来呢？

§3 判断单语素成词的标准

汉语的词从语素构造的角度可以分为单语素词(单纯词)和多语素词(合成词)。本节讨论单个语素的成词问题,下一节讨论多语素组合的成词问题。

3.1 对"单用"的定义

判断一个语素能否单用,实际上就是判定这个语素能否成词。一方面,它涉及到语素和词的划界问题,另一方面它涉及到词和短语的界限问题。

从理论上说,如果一个语素能够单用就是成词语素,否则就是不成词语素。这看起来很容易,其实不然,主要是大家对"单用"的理解各不相同。吕叔湘(1979)明确地区分了"单用"和"单说"的概念:"语素有能单用的,有不能单用的。能单用的又有两种:一种是能单说的,如"来";一种是不能单说但是也可以不跟别的语素组成一个词的,如"再"。这两种都是词。"对于能单说(或单独成句)的语素大家的认识是绝对没有分歧的,关键是第二种。如何判定一个不能单说的语素是否单用成为语素和词划界的焦点,也是对"单用"这个概念下定义的关键所在。吕先生对这个问题的说明在逻辑上存在循环定义的毛病,因为定义"单用"这个概念正是为了区分语素和词,而吕先生却用"词"这个概念来给"单用"下定义。

我们认为,"单用"这个概念是对语素功能属性的描述,因此给它下定义必须从功能角度着眼。我们可以这么来推理:a. 如果说一个语素能单用,就说明它是一个词;b. 一个词一定属于一个语法功能类(即词类);c. 一个语法功能类中的词一定具有该类的主要功能特征。

例如,如果是名词,它就应该:(1)能受数量词的修饰;(2)能作主宾语。如果是动词,它就应该:(1)能作谓语;(2)能受副词的修饰;(3)能在后面加上"着""了""过"等。

根据以上三个命题,我们就可以反过来推理:d. 如果一个语素 X 具有某类词的典型特征,X 一定属于这一词类;e. 如果 X 属于某一词类,那么 X 一定是词;f. 如果 X 是词,那么 X 一定是能单用的。

由命题 d 到 f 我们可以很自然地得出如下结论:

如果一个语素 X 具有某类词的典型功能特征,那么 X 一定是能单用的,否则 X 就是非单用的。

3.2 语素的同一性问题

判断一个语素能否单用,首先应该辨明语素,即解决语素的同一性问题。语素是最小的音义结合体,所以判定语素是否同一,应该从语音和语义两个方面来考虑。

在语音方面,同形异音的语素比较容易区别,如:在"咀嚼"里的"嚼"(音 jue)和"嚼口香糖"的"嚼"(音 jiao)读音不同,因此它们不具有同一性。比较容易被忽视也是比较难以处理的是轻声、儿化和后缀"子"对于语素单用性的影响。例如有的语素在口语中必须儿化或加"子"才能单用,在组合体中如果不儿化或加"子"则视为不单用,如"声"(表示"声音")在口语中只有儿化才能单用,但在"声音"里,"声"不能儿化,故不单用。

汉语的多义语素(即一字多义)是一个相当普遍的现象,有的语素甚至有十几、二十几个义项,这就需要我们在辨别语义的时候特别小心。对于语法功能上有差异的还比较容易认定,如"一本破书"的"破"(形容词)和"破坏"的"破"(动词性的)显然不同。比较困难的是那些语法功能一致,只是搭配不同的那些语素。如"当"在"担任"义上可以单用,如"当干部",但在"掌管"义上就不能单用,如"当家、当权"等。最困难的是辨别一个结构体中的一个语素到底属于诸义项中的哪一个,特别是在诸义项之间存在复杂的引申关系的时候。如"带头"的"头"到底是什么意思?《现代汉语词典》在语素义的分别方面做得比较细致准确,可以作为这一工作的基础。

3.3 语体对确定语素单用的影响

我们所分析的大多数是书面语料。如前所述,现代汉语的书面语内部是极不均匀的,它好比一个混合物,里面有现代口语的成分,也有文言的成分,有的语料口语色彩浓一些,有的书面语色彩浓一些。在考察语素的语法功能时,只能以它在现代口语中的用法为准,否则就无法确定语素的单用性,因为文言中几乎所有的语素都能成词,但是在现代口语中却并非如此。例如以下加点的语素在口语中都是不单用的:

不乏　不公　飞抵　见状　貌美　偶有　驱车　莅会　驰入　道出　我校

但是这些现象都是出现在现代汉语的书面语中的。我们认为这些现象可以看成是文言词语在现代汉语中的遗留,在特定的上下文中我们还是承认它们是词。这样会不会因为几乎所有的单音节语素都会成为词从而迫使我们放弃关于确定语素单用的所有原则呢?不会的,因为这些文言词语在现代汉语中作为词是有条件的,即必须加上语体标记。

§4 判断多语素组合体成词的标准

4.1 关于"扩展法"

扩展的意思是:如果一个结构体AB,中间可以插入一个语言单位C,使得ACB仍然成为一个结构体,并且C与A或B直接组合,且AC或CB也是结构体(中缀成分"得"、"不"除外)。按照"扩展法"的理论,如果AB能扩展则AB是短语,否则AB是词。

由于扩展法客观性强,易于操作,所以已成为语法学界公认的判别词和短语的方法,但是在实际应用中我们发现扩展法还存在着不少问题。一方面,一些明明是短语的例子却不能扩展,例如由一个名词加上一个单纯方位词构成的方位结构,如"树上、门外、食堂里、容器内、拖拉机上"等,恐怕谁都不会认为它们是词,因为这类结构可以说要多少有多少,但是它们都不能扩展。另一方面,一些明明是词的例子却能够扩展,例如"洗澡、游泳、理发、毕业"等,这些词被称为"离合词",因为大家都承认它们是词,但是可以分离,即能够扩展。由此可见,"扩展法"不能作为区分词和短语的唯一标准。

4.2 扩展法的局限性与结构类型的关系

我们通过研究发现,扩展法的运用跟结构类型很有关系。扩展法对于并列结构比较灵,不

管是双音节组合，还是三音节组合，只要内部是并列关系，基本上就可以根据扩展法来判定它是词不是。但对于偏正结构（包括定中结构和状中结构）、述补结构和动宾结构就不灵了。具体来说，扩展法对偏正结构和动补结构显得"过严"，就是说，一些明显是短语的例子却不能扩展，如果按扩展法就得归入词中，例如以上所举的方位结构。相反，它对述宾结构则显得"过松"，一些看起来是词的例子却能够扩展，如果按照扩展法就得归入短语，如"睡觉、理发、打仗"之类。

由此可以看出，汉语的偏正结构呈现出结合性，而动宾结构则呈现出分离性的特点。对于偏正结构来讲，能扩展的一定是短语，但不能扩展的未必不是短语；对于动宾结构来讲，不能扩展的一定是词，但能扩展的未必不是词。

4.3 可类推性

我们认为扩展法只能作为鉴别词和短语的一个比较方便的工具，它本身不能成为自足的方法。要区分词和短语还应当寻找最本质的东西。这种最本质的区别就是可类推性。短语的构成是可类推的，即可以用规则来描述。而词的构成是不可类推的，即不能规则来描述。如前所述，词是一种句法·语义范畴，所以这里的可类推性应包含句法和语义两个方面。句法的可类推性比较容易理解，如上面所举的量词重叠和方位结构的例子。下面举例说明语义的可类推性。

在汉语分词中有一类结构很不好处理，这就是由单音节名词加单音节名词构成的名词性结构，如："帆船、饭店、风车、海参、车闸、门票、面粉、肉菜、鸡汤"等。这类结构都符合名词加名词构成名词性短语（NP）的句法规则，但我们不能据此把它们都统统划入短语。从语义构造来看，其内部的语义关系是多种多样的，非常复杂，陆志韦（1956）曾列举了十四种语义类型，而且没有考虑因前后位置的不同而产生的新类型，实际的类型比这还要多。我们发现其中有几种语义关系类型明显具有可类推性，能产性非常强，如：

A. 材料＋制成品，如：布鞋、草鞋、米饭、米粥、纸灯、瓦房、砖墙

B. 整体＋部分，如：牛肉、羊肉、鱼头、狗腿、鸡皮、牛角、羊肝

C. 被容纳物＋容器，如：药锅、饭盒、酒瓶、花盆、水桶、汤锅、鞋盒

以上这些类型就应该看成是短语，至于象"帆船、门票、风车、饭店"之类就找不到可以简单类推的语义规则。虽然它们的构成成分都能单用，而且它们的语法构造符合句法规则，但仍应看成词。

参 考 文 献

[1] 陆志韦（1956）：《北京话单音词词汇》，科学出版社。

[2] 陆志韦（1957）：《汉语的构词法》，科学出版社。

[3] 吕叔湘（1979）：《汉语语法分析问题》，商务印书馆.

[4] 刘源等（1994）：《信息处理用现代汉语词规范及自动分词方法》，清华大学出版社.

[5] 陆俭明（1988）：《名词性"来信"是词还是短语》，《中国语文》1988 年第 5 期.

[6] 王洪君（1994）：《从字和字组看词和短语》，《中国语文》1994 年第 2 期.