

确定切词单位的某些语法因素

冯志伟 (语言文字应用研究所)

关键词 :理论词 ;形式词 ;替代测定法 ;插入测定法 ;黏附性测定法 ;功能完备性测定法

摘要 :本文指出了切词单位的确定在语言学理论方面的缺陷 ,提出了形式词的概念 ,系统地研究了确定形式词的语法因素 ,分析了基于语法因素的替代测定法、插入测定法、黏附性测定法、功能完备性测定法在确定切词单位中的效用。

Some grammatical factors to determine the segmentation element

FENG Zhiwei

Keywords: theoretical word; formal word; the substitution test approach, the insert test approach, the adhesion degree test approach, the functional integrity test approach

Abstract : In this paper, the author points out the weakness for the determination of segmentation element in linguistic theory. He proposes the conception of formal word, and then systematically analyzes the grammatical factors and approaches to determine the formal word. They are the substitution test approach, the insert test approach, the adhesion degree test approach and the functional integrity test approach.

在汉语书面文本的自动切分中,切分单位的确定是一个关键而困难的问题。之所以说这是“关键”问题,是因为如果切分单位不合理,将严重影响自动切分的效果和应用的前景;之所以说这是“困难”问题,是因为切分单位的确定常常令研究人员举棋不定,分词规范中提出的“结合紧密,使用稳定”的原则,显得过于笼统和含混,难于操作,而语言学的理论上,又划分不清语素、词和词组的界限,使得研究人员无所适从。

在语言学界,对于什么是词,如何确定语素、词和词组的界限,一直是一个长期议而不决的难题,语言学界未能提出切实可行的原则作为确定切分单位的理论依据,而且在关于语素、词和词组的基本理论方面,存在着相互矛

盾、不能自圆其说的严重缺陷。

一、理论词的概念在语言学上的缺陷

我们把语言学上的词叫做“理论词”(theoretical word),这样的理论词的概念,在语言学理论上与语素和词组划水难分,存在着严重的缺陷。

在语言学中把语素分为自由语素和黏附语素两大类。“自由语素”是活动能力很强、不仅可以与其他语素组合成词、而且还可以单独成词使用的语言中的最小的造句单位。如“地、跑、红”等都是自由语素。黏附语素的活动能力不强,不能单独成词,它们要与其他语素相组合而成词。如“机、劳、老、小、子、者、然”。

语言学中又把词分为单纯词与合成词两大类。单纯词是由一个语素构成的词。合成词是由两个或两个以上的语素构成的词。

由于单纯词只由一个语素构成,所以,这个构成单纯词的语素必定是自由语素。这样一来,“地、跑、红”等都是自由语素,它们同时又可以看成单纯词。从语素的角度看是自由语素,从词的角度看是单纯词。观察的角度不一样,名称不同,实质则是一样的。

在语素与词这两个集合之间,有一个交集(intersection),这个交集就是自由语素,如果从词的角度看,它们又可以叫做单纯词,如图1所示。

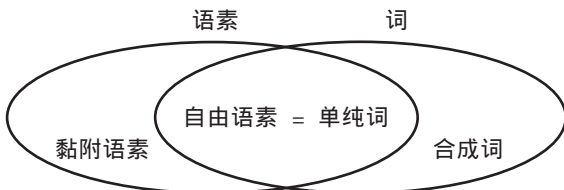


图1 语素和词之间的交集

由此可见,黏附语素和词之间的界限是可以区分清楚的,黏附语素绝不可能是词;语素和合成词之间的界限也是可以区分清楚的,合成词不可能是单个的语素。然而,在语素和词之间有一个交集,这个交集,从语素的角度看是自由语素,从词的角度看是单纯词。由于自由语素和单纯词名异而实同,导致了合成词和词组之间的界限不清。

下面我们进一步从结构方面说明这种界限不清的情况。合成词的构成方式与词组的构成方式有许多一致的地方。由语素和语素组成的合成词,构成方式主要有以下7种:

① 并列式:两个语素并列在一起组成合成词,形成一种并列关系,例如朋友、泥土、松散、东西、开关。

② 偏正式:合成词中的两个语素有主有从,后一个语素为主体,前一个语素修饰或限制后一个语素,形成一种偏正关系,例如火车、工

业、微笑、雪白、飞快。

③ 支配式:合成词中的两个语素,前一个表示动作,后一个表示动作涉及的事物,形成一种支配和被支配的关系,例如领队、司机、主席、签名、悦耳。

④ 补充式:合成词中的两个语素,前一个表示动作,后一个补充说明动作的结果,形成一种补充关系,例如提高、说明、扩大、缩小、改善。

⑤ 陈述式:合成词中的两个语素,前一个是陈述的对象,后一个是陈述的内容,形成一种陈述和被陈述的关系,例如地震、日食、祖传、法定、雪崩。

⑥ 附加式:合成词中的两个语素,只有一个表示实在的意义,另一个不表示实在的意义,只是作为一个辅助成分,附加在表示实在意义的语素之前或之后,形成前缀或后缀,例如老虎、第一、非法、桌子、椅子。

⑦ 重叠式:合成词是由单音节语素重叠而构成的,例如星星、茫茫、纷纷、巍巍、翩翩。

词组(phrase)是由词和词组合而成的。汉语词组的构成方式主要有以下7种:

① 联合结构:词组中的两个词是并列的,形成一种并列关系,例如风俗习惯、调查研究、勤劳勇敢、植树造林、轻松愉快。

② 偏正结构:词组中的两个词,前一个是修饰语,后一个是中心语,形成一种偏正关系,例如茫茫大海、劳动热情、衷心祝贺、十分灵敏、热烈欢迎。

③ 述宾结构:词组中的两个词,前一个是述语,后一个是宾语,形成一种述语对宾语的支配关系,例如热爱祖国、欣赏音乐、发射导弹、供养父母、行使职权。

④ 述补结构:词组中的两个词,前一个是述语,后一个是补语,形成一种补充关系,例如解释清楚、举起来、洗干净、讲明白、扔出去。

⑤ 主谓结构:词组中的两个词,前一个是主语,后一个是谓语,形成一种陈述关系,例如小

孩咳嗽、姑娘唱歌、天气热、月亮圆、今天星期日。

⑥附加结构：“的字结构”和“所字结构”都可以看成是附加了“的”字或“所”字的结构，形成一种附加关系，例如当兵的、掌柜的、当家的、所看到、所研究、所驱使。

由后缀“者”构成的一些长结构也可以看成附加结构的词组，例如屡教不改者、成绩不合格者、申请移民者、患心脏病者、诺贝尔奖金获得者。

⑦重叠结构：词组中的两个词，后一个词是前一个词的重叠，形成一种重叠关系，例如研究研究、练习练习、讨论讨论、总结总结、复习复习。

可以看出，汉语的合成词与词组的构成方式存在着整齐的对应，而且每种对应的结构所表示的关系是相同的。这种对应关系如表1所示：

表 1

构成方式		表示的关系
合成词	词组	
并列式	联合结构	并列关系
偏正式	偏正结构	偏正关系
支配式	述宾结构	支配关系
补充式	述补结构	补充关系
陈述式	主谓结构	陈述关系
附加式	附加结构	附加关系
重叠式	重叠结构	重叠关系

合成词的构成方式与词组的构成方式的这种一致性，使得汉语的语法规则易学易记，这对汉语的学习是有好处的，可是，这种一致性也往往导致合成词与词组的界限不甚分明，使我们难于判断一个结构究竟是合成词还是词组。如果一个结构由两个黏附语素构成，必定是合成词，不可能是词组。例如，“劳”是黏附语素，“损”也是黏附语素，它们结合而成的“劳损”必定是合成词，不可能是词组。而如果一个结构由一个黏附语素和一个自由语素构成，必定是合

成词，不可能是词组。例如，“劳”是黏附语素，“动”是自由语素，他们结合而成的“劳动”必定是合成词，不可能是词组。同样，含有前缀的“老师”、“老虎”等结构，也必定是合成词，不可能是词组，因为前缀是黏附语素。含有后缀的结构“桌子”、“作者”、“忽然”，除了后缀“者”有时可以附加在多音节结构之后构成词组之外，在一般情况下，也必定是合成词，不可能是词组，因为后缀是黏附语素。但是，如果一个结构由两个自由语素组成，问题就比较复杂了：

如果组成结构的两个自由语素都是双音节语素或多音节语素，那么，它们必定是词组，不是合成词。例如，“模糊”是双音节自由语素，“逻辑”也是双音节自由语素，由它们构成的“模糊逻辑”必定是词组，不是合成词；

如果组成结构的两个自由语素，一个是双音节语素，一个是单音节语素，那么，就不容易判定这个结构是合成词还是词组。例如，“坦克”是双音节自由语素，“车”是单音节自由语素，由它们结合而成的“坦克车”，有人认为应该是合成词，因为它表示一个整体概念。但是，“开”是单音节自由语素，“坦克”是双音节自由语素，由它们构成的“开坦克”却很难认为是一个合成词，有许多人认为它是一个述宾结构的词组。

可见，当构成结构的两个自由语素中，有一个单音节语素，就可能使合成词和词组的界限变得模糊起来，难于判定。而如果构成结构的两个自由语素都是单音节语素，那么，合成词和词组的界限就更加模糊，更加难于判定。例如，当单音节自由语素“大”与另外的单音节自由语素“会、军、陆、脑、好、红”组成“大会、大军、大陆、大脑、大好、大红”时，有人会认为前后语素之间结合得很紧密，应该是合成词。但是，当“大”与另外的单音节自由语素“鱼、河、船”组成“大鱼、大河、大船”时，可能就会有人觉得前后语素之间结合得不很紧密，它们不太像合成词，而似乎应该是词组了。

又如，表示陈述关系的结构像洗澡、鞠躬，

游泳,理发等,看来似乎是合成词,可是,有时其中的语素可以分离开来,诸如:洗澡——洗了一次澡,鞠躬——鞠了一个躬,游泳——游了一次泳,理发——理了一次发。这时,它们似乎又不像是合成词。究竟是合成词还是词组,难于判定。

我们可以对语素、词和词组的性质加以区别和比较,如表2所示。

表 2

结构单元	性 质				
	是否有意义	是否为最小单位	能否独立运用	包含语素数	包含单词数
黏附语词	+	+	-	1	0
自由语素	+	+	+	1	1
单纯词	+	+	+	1	1
合成词	+	-	+	≥2	1
词组	+	-	+	≥2	≥2

从表2中可以看出:

① 任何一个结构单元,可以根据“是否有意义”,“是否为最小单位”,“能否独立运用”,“包含语素数”,“包含单词数”等5个性质来鉴别。这5个性质之间的关系是逻辑上的合取关系(∧),也就是说,每一个结构单元,要同时根据这5个性质来鉴别,如果仅仅根据其中的某一个性质或者某几个性质,是不可能鉴别清楚的。

② 自由语素与单纯词的性质完全一样,它们在实质上是一个东西。

③ 合成词与词组的前面4个性质都相同,只有最后一个性质(即“包含单词数”)不同,合成词只包含一个单词,而词组则包含两个或两个以上的单词;可是,由于自由语素同时又可以看成单纯词,因此,当合成词由两个自由语素组成时,也可以把它看成是由两个单纯词组成的,这样,合成词就变成词组了。

可见,在语言学的理论上,合成词与词组的分界问题并没有解决。这种理论上的缺陷,必然会在汉语文本自动切分的实践中,引起种种的矛盾和困难。

二、形式词的概念

由于词是汉语句法和语义自动分析的基本单位,因此,当中文信息处理从“字处理”阶段过渡到“词处理”阶段时,必须对由连续的汉字流构成的、单词之间无空白的汉语书面文本进行自动切分。所谓“自动切分”,就是在汉语书面文本中,自动地把词切分出来,这是中文信息处理的一个难题。在汉语书面文本中把词切分出来之后,才有可能对它进行更为深入的加工和处理。因此,从自动切分的角度,我们可以把词定义为“在切分好的汉语书面文本中用空格分开的连续的汉字串(也可以是一个汉字)”。这样定义的词,叫做形式词(formal word)。

其实,在汉语书面文本长期发展的过程中,人们早就感到了这种形式词的存在,常常给词赋予某种形式,使之更加鲜明醒目。例如,《说文解字》中说,“𧈧,𧈧蛄也”。蛄蛄中的“蛄”字和“蛄”字在书面上都是“虫”字旁,以表示它们是一个词。又如,“伙伴”原来写为“火伴”,后来在“火”字上仿照“伴”字加了“亻”字旁,以表示它们是一个词。“婚姻”原来写为“昏姻”,后来在“昏”字上仿照“姻”字加了“女”字旁,以表示它们是一个词。“凤凰”原来写为“凤皇”,后来在“皇”字上仿照“凤”字加了个帽子,写成“凰”字,以表示它们是一个词。

在汉语书面文本的自动切分中,我们给词赋以特定形式的方法,就是把形式词与其前后的其他形式词用空格分开,实行切分,这就要确定“切分单位”。确定切分单位的形式因素,就是把形式词从一种形式上表现出来的形式手段。

三、确定切分单位的语法因素

由于在理论上合成词与词组的界限问题没

有彻底解决,我们在讨论如何确定切分单位的问题时,只有从实践中逐步摸索和探讨确定切分单位的形式因素。因为没有坚实的理论基础,于是只好“摸着石头过河”。在这种“摸着石头过河”的自动切分工作中,尽管我们没有能力在理论上解决合成词与词组的界限问题,但是,我们可以吸取汉语研究的一个局部性成果,来找出确定切分单位的一些形式因素。

从语言学的角度来看,确定切分单位的形式因素可从以下三个方面入手:第一是语法因素,第二是语义因素,第三是语音因素。它们是确定切分单位的主要形式因素。本文主要讨论语法因素。

在语法因素的方面,可以使用如下几种测定方法:

① 替代测定法 (the substitution test approach)

用性质相近的别的自由语素来替代待测结构中的自由语素,如果能够替代,就可以判定为词组,而不是合成词。例如,在“吃饭”中,“吃”和“饭”都是自由语素,要测定“吃饭”是合成词还是词组,先用与“吃”性质相近的自由语素“盛”、“煮”(它们都表示动作)替代“吃”,说成“盛饭”、“煮饭”;再用与“饭”性质相近的自由语素“面”、“粥”(它们都是食品),说成“吃面”、“吃粥”。由于前后两个自由语素都能被替代,就可以判定“吃饭”是词组,而不是合成词,应切分为“吃/饭”。

替代测定法是不十分可靠的,因为这种方法容易引出不合常识的错误结论。例如,“驼绒”可以被替代之后成为“驼毛”、“驼肉”、“鸭绒”、“鹅绒”,但是,从语感上显然“驼绒”不可能是一个词组,而是一个合成词。这说明用替代测定法得出的结论,与人们的语感差别太大。所以,替代测定法只能作为确定切分单位的一种参考,而不能作为可靠的依据。

② 插入测定法 (the insert test approach)

用特定的自由语素(如“的”)插入待测定的

结构中,如果能插入而不改变该结构的意义,就判定为词组,而不是合成词。

对于“形+名”的偏正结构,其切分的分合问题,就可以用插入测定法来确定。

——“形”(单音节)+“名”(单音节):

“新鞋”中插入特定的自由语素“的”,形成“新的鞋”,意义没改变,故可判定“新鞋”为词组,不是合成词,应切分为“新/鞋”。同理,“小床”应切分为“小/床”,“白花”应切分为“白/花”。

“白菜”中插入特定的自由语素“的”,形成的“白的菜”,其意义与“白菜”不同,故可判定“白菜”不是词组,而是合成词,不能切分。同理,“红花”(一种药材)、“花布”、“苦瓜”、“红茶”、“红旗”、“白旗”都不能切分。

——“形”(单音节)+“名”(双音节):

“白砂糖”中插入“的”,形成“白的砂糖”,意义没有改变,可判定“白砂糖”为词组,应切分为“白/砂糖”。同理,“甜点心”应切分为“甜/点心”,“香橡皮”应切分为“香/橡皮”。

“小媳妇”中插入特定的自由语素“的”,形成“小的媳妇”,其意义与“小媳妇”不同,可判定“小媳妇”不是词组,而是合成词,不能切分。同理,“老姑娘”、“老革命”、“高帽儿”都不能切分。

——“形”(双音节)+“名”(单音节):

“贫困县”中插入“的”,形成“贫困的县”,意义没有改变,故可判定“贫困县”为词组,应切分为“贫困/县”。同理,“富裕村”应切分为“富裕/村”,“先进队”应切分为“先进/队”。

“美丽岛”中插入“的”,形成“美丽的岛”,其意义与“美丽岛”(一个地名)不同,可判定“美丽岛”不是词组,而是合成词,不能切分。同理,“牡丹江”、“横断山”、“橄榄绿”(一种颜色)也不能切分。

插入测定法比较客观,适用范围比较广,但是,有时也会得出一些不合常识的结论。例如,北京话中可以说“鸡”,不可说“鸭”,而要说成

“鸭子”。如果我们用插入自由语素“的”的方法来测定“鸡蛋”和“鸭蛋”，“鸡蛋”可以改说成“鸡的蛋”，“鸭蛋”不可以改说成“鸭的蛋”，于是得出结论：“鸡蛋”是词组，“鸭蛋”是合成词，这种结论与人们的语感相差太大。事实上，人们普遍认为“鸡蛋”和“鸭蛋”都不是词组，而是合成词。

可见，插入测定法并不是万能的，使用时要考虑到各种复杂情况。除了插入“的”之外，还可以插入其他成分来确定切分单位。在自动切分中，可以使用插入“得”或“不”的方法来确定某些述补结构的分合问题。

某些由动词加动词或动词加形容词构成的述补结构，它们的分合常常令我们举棋不定。使用插入测定法，可以规定，双音节的述补结构中间，如果可以插入“得”或“不”，则一般应予切分。例如：

“走到”可以插入“得”或“不”（走/得/到，走/不/到），因此，“走到”应切分为“走/到”。

“安上”可以插入“得”或“不”（安/得/上，安/不/上），因此，“安上”应切分为“安/上”。

“撞上”，可以插入“得”或“不”（撞/得/上，撞/不/上），因此，“撞上”应切分为“撞/上”。

“抓住”，可以插入“得”或“不”（抓/得/住，抓/不/住），因此，“抓住”应切分为“抓/住”。

“调好”，可以插入“得”或“不”（调/得/好，调/不/好），因此，“调好”应切分为“调/好”。

“坐稳”，可以插入“得”或“不”（坐/得/稳，坐/不/稳），因此，“坐稳”应切分为“坐/稳”。

“打坏”，可以插入“得”或“不”（打/得/坏，打/不/坏），因此，“打坏”应切分为“打/坏”。

如果述补结构中间不能插入“得”或“不”，则不予切分，作为一个切分单位。例如鼓动，揭露，震动，加深，毁坏。

在有“得”或“不”的述补结构中，如果去掉“得”或“不”后，前后两个字不构成一个词的，则

不切分，作为一个切分单位。例如“来得及，来不及”，“对得起，对不起”，“说得过去，说不过去”，“了不起”。

语言学中的“词汇完整性假设”（lexical integrity hypothesis）指出，句法规则不能影响到词汇内部的任何成分。在上述的插入测定法中，把一些自由语素插入到待测的结构中，实际上是通过插入这种方法来观察句法规则能否影响到待测结构的内部，如果不能插入，就说明句法结构不能影响到待测结构的内部，从而判定待测结构是合成词而不是词组。所以插入测定法实际上就是利用“词汇完整性假设”，根据词汇的“可拆性”（separability）来区别合成词与词组的。

③黏附性测定法（the adhesion degree test approach）

测定组合成分的黏附性，如果在一个组合中出现黏附语素，则不能切分，应确定为一个切分单位。例如，含有前缀、后缀的词，都是一个切分单位，不能切分。例如，“阿哥”，“老鹰”，“非金属”，“超声波”（含前缀）；“科长”，“木头”，“学者”，“科学家”，“革命性”，“理发员”，“标准化”（含后缀）。

如果词中含有多个后缀，仍然算为一个切分单位，不能切分。例如，“物理学家”，“语言学界”，“拖拉机手”，“马克思主义者”都不切分。

但是，当某些前缀的管辖范围超出了一个单词之外，则仍然应该切分。例如，“非/国家/工作/人员”，“非/本市/注册/车辆”。

④功能完备性测定法（the functional integrity test approach）

上面在插入测定法中提到的“词汇完整性假设”说明了合成词应该具有完备的功能，而词组则不一定具有像合成词那样完备的句法功能。“词汇完整性假设”是词汇的“功能完备性”的一种反映，词汇的“功能完备性”意味着：句法规则只表现于词与词之间；单词具有完备的句法功能，而词组不能具有单词能够具有的那样

完备的句法功能。因此我们可以利用待测对象功能的完备性来判定其是否为形式词。功能完备的是合成词,功能不完备的是词组。

除了上述的“可拆性”之外,功能完备性还表现在如下几方面:

——动词的“及物性”(transitivity):“及物性”是动词的重要句法规则,动词合成词后面可以直接插入宾语,而动词词组后面则不能直接插入宾语。具体地说,

[1+1]双音节的[动+宾]式组合后能直接带宾语,可判定为合成词。例如,“得罪”后面可以带直接宾语(“得罪人”),因而可判定“得罪”是合成词。同理,可判定“抱怨(抱怨人),关心(关心他),担心(担心他),进口(进口货物),留神(留神钱包)”是合成词,不能切分。

[1+2]三音节的[动+宾]式组合后不能直接带宾语,可判定为词组。例如,“开玩笑”后面不能直接带宾语(*开玩笑人),因而可判定“开玩笑”是词组。同理,可判定“动手术(*动手术他),咬耳朵(*咬耳朵他)”是词组,应该切分。

[1+1]双音节的[动+补]式组合后能直接带宾语,可判定为合成词。例如,“想透”后面可直接带宾语(想透问题),因而可判定“想透”是合成词。同理,可判定“哭哑(哭哑嗓子),摆齐(摆齐桌子),绑好(绑好绳子),写出(写出文章)”是合成词,不能切分。

[1+2]三音节的[动+补]式组合后不能直接带宾语,可判定为词组。例如,“想透彻”后面不能直接带宾语(*想透彻问题),因而可判定“想透彻”是词组。同理,可判定“哭嘶哑(*哭嘶哑嗓子),摆整齐(*摆整齐桌子),绑结实(*绑结实绳子),写通顺(*写通顺文章),说流利(*说流利汉语)”是词组,应该切分。

——形容词前加“非常、特别”修饰:

[1+1]双音节的[“可”+动]式形容词,前面能加“非常、特别”等副词修饰,可判定为合成词。例如,“可爱”前面可加“非常、特别”修饰(非常可爱、特别可爱),因而可判定“可爱”是合

成词。同理,可判定“可恨(非常可恨、特别可恨),可悲(非常可悲、特别可悲),可耻(非常可耻、特别可耻),可疑(非常可疑、特别可疑)”是合成词,不能切分。

[1+2]三音节[“可”+动]式形容词,前面不能加“非常、特别”等副词修饰,可判定为词组。例如,“可喜爱”前面不能加“非常、特别”修饰(*非常可喜爱、*特别可喜爱),因而可判定“可喜爱”是词组,应该切分。同理,可判定“可痛恨(*非常可痛恨、*特别可痛恨),可悲哀(*非常可悲哀、*特别可悲哀),可羞耻(*非常可羞耻、*特别可羞耻),可怀疑(*非常可怀疑、*特别可怀疑)”是词组,应该切分。

由此可见,词组往往会失去单词所具有的完备的句法功能,因此我们可以使用功能完备性测定法来判定切分单位。

除了语法因素之外,确定切词单位的因素还有语义因素、语音因素等非语法因素以及其他的非语言学因素。但是,这已经超出了本文的讨论范围,我们将在另外的文章中进一步讨论这些问题。

参考文献

- [1] 冯志伟. 自然语言的计算机处理. 上海: 上海外语教育出版社, 1996
- [2] 冯志伟. 应用语言学综论. 广州: 广东教育出版社, 1999
- [3] 俞士汶等. 现代汉语语法信息词典详解. 北京: 清华大学出版社, 1998
- [4] 国家标准 GB13715. 信息处理用现代汉语分词规范. 北京: 中国标准出版社, 1992

