# 确定切词单位的某些非语法因素*

冯志伟**

（教育部语言文字应用研究所 北京 100010 北京大学计算语言学研究所 北京 100871）

摘要：本文提出了“形式词”概念，并在形式词的基础上，进一步研究了确定切词单位的非语法因素，包括语义因素、语音因素。在语义因素方面，研究了意义单纯性测定法、意义紧密性测定法、引申意义测定法、常用性测定法；在语音因素方面，研究了停顿判定法、双音节化判定法。最后，提出了视读原则、多元化原则、领域针对性原则等确定切词单位的非语言学的原则。

关键词：理论词；形式词；意义单纯性测定法；意义紧密性测定法；引申意义测定法；停顿判定法；双音节化判定法；视读原则；多元化原则；领域针对性原则

中图分类号：TP391

# The non-grammatical factors to determine the segmentation element

FENG Zhi-wei

（Ministry of Education, Institute of Applied Linguistics Beijing 100010
Peking University, Institute of Computational Linguistics Beijing 100871）

**Abstract**: Based on the conception of formal word, the author inquires into the non-grammatical factors for determination of segmentation element. They include the semantic factors, phonetic factors. He put forward the corresponding approaches: the semantic simplicity test approach, the semantic tightness test approach, the semantic derivation test approach, the pause test approach, the disyllabism test approach. Some non-linguistic principles are studied: easy-read principle, pluralism principle, domain-oriented principle.

**Keywords**: theoretical word; formal word; the semantic simplicity test approach; the semantic tightness test approach; the semantic derivation test approach; the pause test approach; the disyllabism test approach; easy-read principle; pluralism principle; domain-oriented principle

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

* 收稿日期：2001－02－26；修改稿收到日期：2001－04－18
基金项目：国家自然科学基金（69483003）；973 计划（G1998030507－4）；北大 985 计划
** 作者冯志伟，男，1939 年生，研究员，博士生导师，主要研究机器翻译、计算语言学。

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，无所适从，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。我们认为，导致这种困境的根源在于语言学中对于“词”的定义。在语言学理论上，把词定义为“语言中能够自由运用的最小单位”，这样定义的词，我们把它叫做“理论词”（theoretical word）。这样定义的理论词，在理论上存在着相互矛盾、不能自圆其说的严重缺陷，使得语素、词和词组的界限划水难分。

为了摆脱理论词的困境，我们从自动切分的角度，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）。形式词也就是切词单位，确定切词单位（即形式词）的因素是有形式规律可循的，比较容易操作，这些形式因素有语法方面的，也有非语法方面的。

在语法因素方面，确定切词单位的测定方法有替换测定法、插入测定法、黏附性测定法、词汇完整性测定法等，本文不作赘述。由于篇幅所限，本文主要讨论确定切词单位的非语法因素。这些非语法因素包括语义因素、语音因素以及一些语言学之外的原则。

## 一、确定切词单位的语义因素

在确定切词单位的语义因素的方面，主要有意义的单纯性、意义的紧密性、意义的可引申性等。据此，我们可以提出如下的判定方法：

1. 意义单纯性判定法（the semantic simplicity test approach）

根据待测结构中两个语素意义结合而成的总体意义的单纯性来判定切词单位。总体意义单纯的判定为合成词，总体意义不单纯的判定为词组。

例如，“城市”的总体意义单纯，是合成词，是一个切分单位；“夫妻”的总体意义不单纯，它的意义等于“夫”与“妻”的意义的总和，是词组，应切分为“夫/妻”。

“长短”这个结构有歧义。当它的意义表示一个人的优缺点时（“不要议论别人的长短”），意义单纯，是合成词，作为一个切分单位；当它的意义表示“长”和“短”时，这个意义等于“长”和“短”的总和，意义不单纯，是词组，应切分为“长/短”。

“动（单音节）＋名（双音节）”结构是有歧义的，当它是偏正关系时，只表示一种事物，意义比较单纯，不应切分；当它是述宾关系时，涉及到行为及其对象，意义不单纯，应该切分。例如，

我/喜欢/吃/烤白薯。（“烤白薯”不切分）

我们/来/烤/白薯/吃。（“烤/白薯”切分）

“介（单音节）＋名（单音节）”的结构也有歧义，当它表示一个事物时，意义单纯，不能切分；当它是介宾结构时，涉及到行为的对象，意义不单纯，应该切分。例如，

这/个/把手/是/木制/的。（“把手”不切分）

把/手/抬/起来。（“把/手”切分）

2. 意义紧密性判定法（the semantic tightness test approach）

根据待测结构中两个或诸个语素意义结合的紧密性来判定，意义紧密的判定为合成词，不切分，意义松懈的判定为词组，切分。

例如，“爱国”中的两个自由语素“爱”与“国”中间不能插入别的成分，意义结合得很紧密，判定为合成词，不切分。“读书”中的两个自由语素“读”和“书”之间可以插入别的成分：“读了一本书”，意义联系松懈，判定为词组，应切分为“读/书”。

国名具有唯一性，其组成成分的意义结合紧密，是一个切分单位，不应切分。例如，“中华人民共和国”，“美利坚合众国”，“德意志联邦共和国”，都不切分。

菜谱名中的各个成分，如切分后意义相差甚远，说明其意义结合紧密，则不切分。例如，“宫保肉丁”，“木樨肉”，“红烧肉”，“松鼠鳜鱼”，都不切分。但是，如果菜谱名的意义是它的各个成分的意义的简单组合，意义结合不紧密，则切分。例如，“鸡蛋/汤”，“肉丝/面”，“芝麻/糊”。

缩写词中诸成分结合紧密，也不切分。例如，“四化”，“水电”，“石化”，“环保”，“科技”，“奥运会”，“工农业”，“中西方”，“港澳台”，“教科文”，“爱委会”，“零部件”，“离退休”，“农林牧副渔”。但是，当在有顿号隔开时，则切分。例如，“港/、/澳/、/台/同胞”。

四字成语和习惯用语，各成分意义结合紧密，难以拆开，不切分。例如，“胸有成竹”，“一衣带水”，“匹夫有则”，“众所周知”，“春夏秋冬”，“充其量”，“由此可见”，“喝西北风”，“闲人免进”。

超过四个字的成语和惯用语，各成分意义结合紧密，也不切分。例如，“一年之计在于春”，“不管三七二十一”。但是，当有标点符号隔开时，则切分。例如，“人心/齐/，/泰山/移”。

3.引申意义判定法（the semantic derivation test approach）

根据待测结构的意义是否为引申意义来判定，是引申意义的判定为合成词，而保持本义的就可判定为词组。

例如，“吃饭”的本意是进餐，判定为词组，切分为“吃/饭”；但是，在句子“靠自己的劳动吃饭”中，“吃饭”的意义引申为“生存”，就判定为合成词，不切分。

同样地，“吃醋”的本义是“喝醋”，应判定为词组，切分为“吃/醋”；但是，当引申为“产生嫉妒情绪”时，就判定为合成词，不切分。

又如，“领”与“袖”两个名素构成的并列式名词“领袖”，表示“带头人物”，其含义与名素“领”与名素“袖”的含义完全不同，是“领”与“袖”含义的很远的引申，应判定为合成词，不切分。

在“妇女能顶半边天”中的“半边天”（指新社会的妇女），“他真小气，像个铁公鸡”中的“铁公鸡”（比喻一毛不拔），“银行的工作是铁饭碗”中的“铁饭碗”（比喻非常稳固的职位），“他在那里泡蘑菇”中的“泡蘑菇”（比喻故意纠缠，拖延时间），都具有引申意义，不切分。

4.常用性判定法（the frequency test approach）

根据待测结构的常用性来判定，常用的判定为合成词，算一个切分单位，不常用的判定为词组，切分。

“名词（单音节）+方位词（单音节）”的方位词组，一般应该切分。例如，“饭/前”，“树/上”，“包/里”，“床/下”。但是，某些这样的方位结构使用频度很高，事实上已经转化成处所词或时间词，不应切分。例如，“桌上”，“胸前”，“身上”，“晚上”，“午后”，“国外”。

“分之”是常见的表达分数的词语，不切分。

一些常见的并且已经收入词典中的书籍名、报刊名，也不切分。例如，“红楼梦”，“西游记”，“水浒传”，“儒林外史”，“人民日报”，“光明日报”。

## 二、确定切词单位的语音因素

在语音因素的方面，主要有停顿、双音节等，据此，我们提出了如下的判定方法：

1.停顿判定法（the pause test approach）

在一些包含多个汉字的词组中，构成词组的自由语素间常有停顿，可作为切分的参考。

例如，“全国信息技术标准化委员会”这个结构中的停顿情况是：“全国·信息·技术·标准化

·委员会”，语素之间有停顿，判定为词组，切分为“全国/信息/技术/标准化/委员会”。

2. 双音节化判定法（disyllabism test approach）

汉语的单词有双音节化（disyllabism）的倾向。双音节化导致音节之间出现两种相反的现象：一种是相吸，另一种是相拒，汉语双音节化的基本规律是：“单单相吸”，“双双相拒”，“吸单拒双”。

（1）“单单相吸”：所谓“单单相吸”，是指两个单音节的自由语素相吸而连结成一个合成词，不切分。例如，“人”和“民”相吸而连结成合成词“人民”，不切分；“香”和“烟”相吸而连结成合成词“香烟”，不切分。

单音节的区别词和单音节名词构成的组合，单单相吸而不切分。例如，“雄鸡”，“男人”。

单音节代词“本、每、各、诸”后接单音节名词时，单单相吸而不切分。例如，“本社”，“每人”，“各位”，“诸位”。但是，当它们后接双音节名词时，就排斥双音节名词而切分为两个单位，表现出一种“吸单拒双”的倾向。例如，“本/公司”，“各/部门”。

单音节名词的重叠式，单单相吸而不切分。例如，“人人”，“家家”。

单音节动词重叠式，单单相吸而不切分。例如，“走走”，“看看”。

单音节形容词重叠式，单单相吸而不切分。例如，“红红”，“久久”。

单音节量词重叠式，单单相吸而不切分。例如，“件件”，“个个”。

单音节副词重叠式，单单相吸而不切分。例如，“常常”，“仅仅”。

（2）“双双相拒”：所谓“双双相拒”，是指两组双音节结构往往有相拒的倾向而分写为词组。例如，“讨论”是一个双音节结构的合成词，它的 ABAB 型的重叠形式是“讨论讨论”由两个双音节结构组成，这两个双音节结构彼此相拒，应分写为词组，分写为“讨论/讨论”。

双音节形容词的 ABAB 型重叠式，双双相拒而切分为“AB/AB”。例如，“高兴/高兴”。“热闹/热闹”。

双音节状态词的 ABAB 型重叠式，双双相拒而切分为“AB/AB”。例如，“碧绿/碧绿”，“雪白/雪白”，“浅黄/浅黄”。

双音节数词的 ABAB 型重叠式，双双相拒而切分为“AB/AB”。例如，“许多/许多”，“很多/很多”。

双音节数量词的 ABAB 型重叠式，双双相拒而切分为“AB/AB”。例如，“一个/一个”。

但是，双音节动词的 AABB 型重叠式，由于 AA 和 BB 切分后意义发生变化，算一个切分单位。例如，“勾勾搭搭”，“比比划划”。

双音节形容词的 AABB 型重叠式，由于 AA 和 BB 切分后意义发生变化，算一个切分单位。例如，“高高兴兴”，“热热闹闹”。

双音节名词的 AABB 型重叠式，由于 AA 和 BB 切分后意义发生变化，算一个切分单位。例如，“山山水水”，“方方面面”。

双音节数词的 AABB 型重叠式，由于 AA 和 BB 切分后意义发生变化，算一个切分单位。例如，“多多少少”，“许许多多”。

（3）“吸单拒双”：所谓“吸单拒双”，是指当双音节结构与单音节结构相遇时，这个双音节结构能够把单音节结构吸引过来而形成合成词，而当双音节结构与另一个双音节结构相遇时，这个双音节结构往往会排斥另一个双音节结构而新形成词组。例如，“图书”是个双音节结构的合成词，当它与单音节语素“馆”相遇时，能够把这个单音节语素“馆”吸引过来，形成“图书馆”这个合成词，是一个切分单位；但是，当它与双音节结构“目录”相遇时，却排斥这个双音节结

构，而形成一个词组“图书目录”，应分写为“图书/目录”两个切分单位。有时，三音节结构也会把它后面的单音节语素吸引过来而形成合成词，也具有“吸单拒双”的规律。例如，“天文学”这个三音节结构，与单音节语素“书”相遇时，会把这个单音节语素吸引过来而形成合成词“天文学书”，是一个切分单位；而当三音节词“天文学”后接双音节词“理论”时，则表现出排斥的倾向，应该切分为“天文学/理论”。如前所述，单音节代词后接名词时，也表现出这种“吸单拒双”的倾向。所以，“吸单拒双”的倾向不仅是双音节词的特性，而且三音节词和单音节词也表现出这种“吸单拒双”的倾向。这也许是汉语书面文本自动切分在语音方面的一个普遍规律。

这里需要注意的是，双音节词“吸单拒双”中的“吸单”，是指前面的双音节词吸引它后面的单音节词，是“前双吸后单”；单音节词“吸单拒双”中的“拒双”，是指前面的单音节词拒绝后面的双音节词，是“前单拒后双”。虽然两者都是双音节词与单音节词相遇，但由于前后位置不同，吸引或拒绝的情况也就大不一样。所以我们不能笼统地说双音节词与单音节词之间的相吸或者相斥，而应该注意它们前后位置的不同对于相吸相斥规律的影响。

这种“吸单拒双”的倾向，在地名的切分中也表现出来。

当地名后有“省”、“市”、“县”、“区”、“乡”、“镇”、“村”、“旗”、“州”、“都”、“府”、“道”等单音节的行政区划名称时，马上把单音节名称吸过来，形成单独的切分单位。例如，“四川省”，“天津市”，“正定县”，“海淀区”，“东升乡”，“双桥镇”，“南化村”，“俄亥俄州”，“东京都”，“大阪府”，“北海道”，“长野县”，“开封府”，“宣城县”。

当地名后的行政区划名称为双音节时，则排斥双音节的名称，形成两个切分单位。例如，“芜湖/专区”，“深圳/特区”，“香港/特区”，“华盛顿/特区”。

当地名后有表示地形地貌的单音节的普通名词“江、河、山、洋、海、岛、峰、湖”时，则相吸而形成单独的切分单位，不予切分。例如，“鸭绿江”，“亚马逊河”，“喜马拉雅山”，“大西洋”，“地中海”，“塞普路斯岛”，“珠穆朗玛峰”，“洞庭湖”。

当地名后有表示地形地貌的双音节的普通名词时，则相拒而成为两个切分单位，例如，“台湾/海峡”，“华北/平原”，“帕米尔/高原”，“青藏/高原”，“南沙/群岛”。

当地名后有表示自然区划的单音节的“村，街，路，道，堡，巷，里，庄”等普通名词时，则相吸而形成单独的切分单位，不予切分。例如，“中关村”，“长安街”，“学院路”，“马场道”，“吴家堡”，“北菜市巷”，“三元里”，“庞各庄”。

当地名后有表示自然区划的双音节普通名词时，则相拒而切分为两个切分单位。例如，“米市/大街”，蒋家/胡同“，”陶然亭/公园”。

这种“吸单拒双”的倾向，在民族名称、语言文字名称的切分中也表现出来。

民族名称后面的单音节词“族”一律不切分，整个民族作为一个切分单位。例如，“蒙古族”，“朝鲜族”，“哈萨克族”，“维吾尔族”。但是，如果后面接双音节的词“民族”，则切分。例如，“蒙古/民族”，“朝鲜/民族”，“中华/民族”。

语言文字名称后面的单音节词“语”和“文”一律不切分，整个语言文字名称作为一个切分单位。例如，“蒙古语”，“维吾尔语”，“斯拉夫语”，“日耳曼语”，“蒙古文”。但是，当后面接双音节词“语言”和“文字”时，则切分为两个单位。例如，“印欧/语言”，“吐火罗/文字”。

由此可见，“双音节化判定法”是确定汉语文本自动切分的切分单位的一个非常重要而且行之有效的方法。这种“双音节化”反映了汉语韵律系统(Chinese prosodic system)的特征，汉语韵律的基本形式是双音节，这种双音节，就是汉语韵律的音步(prosodic step)，音步是汉语韵律的单位，也是汉语书面文本的切分单位，只要满足音步，就可以判定为词。如果某一字符串

等于韵律单位，那么，该字符串就被韵律“压”成词；如果某一字符串大于韵律单位，那么，该字符串就往往会被韵律“押”为词组。在现代汉语中，存在着“韵律压词，韵律押语”（“语”就是短语，也就是词组）的规律。

我们在讨论语法因素时常常考虑“双音节化”的规律对于语法因素的制约作用。看来，在确定切分单位的各种因素中，“双音节化”的韵律起着举足轻重的关键作用。韵律是我们在确定切分单位时首先应当考虑的因素。以韵律因素为主，辅之以语法因素和语义因素，可能是确定切分单位的有效办法。

当然，确定了韵律因素为主，并不意味着忽视其他因素。事实上，在汉语书面文本的自动切分研究中，我们不能只采用一种方法来确定切分单位。比较切合实际的办法是综合运用上述各种方法来进行判断，各种方法之间应该相互补充，相互校正。

## 三、确定切分单位的非语言学原则

形式词是理论词在汉语文本自动切分中的进一步拓广，它的外延比理论词更为广泛，因此，除了如上所述的语言学上的语法、语义、语音等三个形式因素之外，还应该考虑以下的非语言学原则。

1. 视读原则（easy-read principle）

切分以后的汉语书面文本是一种视读实体，最好应该满足视觉形象方面的要求。

但是，根据认知心理学的研究，人对信息的感知广度以 7 左右为限。我们数苹果，五个五个地数比较容易，十个十个地数就很难。据说象棋大师对于不成布局的、阵势较乱的棋盘，粗看一下之后，至多也只能记住 7 个棋子的位置。根据这样的原理，切分出来的形式词中所含的汉字数目以不多于 7 个为佳，要尽量使汉字数目超过 7 个的形式词不要太多。例如，“同步稳相回旋加速器”含有 9 个汉字，如果连写为一串长龙不便阅读，根据视读原则，可切分为“同步/稳相/回旋/加速器”4 个形式词。

一些长的地名和机构名如果不切分也不便于视读，应该切分。例如，“河北省/正定县/西平乐乡/南化村”，“云南省/昆明市/五华区/大观街”，“国家/语言/文字/工作/委员会/语言/文字/应用/研究所/计算/语言学/研究室”。

新闻报道中的活动名称不宜太长，对于那些太长的活动名称，也应该切分开来，以便视读。例如，“庆/回归/公益/千万/行”，“第三/次/横田/基地/噪音/诉讼”。

“者”是名词的后缀，属于黏附语素，根据“黏附性测定法”，后缀“者”前面的部分不应与“者”切分。但是，有时“者”前面的部分很长，不便于视读，也应该切分。例如，“经过/苦苦/追求/而/获得/幸福/者”，“不/顾/劝告/而/执意/闹事/者”，“多/次/判刑/而/屡教不改/者”。

“非”是前缀，属于黏附语素，根据“黏附性测定法”，前缀“非”后面的部分不应该与“非”切分。但是，有时“非”后面管辖的范围太长，不便视读，也应该切分。例如，“非/本市/注册/车辆”。

认知心理学还证明了，形式词的汉字序列中首尾两头的汉字较容易辨认。个别的一些长词，如果看一看它们的两头，再加上前后文的提示，则中间的汉字不必细看也可以辨别出这个词来。根据这样的原理，在自动切分时，可以把多音节后缀“－主义”、“－主义者”同前面的汉字连写，反而比分写容易辨认。例如，“马克思列宁主义者”。当然，这样的长词不宜过多，长词

的数目要加以严格的控制。如果长词数目太多,其可辨识性就会随长词数目的增加而降低。

在确定形式词的时侯,我们应该考虑到这些视读方面的原则。

2. 多元化原则(the pluralism principle)

从汉语书面文本自动切分的实际情况来看,切分单位不仅仅是上述的词,还可能是比词更大的单位(如成语和习惯用语),也可以是比词更小的单位(如黏附语素和非语素字)。所以,本文中所说的形式词除了一般意义上的词之外,还包括比词更大以及比词更小的单位。形式词也就是切分单位。

作为切分单位的成语和习惯用语有如前述。

黏附语素和非语素字也可以是切分单位。

某些离合词("洗澡,鞠躬,游泳,理发,出差")在实际文本中可能分离出黏附语素,这时,这些分离出来的黏附语素就成为了切分单位。例如,

洗/了/一/次/澡

鞠/了/一/个/躬

游/了/一/次/泳

出/了/一/次/差

其中的"澡、鞠、躬、泳、差"都是黏附语素,然而,它们都是实实在在的切分单位,也就是我们的形式词。

某些非语素字也可以成为切词单位。例如,

葡萄/的/葡/字/怎么/写/?

鹦鹉/的/鹦/有/什么/意思/吗/?

其中的"葡"和"鹦"都不是语素,它们是没有意义的非语素字,然而,它们都可能成为切分单位。

标点符号也应该是切分单位,从这个意义上说,标点符号也是一种特殊的形式词。

科学技术文章中的公式和符号,也应该是切分单位,也可以看成一种特殊的形式词。

由此可见,我们对于形式词的理解应该是多元化的,形式词不仅仅是词,还可以是成语、惯用语、黏附语素、非语素字,甚至还可以是标点符号、公式或其他符号、数字串、外文字母串,等等。我们应该遵从多元化的原则,对于形式词作广义的理解。从中文信息处理的实际需要来看,我们完全有必要在自动切分中把"理论词"的概念加以扩展,引入"形式词"的概念。

国家标准 GB13715《信息处理用现代汉语分词规范》中,给"分词单位"下的定义是:"汉语信息处理使用的、具有确定的语义或语法功能的基本单位"。我们在本文中提出的"形式词"的外延比这个定义所界说的"分词单位"要广泛一些,这个"形式词"的概念更加适合于中文信息处理的需要。

3. 领域针对性原则(the domain-oriented principle)

我们还可以根据中文信息处理其他领域的实际需要,把形式词的概念引入机器翻译、信息检索、自动分类、自动文摘、语音识别等领域中去,针对不同领域的实际需要,建立不同领域的形式词系统,以弥补语言学中由于"理论词"在理论方面的缺陷而引起的各种困难和矛盾。

例如,在汉语翻译成外语的机器翻译中,词组型的科学技术术语最好不要切分,可以整个儿地翻译为相应的外语术语,这样可以减轻汉语分析的负担。例如,地理学术语"沙漠卵石覆盖层",可以直接翻译为英语的"desert pavement",如果切分开来翻译,译文可能会不知所云。在信息检索中,这样的长术语也最好不要切分,以提高检索系统的查准率。但是,如果在研究

(下转 51 页)

## 参 考 文 献

[1] Tang Haijiang, Pascale Fung. A multi-path syllable to word decoder with language model optimization and automatic lexicon augmentation. 2000 International Symposium on Chinese Spoken Language Processing, Beijing, China, Oct 2000

[2] Chien Lee-Feng. PAT-tree-based adaptive keyphrase extraction for Intelligent Chinese Information Retrieval. Information Processing and Management, 1999, 35: 501－521

[3] Yang Kae-Cherng, Ho Tai-Hsuan, Chien Lee-Feng, *et al*. Statistics-based segment pattern lexicon-a new direction for Chinese language modeling. IEEE, 1998 International Conference on Acoustics, Speech and Signal Processing, Seattle, WA, 1998, 169－172

[4] Gao Jianfeng, Wang Hai-Feng, Li Mingjing, *et al*. A Unified Approach to Statistical Language Modeling for Chinese. IEEE, 2000 International Conference on Acoustics, Speech and Signal Processing, 2000

[5] Wong Pad-Kwong, Chan Chorkin. Chinese word segmentation based on maximum matching and word binding force. The 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 1996, 200－203

[6] Witten I H, Bell T C, The zero-frequency problem: estimation the probabilities of novel events in adaptive text compression. IEEE trans. On inform. Theory, 1991, 37(4): 1085－1094


---

(上接 14 页)

汉语科技术语结构的术语数据库中，为了表示科技术语的结构，就有必要加以切分。不同的领域对于切分的要求是有差别的，我们有必要针对不同的领域建立不同的形式词系统，以满足不同领域的不同要求。

显而易见，针对不同领域的形式词系统应该既有“大同”，又有“小异”。“大同”反映了不同领域的形式词的共性，“小异”反映了不同领域形式词的特性，我们应该把共性和个性结合起来，建立计算语言学中“形式词”的新概念。

形式词研究是计算语言学理论建设的一项基础工作，希望引起学术界的进一步讨论，我们的文章仅是抛砖引玉而已。


## 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] 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

[3] 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

[4] 冯胜利. 从韵律词看汉语“词”“语”分流之大界. 中国语文, 2001, (1): 27－38