

汉语理解的基本单位

石定栩

香港理工大学

摘要：一般都认为机器理解及处理的最小单位是词。但是切词在技术上有很大的困难，而且词库往往规模失控。本文主张汉语的机器处理不必仿照句法分析，应该按照汉语的特点与机器处理语言的要求自找出路。汉语机器理解的最小单位是语素，按词法规则形成的复合词，以及成语性的词和短语。这些单位在句法和语义上不能再分割，而且形成封闭系统，因而可以用语素库和固定搭配组合库的形式加以穷尽。这些单位按照句法规则层层组合，最终形成句子。

The Smallest Unit in Processing Chinese

Dingxu Shi

Dept. of Chinese and Bilingual Studies The Hong Kong Polytechnic University Kowloon, Hong Kong

email: ctdshi@polyu.edu.hk

ABSTRACT: Word is commonly assumed to be the smallest unit of Chinese in natural language processing and corpus building, following the general trend in the syntactic analysis of Chinese. It is very difficult, however, to distinguish morpheme and word or word and phrase in actual segmentation of raw data. The proposal is to use morpheme as the smallest unit for machine processing of Chinese. The lexicon for building Chinese corpus should consist of three parts: one for morphemes, one for morphologically built words and one for fixed expressions. All three parts contain closed class items only and the lexicon will not become infinitely large. The internal structure of compound words should be taken care of by syntax.

0. 引言

自然语言的理解和处理是当前计算机科学的重大课题之一。迄今为止，汉语的理解和处理大致上沿用普通语言学的基本观点，以词作为句子分析处理的最小单位，词上面是短语，然后是句子，句群，篇章等各级单位。自然语言理解和机器翻译等工作都离不开对句子的分析，都必须先将句子分拆成最小单位，然后逐步往上归并，切词因而是语言处理必不可少的第一步。切词的最大困难是词和短语的界线不清，附带的困难之一是最终建立的词库规模失控。目前已有的切分办法，都包括一些硬性规定，其实就是将解决不了的问题掩盖起来。

本文从另一个角度观察问题，试图从计算机分析的实际出发，为汉语的理解和处理找出一条比较可行的路子。

1. 汉语切分的最小单位应该是语素

一般的汉语语法书都将汉语的语法成分分为语素，词，短语，句子，句群这几种。语素是音义结合的最小单位，词是可以在句子中单独发挥作用的最小单位。自由语素可以独立成词，而非自由语素则必须同其它语素结合才能成词。词同词相结合成为短语，短语可以自主成句，也可以同其它短语结合成句。这些规则同普通语言学的一般认识相仿，用来分析汉语句法现象也似乎无可厚非。可是，一到实际的语料分析，就总是困难重重。特别是词同语素的分界，以及词同短语的区分，怎么也说不清楚，只好再加上两条具体判定的标准。一是用能否独立使用作为词和语素的分界，能独立充当句子成分的是词，不能充当句子成分的就是语素（吕叔湘 1982）。“朋”和“友”都不能独立使用，比如不能作为句子的宾语。不能说“找朋”或“找友”，只可以说“找朋友”，所以“朋”和“友”都是语素，而“朋友”是词。二是以能否扩展作为词和短语的区别依据。不能扩充的是词，能够扩充的是短语（吕叔湘 1982）。“白菜”不等于“白的菜”，所以“白菜”是词。“白布”可以扩展成“白的布”，因而是个短语。

这两个标准的好处是简单明了，对任何一个具体例子都能直接做出判断，很少会有模棱两可的情况。但是对实际语料进行切分时，就会发现问题在于得出的结果不可靠。早就有人指出过（符淮青 1984），“楼”在北京话里不能单说，只能说“大楼”、“楼房”及“主楼”等。但是，“楼”可以用在句子里作为宾语。如果有人问，“那里在盖什么？”回答可以是“在盖楼”。按照前一种情况，“楼”是语素，按后一种情况，“楼”就应该是词了。又比方说很多双音节的动宾组合是不能拆开说的，“游泳”可以说，但“游”和“泳”都不能独立使用，所以都是语素。但是，“游了好几次泳”显然是可以说的，所以又想个办法来弥补，将这一类组合归成一类，起个含含糊糊的名字叫“离合词”，也就是承认里面的成分既是不能成词的语素，又在一定条件下可以具有词的地位。类似的情况在动补结构里也很普遍，“享福”里的“享”通常情况下不能单说，是个典型的非自由语素。可是，“他享福享够了”又完全可以说，“够”是个自由语素，是词，那么这里的“享”是不是词就成问题了。

以能否扩展作为鉴定词与短语的依据，王力（1946，1957）赵元任（Chao 1947，1968）早就提出来了，也一直是汉语语法界的主流意见（如李济中、姚锡远 1997，张斌 1998）。<信息处理用现代汉语分词规范>（GB13715）里的划词标准“结合紧密，使用稳定”，大致上源于这一认识。目前的分词工作，大部分也都沿用这一标准。从理论上说，任何一个组合只要能找到扩展的例子，就可以从词的范围排除出去，很少有分不清的例子。但是，实际语料切分的结果则问题多多。比如说，“第一”、“第五”、“第十”这些组合都不能扩展，看成词当然毫无问题。可是，“第一千五百六十一”和“第一百万零一”这样的结构是按类似的方式组合起来的，而且也不能扩展，照此办理的话也只能算是词，能接受这种分法的人就不会太多了。实际切分时，往往就只好再创造一个标准，硬性规定在某个地

方切一刀，例如规定“第十”以下是词，而从“第十一”起都是短语（汤志祥 1997）。

另一个问题是词库的规模。很多从文言文里继承下来的语素，如“机”、“器”、“剂”、“农”、“员”、“锻”、“曲”等，不能独立成词，只有在同其它的语素或词组合以后才能起作用。相关的组合都不能扩展，也就都只能当做是词。问题是这些语素非常能产，可以按照一定的规律不断创造出新的组合来。如果把有关的组合都收到词库里去，一方面会跟不上新组合的创造速度，词库总是落后于形势；另一方面由于包括了太多的开放系统，词库的规模会失去控制，变得大而无当。

当然，上面说的都是老问题，汉语处理的专家们也已经想了很多补救的办法（参见 张普 1992，宋柔等 1998，刘群与俞士汶 1998），取得了不少进展。可是，与其修修补补，被动应付，还不如找一个根本的解决办法，完全避开这个说不清，理还乱的问题。事实上，吕叔湘（1982）先生早就指出过，这两个判断标准本身存在着不合理的地方，只是找不出更好的办法来代替它们而已。比如以扩展的方式来判定词和短语，从句法分析的角度来说其实是没有任何根据的。以受谓词成分修饰的体词性成分为例，能否扩展实际上只牵涉到“的”字结构。如果不先引入“的”字，所有的谓词+体词组合都无法扩展。“很白菜”同“很白布”一样都不能说：“白那种菜”和“白那种布”同样都不可接受。只有在“白布”用“的”加以扩展之后，“白的布”才真正成为短语，才可以说“很白的布”和“白的那种布”。换句话说，只有包含“的”字结的组合才是不折不扣的短语，“白菜”和“白布”的句法地位实际上是相等的，能否用“的”字扩展，其实应该说成某个复合词是否有相应的由“的”字结构修饰的短语。

至于“白菜”为什么没有相应的短语“白的菜”，其根本原因在于“白菜”的意义有了转变，不再表示白颜色的菜，而是一类同“菘”相似的蔬菜，说成“白的菜”就改变了“白菜”的意思。而诸如“演说家”之类的组合之所以不能说成“演说的家”，则是因为“的”字结构是短语，而这个意义的“家”是非自由语素，不能直接受短语修饰。如果换成“演说人”就没问题了，“演说的人”同样可以说。“演说家”和“演说人”的内部结构和意义都完全相等，硬要分到两个不同的大类里去，实在有点说不过去。

这些问题牵涉的数量极大。如果将以往认为是词的各种组合分类的话，可以大致上得到四个类别。一是语素词，即由单个自由语素形成的词，如“牛”、“羊”、“骆驼”和“葡萄”之类。二是由词法过程构成的词，即由前缀或后缀加其它成分构成的词，象“老虎”、“砖头”和“桌子”等。三是由修辞手法构成的词，就是意义发生了转变，整体意思不再由构成该词的语素意义合成的，成语性的词汇，诸如“丹青”、“炒鱿鱼”及“杯弓蛇影”之类。四是由句法过程构成的词，即通常所说的复合词（任学良 1981）。这四类词中，前三类基本上是封闭系统，即使数量有些增减，速度也很慢。只有复合词是能产的，形成一个开放系统，因而可以大批创造，也会随时退出流通范围，数量无法控制。上面说到的问题，牵涉到的词虽然数量极大，其实只同复合词有关，因而应该可以找到解决的办法。

复合词的构成方式可以是“语素+语素”，“词+语素”，“语素+词”也可以是“词+词”。成分之间的关系可以是并列，同位，偏正，主谓，动补或者动宾，也就是包括了常见的所有句法结构关系（朱德熙 1984，任学良 1981）。显而易见，复合词的构成，与通常所说的词与词搭配成短语采用的是类似的方式，只是牵涉到的单位有所不同而已。如果将以词法过程构成的词和成语性的词看成不可分割的整体，汉语句法关系的最小单位应该是语素

和一些不能再分割的整体组合。这就为计算机处理汉语提供了另一条出路。

2. 双库制：语素库+固定搭配库

上面的这种分析并非否定词在汉语句法中的地位，而是要说明词同语素，词同短语之间的界线十分模糊，同一个组合很在不同的环境中可能具有不同的句法地位，一定要将某个组合判定为词或短语，有时候无法做到。另一方面，如果企图将所有的复合词都放到词库里去，必然无法穷尽，而且会造成词库规模失控。

自然语言的机器处理其实并不一定要完全照抄理论句法分析的模式。大部分印欧语言的复合词是由词组合而成的，词不但是在句子中发挥作用的最小单位，也是句法过程所牵涉到的最小单位。而汉语复合词的构成单位可以是语素，其组合同样牵涉到句法过程，语素就成了句法过程的最小单位，所以汉语句子中发挥作用的最小单位和句法过程的最小单位并不等同。如果能替计算机建立一个汉语句法关系分析系统的话，语料的机器分析以词为单位和以语素为单位应该差不多，只是相差一个层次而已，难度的区别不会太大。但是从语素做起有个显著的额外好处。一旦不需要以词为单位进行分析，建立词库也不再是必由之路，切词的烦恼可以大大减少，词库的不可穷尽性和庞大规模也不再是致命伤了。

以语素为句法分析的最小单位自然就要建立语素库。汉语的语素在某一特定时刻是个可以穷尽的封闭系统，基本语素增加和消失的速度都极为缓慢，因而建立完整的汉语语素库可行性极高，语素库的规模也不会是天文数字。外语借词的不断引进是建立语素库的一大问题，语素的句法性质标注也会遇到一些困难，但这些困难都可以借鉴以往词库建设的经验加以解决。

当然，只有语素库并不能解决所有的问题，有些词的整体意义同内部语素的个别意义无关或者关系不大，完全按照句法关系分析这种词会有解困难。这就需要建立一个固定搭配库，专收两类组合，由词法过程构成的词和已经转意或成语性的词和短语。前者是由语缀（前缀，后缀和中缀）加语素或词构成的词。这些词是否应称为复合词尚有争议（张斌 1998），但其语法地位在建立词库的过程中可以忽略不计。现代汉语很少有由语素加语缀产生的新词（任学良 1981），所以这类词基本上形成个封闭系统，数量也不太大，建立专门词库的困难不算大。

成语性组合的数量较大，特别在社会政治经济发生剧烈变化时会有大量新组合产生。前些年产生的“倒爷”、“大腕”，这两年的“下课”、“下岗”都属于这种情况。另一个问题是成语性组合同一般复合词之间的界线不好划，组合的意义转化到什么程度才算成语不容易说清楚。解决这一问题的最简单办法是宁宽毋窄，只要有一点转意就收进来。这样做当然会把库容搞得很大，但再大也不会比现在的词库更大。

从某种意义上说，语素库和固定搭配组合库的本质大致相同，收录的都是句法分析过程中不可分割的最小单位，也就是汉语机器处理和理解的最小单位。将这些单位按句法规则一层一层地组合起来，就可以最终得到句子。词和短语都是机器分析和理解的中间单位，没有必要再建立词库或短语库了。

3. 同左邻右舍的关系

本文讨论的是汉语机器处理和机器理解的机制，当然也适用于相关的领域如机器翻译和语音输入等等。但是，这里提出的解决办法不一定可以推广到其它领域里去。比如说，本文并不主张取消词的概念，也不认为可以抹杀语素同词的区别，在句法分析里这仍然是两个十分有用的单位，只是句法分析允许同一组合在不同场合以不同身分出现而已。

本文讨论的主要对象是复合词，主张用句法分析来处理复合词内部成分之间的关系，并不是说复合词的内部结构同短语的内部结构完全相同，可以用同一方式从句子到语素一下子分析到底。复合词内部成分的组合方式有自己的特点，无法完全沿用现代句法理论的短语结构来描述，而是必须另找出路，用核心之间的关系来描述。这两种关系都是现代句法中已经有的，不存在标新立异的问题。

这里说的双库制应该适用于机器词典的编撰，但不牵涉到一般字典或词典的编写。机器词典是为计算机处理汉语而设计的，一般的词典是为人使用而设计的，两者可以互相借鉴，但不一定完全按照同一原则设计。人用词典的收词可以滞后一些，等某一个新词稳定下来再收进去。而机器词典的收词原则应该是动态的（张普 1999），有新词就收，但淘汰了的则要保留相当一段时间，不然就无法处理历史文献了。

参 考 文 献

- [1] 符淮青（1985）《现代汉语词汇》北京：北京大学出版社。
- [2] 李济中，姚锡远（1997）《现代汉语专题》北京：中国社会科学出版社。
- [3] 刘群，俞士汶（1998）“汉英机器翻译的难点分析”。《1998 中文信息处理国际会议论文集》507-514 页。北京：清华大学出版社。
- [4] 吕叔湘（1982）《语文常谈》香港：三联书店。
- [5] 吕叔湘（1984）《语文杂记》上海：上海教育出版社。
- [6] 任学良（1981）《汉语构词法》北京：中国社会科学出版社。
- [7] 宋柔，戴伟长，邱超杰，季飞（1998）“现代汉语二字结构”。《1998 中文信息处理国际会议论文集》181-187 页。北京：清华大学出版社。
- [8] 王了一（王力）（1946）《中国语法纲要》上海：开明书店。
- [9] 王了一（王力）（1957）《汉语语法纲要》北京：新知识出版社。
- [10] 张斌（1998）《汉语语法学》上海：上海教育出版社。
- [11] 张普（1992）《汉语信息处理研究》北京：语言学院出版社。
- [12] 张普（1999）《关于大规模真实文本语料库的几点理论思考》《语言文字应用》1999 年第 1 期。
- [13] Chao, Yuen-ren. 1948. *Mandarin Primer*. Cambridge: Harvard University Press.
- [14] Chao, Yuen-ren. 1968. *A Grammar of Spoken Chinese*. Berkeley: University of California Press.