RECENT ADVANCE OF TERMINOLOGY RESEARCH

术语学研究新进展

刘　青　易绵竹　主　编
刘伍颖　温昌斌　副主编

国防工业出版社
National Defense Industry Press

术语学研究新进展

Recent Advance of Terminology Research

刘　青　易绵竹　主　编
刘伍颖　温昌斌　副主编

国防工业出版社
·北京·

内 容 简 介

本书是第五届“中国术语学建设暨术语规范化”研讨会(2013 年 10 月 26~27 日，洛阳)的论文集。书中选录的 53 篇论文是从全国各地学者的投稿中精选出来的。本书内容包括：术语学理论、术语规范化与标准化、科学技术名词审定工作、术语翻译与词典编纂、术语学教育、大数据背景下的多语言多学科术语研究、计算术语学与术语知识工程等。

本书充分展示了国内术语学研究与应用的最新进展，也展示了最近一段时间研究的前沿和动向，对术语学基础研究和产品开发具有重要的参考价值。

本书可供术语学、语言学等专业的科研人员、工程技术人员、大学教师和研究生学习参考。

图书在版编目（CIP）数据

术语学研究新进展/刘青，易绵竹主编. —北京：国防工业出版社，2015.3

ISBN 978-7-118-09739-9

Ⅰ.①术... Ⅱ.①刘... ②易... Ⅲ.①术语学-文集 Ⅳ.①H083-53

中国版本图书馆 CIP 数据核字(2015)第 040627 号

※

国防工业出版社 出版发行

（北京市海淀区紫竹院南路 23 号 邮政编码 100048）

北京京华虎彩印刷有限公司印刷

新华书店经售

*

开本 880×1230 1/16 **印张** 18¾ **字数** 542 千字

2015 年 3 月第 1 版第 1 次印刷 **印数** 1—1500 册 **定价** 48.00 元

（本书如有印装错误，我社负责调换）

国防书店：（010）88540777 发行邮购：（010）88540776

发行传真：（010）88540755 发行业务：（010）88540717

前　言

由全国科学技术名词审定委员会主办的第五届“中国术语学建设暨术语规范化”研讨会，于 2013 年 10 月 26～27 日在洛阳解放军外国语学院举行。“中国术语学建设暨术语规范化”研讨会是国内最权威、最具影响的术语学研究学术会议。经会议程序委员会严肃认真地评审，本届会议最终录用论文 53 篇。本书是本届会议的论文集，内容包括术语学理论、术语规范化与标准化、科学技术名词审定工作、术语翻译与词典编纂、术语学教育、大数据背景下的多语言多学科术语研究、计算术语学与术语知识工程等。术语学研究与规范化工作，对支持科技发展，保障语言健康，传承中华文化，促进社会进步，维护民族团结和国家统一有着不可替代的重要作用和意义。

编撰本书的目的是学习和借鉴各术语学派的理论及术语工作经验，培养、充实和增强术语学研究队伍，提高我国术语学研究与规范化工作的水平，并致力于向国际化方向发展。积极研究国内外术语学理论，结合汉语特点，在总结前人研究成果的基础上，为逐步建立起我国特色的术语学理论而努力。本书的内容非常丰富，与以往历届会议论文集相比，无论是在研究与应用的深度还是在广度上都有了新的拓展。总的感觉是，学者们近年来研究工作的“剑锋”所指，逐步更多地指向了大数据背景下的术语研究，这无疑是一个值得鼓励的正确方向。相信读者从本书中一定会深受启发。

最后，诚挚地感谢会议特邀报告演讲者冯志伟先生和戴昭铭先生，会议 DEMO 演示者刘伍颖博士，以及全体作者对会议和论文集出版的热诚支持。这些共同的努力，确保本届会议论文集的出版取得了圆满成功。

编　者

2014 年 3 月

目　录

用计算机分析术语结构的尝试

冯志伟

（杭州师范大学外国语学院，杭州　311121）

摘　要：本文采用有限状态转移网络的方法来分析单词型术语的结构，采用短语结构语法来分析词组型术语的结构。这是中国学者用计算机自动分析术语的最早尝试。

关键词：单词型术语；词组型术语；有限状态转移网络；短语结构语法；自动剖析。

近年来，在术语学的研究中，开始引进自然语言的计算机处理的方法和技术，出现了“计算术语学”(Computational terminology)这样的学科。在1998年的计算语言学国际会议（COLING-ACL’98）上组织了世界上第一次计算术语学的讨论会(First Workshop on Computational Terminology)，这次讨论会首次使用的计算术语学这个学科名称。这次讨论会讨论的问题主要如下：

（1）如何抽取术语以满足信息检索的需要；

（2）如何抽取术语以便使用双语语料库来进行翻译；

（3）如何进一步完善和原有术语抽取的工作(例如，如何建立概念层级网络，如何搜索语义信息或概念信息)。

1998年的这次讨论会成为了计算术语学发展的催化剂，从此，计算术语学便成为一个新兴的术语学的学科，活跃在当代科学技术的百花园中，并且一天天地成熟起来，初步具备了系统的理论和有效的方法，值得我们特别地关注。

在计算术语学这个名称出现10年之前，我国冯志伟在1988年就注意到术语的自动处理问题。他在德国夫琅禾费研究院(Fraunhofer Institute)使用计算机对汉语的单词型术语和词组型术语进行了自动结构分析，是国际上最早进行计算术语学研究的学者之一。

本文介绍冯志伟在24年以前进行的这次尝试。

一、单词型术语的结构自动分析

术语中的语缀比普通语言丰富得多，原因在于术语倾向于使用数量有限的希腊语和拉丁语前缀、后缀和词干构成大量的派生词和复合词，而这些词汇成分在普通语言中很少使用。

术语的形态特征研究同时表明，不同类型的希腊、拉丁语缀功能各异。

前缀的主要功能在于促进术语结构系统化。特定的前缀，有助于领域专家对术语进行分类，建立不同的术语集。

例如，infra-red(红外线)，infrasonic(次声的)；submarine(海下的)，subtropical(亚热带的)；semicircle(半圆)，hemisphere(半球)。

而后缀在通过限定方式改变词类，发挥句法作用的同时，也表达了概念不同方面的内容以及术语的语义类别。

例如，booklet(小书)，leaflet(小叶子)，starlet(小星星)；impressionism(印象派)，racism(种族主义)。

单词型术语是由一个单词构成的，其中仅仅包含一个单词。一般地说，单词可以由词根、词缀和词

尾构成，词根和词缀可以组成词干，词根也可以单独成为词干，因此，我们用图 1 所示的有限状态转移网络来表示一个单词的词法分析过程。

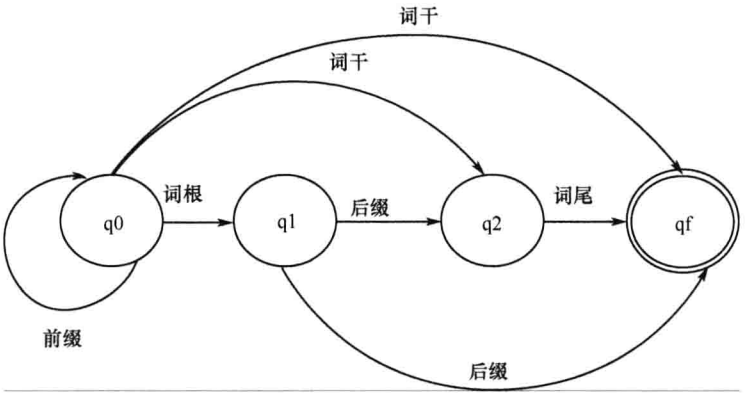

图 1　有限状态转移网络作词法分析

在图 1 中，如果一个单词只包含词干，则其遍历过程是 q0→qf。 如英语的 form (“形式”)。

如果一个单词包含前缀、词干，则其遍历过程是 q0→q0→qf。 如英语的 reform (“改革”，re-是前缀，form 是词干)。

如果一个单词包含词根、后缀，则其遍历过程是 q0→q1→qf。如英语的 formation (“形成”，form 是词根，-ation 是后缀)。

如果一个单词包含前缀、词根、后缀，则其遍历过程是 q0→q0→q1→qf。如英语的 reformation (“革新”，re-是前缀，form 是词根，-ation 是后缀)。

如果一个单词包含词干、词尾，则其遍历过程是 q0→q2→qf。如英语的 forms (form 是词干，-s 是词尾)。

如果一个单词包含前缀、词干、词尾，则其遍历过程是 q0→q0→q2→qf。如英语的 formations (form 是词根，-ation 是后缀，-s 是词尾)。

如果一个单词包含前缀、词根、后缀、词尾，则其遍历过程是 q0→q0→q1→q2→qf。如英语的 reformations (re-是前缀，form 是词根，-ation 是后缀，-s 是词尾)。

由此可见，采用有限状态转移网络，可以非常清楚地描述屈折型语言单词的词法分析过程。

应该指出的是，在词根与后缀相连接时，有时会发生音变。例如，英语的词根 decide 与后缀 -ion 连接成 decision 时，-de-变为 -s-，decide 中的元音 i 读为 [ai]，在 decision 中变为 [i]。但是，英语的词根 deny 与后缀-able 连接成 deniable 时，-y 在书写形式上变为-i，deny 中的 y 读为[ai]，在 deniable 中变为-i-之后，读音仍然为[ai]。对于这些复杂的音变问题，在用有限状态转移网络来进行单词的词法分析时，应该建立相应的音变规则来处理。

下面，进一步举例说明如何用有限状态转移网络来进行德语、法语单词的结构分析。

德语屈折变化丰富，名词、形容词、冠词和指示词有性、数、格的变化，动词有变位形式。

德语中存在大量的派生词，一个单词的词干加上前缀可构成许多新的单词。最常见的是由动词加前缀构成新的动词，由名词和形容词加后缀构成新的名词和形容词。

由动词加前缀构成的动词，如由 rufen (叫)加前缀 aus-构成 ausrufen(呼喊)，aus-是前缀，ruf 是词干，-en 是词尾，也可以用图 1 中的有限状态转移网络来进行词法分词，其遍历过程是 q0→q0→q2→qf。

由名词和形容词加后缀构成新的名词和形容词，如由名词 Kunst(艺术)加后缀-ler 构成的名词 Kunstler(艺术家)，由名词 Stern(星)加后缀-artig 构成的形容词 sternartig(星状的，stern 是词根，-artig 是后缀)，由形容词 neu(新的)加后缀-artig 构成的形容词 neuartig (新型的，neu 是词根，-artig 是后缀)，也

可以用图 1 中的有限状态转移网络来进行词法分析，其遍历过程是 q0→q1→qf。

在德语中还经常使用复合词，这种复合词由限定词加上基本词构成，基本词位于复合词的后部，复合词的性和数由基本词决定，基本词还决定复合词的基本含义，限定词对基本词起修饰和限定的作用。例如，在 Intelligenztest(智力测验)这个复合词中，基本词是 Test(测验)，限定词是 Intelligenz(智力)，它进一步限定了基本词 Test 的确切含义。

图 1 中的有限状态转移网络不能分析这样的复合词，我们必须加以改进，使它在分析了复合词中的限定词之后，还能进一步分析复合词中的基本词。为此，我们从终极状态 qf 出发，再加一条指向初始状态 q0 的弧，并标以#，使之从状态 qf 跳回 q0，再进一步分析复合词中的基本词，如图 2 所示。

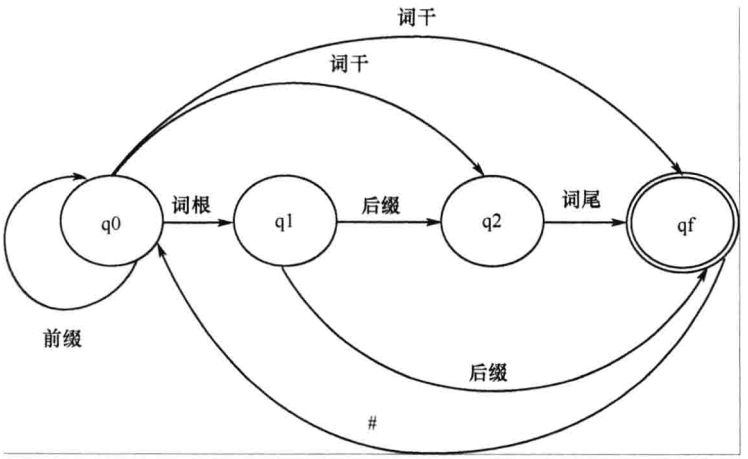

图 2　可以分析复合词的有限状态转移网络

例如，Weltgeschichtlich (世界历史的)这个复合词，由名词 Welt(世界)加形容词 geschichtlich (历史的)复合而成。Welt 是限定词中的词干(这个限定词只有词干)，geschicht 是基本词中的词根，-lich 是基本词中的形容词后缀。这个复合词可利用图 2 中的有限状态转移网络来进行词法分析，其遍历过程是：q0→qf→q0→q1→qf。其中，在 qf 与 q0 之间，进行了一次返回初始状态的“跳跃”。

德语的术语很多是复合词，在许多复合词中，在组合成复合词的各个词之间，往往要加上 -s-，-es-，-en-，-n-，-er-等字母，有的要去掉修饰词的词尾 -e-。例如，Lebenszeichen (生命象征)中，Leben (生命)与 Zeichen (象征)之间加上了 -s-；在 Sinneszelle (感觉细胞)中，Sinn (感觉)与 Zelle (细胞)之间加上了-es-；在 Nervenzelle(神经细胞)中，Nerv (神经)与 Zelle (细胞)之间加上了-en-；在 Sonnenstrahl (阳光)中，Sonne (太阳)与 Strahl (光线)之间加上了 -n-；在 Kinderklinik (儿童诊所)中，Kind (儿童)与 Klinik (诊所)之间加上了-er-；在 Erdgas (天然气)中，去掉了修饰词 Erde (地球)的词尾-e。这些问题，在词法分析时，要建立相应的音变规则来加以处理。

有时，德语的复合词可由两个以上的词组成，这只需在转移到终极状态 qf 之后，再往开始状态 q0 跳跃一次或几次就行了，仍然不难用图 2 中的有限状态转移网络来进行词法分析。但是，当复合词由若干个词组合成的时候，切分时往往会出现莫棱两可、举棋不定的情况，这就需要在各种可能的切分情况中进行选择，确定一种正确的切分，排除不正确的切分。

例如，Bauerlaubnisse (建筑许可)这个复合词，在德语的机器词典中，存有 Bauer (das Bauer，中性名词，鸟笼)，Bau (动词 bauen 的词干，建筑)，Bauer (der Bauer，阳性名词，农民)，Erlaub (动词 erlauben 的词干，许可)，Erlaubnis (die Erlaubnis，阴性名词，许可)，Laub (das Laub，中性名词，树叶)，Nisse (die Nisse，阴性名词，虱子卵)，-se(名词词尾)等语素，因此，可能存在的切分情况有三种。

（1）Bau + erlaubnis + se；

（2）Bauer + laub + nisse；

（3）Bau + erlaub + nisse。

为了在这三种可能的切分中选择出正确的切分，可检查每种切分在语义上的相容性。

在（1）中，其语义的组合情况为

建筑 + 许可 + 名词词尾

切分出来的三个部分的语义是相容的。

在（2）中，其语义的组合情况为

鸟笼 + 树叶 + 虱子卵

或

农民 + 树叶 + 虱子卵

切分出来的三个部分在语义上不相容。

在（3）中，其语义的组合情况为

建筑 + 许可 + 虱子卵

切分出来的三个部分在语义上也不相容。

所以，我们选择语义上相容的第（1）种切分，排除语义上不相容的第（2）（3）两种切分，并确定这个复合词的词义为“建筑许可”。

法语是从拉丁语演变而来的。与拉丁语相比，法语的词形屈折已大大简化，名词没有格的变化，性和数主要通过名词前的冠词、限定词来区别，动词有变位形式，形容词也有性与数的变化，少数形式还比较复杂；法语的词从结构上也可以分为前缀、词干、词根、后缀、词尾几部分，名词、形容词、动词都可以通过加前缀或后缀来派生。

由词干加前缀构成的词，如 contrevent (风窗，contre-是前缀，vent 是词干)，extrafin (纤细，extra-是前缀，fin 是词干)，可用图 1 中的有限状态转移网络来分析，其遍历过程是 q0→q0→qf。

由词根加后缀构成的词，如 mouvement (运动，mouve 是词根，-ment 是后缀)，durable (持久，dur 是词根，-able 是后缀)，可用图 1 中的有限状态转移网络来分析，其遍历过程是 q0→q1→qf。

由词根加前缀和后缀构成的词，如 surproduction (生产过剩，sur-是前缀，product 是词根，-ion 是后缀，telespectateur (电视观众，tele-是前缀，spectat 是词根，-eur 是后缀)，也可用图 1 中的有限状态转移网络来分析，其遍历过程是 q0→q0→q1→qf。

在具体的法语词法分析中，图 1 中的有限状态转移网络显得过于笼统和简单。

在法语中，当名词后缀是-ance，-ation，-ade，-ment 时，其词根一般是动词词根。例如，名词 obeissance (服从)的词根是动词词根 obeiss-,名词 creation (创造)的词根是动词词根 cre-，名词 promenade (散步)的词根是动词词根 promen-，名词 fabrication (生产)的词根是动词词根 fabric-(fabriqu-的音变形式)。

当形容词后缀是 -able，-if 时，其词根一般也是动词词根。例如，形容词 navigable (可通航的)的词根是动词词根 navig-，形容词 pensif (沉思的)的词根是动词词根 pens-。

当名词后缀是-ité，-esse 时，其词根一般是形容词词根，例如，名词 fidelité (忠实)的词根是形容词词根 fidel-，名词 souplesse (柔软)的词根是形容词词根 soupl-。

由形容词词根构成名词时，有时还会发生音变。例如，名词 sottise (笨拙)由形容词词根 sot-(愚笨)和后缀 -ise 构成，而在它们之间，要加辅音字母-t-。

基于这些情况，我们有必要区分构成合成词的词根是动词词根还是形容词词根，从而更加细致地描述名词和形容词的词法分析过程。

另外，分析的方向也不一定总是从左到右，也可以从右到左，先分析词尾、后缀，再分析词根，最后才分析前缀。

为了处理法语中这些复杂的语言现象，我在法—汉机器翻译系统 FCAT 的研制中，提出了如图 3 所示的有限状态转移网络。

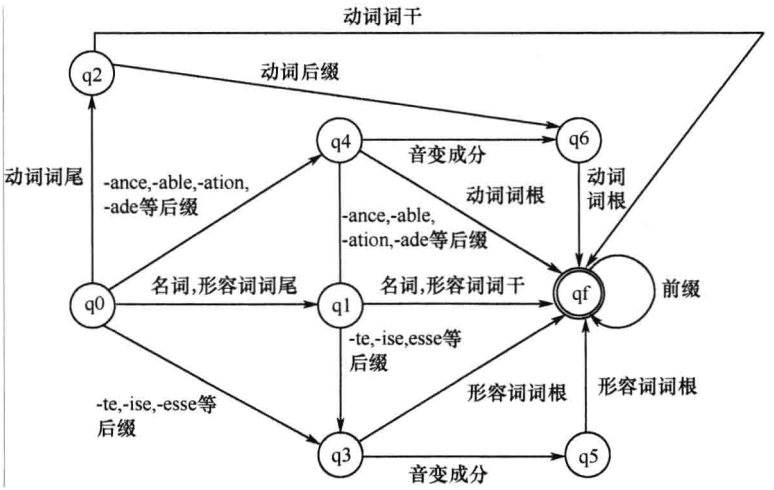

图 3　法语词法分析的 FSTN

这样，词根为动词词根的名词，如果没有音变成分，则其遍历过程是 q0→q4→qf，例如，法语的 creation，先分析后缀-ation，后分析动词词根 cre-。如果有音变成分，则其遍历过程是 q0→q4→q6→qf。例如，法语的 fabrication，先分析后缀-ation，再把音变成分-c-变为-qu-，再分析动词词根 fabriqu-。

词根为形容词词根的名词，如果没有音变成分，则其遍历过程是 q0→q3→q5。例如，法语的 souplesse，先分析后缀-esse，再分析形容词词根 soupl。如果有音变成分，遍历过程是 q0→q3→q5→qf. 例如，法语的 sottise，先分析后缀-ise，再分析音变成分-t-，最后分析形容词词根 sot.

法语的名词、形容词、动词都有词尾屈折变化。如果名词、形容词有屈折变化词尾，则首先还要分析词尾，再分析后缀和词根。无音变时，其遍历过程是 q0→q1→q3→qf 或 q0→q1→q4→qf，有音变时，其遍历过程是 q0→q1→q3→q5→qf 或 q0→q1→q4→q6→qf. 如果动词有屈折变化词尾，则首先分析动词词尾，再分析动词词干，其遍历过程是 q0→q2→qf。

如果名词、形容词、动词还有前缀，则还需在终极状态 qf 分析了前缀之后，再回到这个终极状态 qf. 例如，法语的 prefabrication (预制)，其遍历过程是 q0→q4→q6→qf→qf。首先分析后缀-ation，再把音变成分-c-改变为-qu-，再分析动词词根 fabriqu-，最后再分析前缀 pre-。

汉语单词型术语的结构比较简单，也可以使用图 1 中的有限状态转移网络来分析。

- 只有词干的单词型术语：例如，“速度、能量”，遍历过程是 q0→qf。
- 带前缀的单词型术语：例如，“超导体、非金属”，其中“超，非”是前缀，遍历过程是 q0→q0→qf。
- 带后缀的单词型术语：例如，“电气化、绝缘体”，其中“化、体”是后缀，遍历过程是 q0→q1→qf。
- 带前缀和后缀的单词型术语：例如，“非周期性，反铁氧体”，其中的“非、反”是前缀，“性、体”是后缀，遍历过程是 q0→q0→q1→qf。

汉语的语缀不仅可以附加在词根或单词上，还可以附加在词组上。例如，“非线性规划”，中的附加前缀“非”，“同素异形体”中的附加后缀“体”。对于这样的术语，我们可以按照德语单词型术语中复合词的结构分析方法来处理，使用图 2 中的有限状态转移网络来进行分析。使用这样的有限状态转移网络，“非线性规划”的遍历过程是 q0→q0→qf→q0→qf，“同素异形体”的遍历过程是 q0→qf→q0→q1→qf。

根据有限状态转移网络的原理，单词型术语经过自动分析之后，就可以输出与该单词型术语有关的形态信息。例如，以 beauty 为词干的单词型术语 beautified 经过形态分析之后，可以得到如下的分析结果：

beatified：　　<<<*>N + ify>V + ed>A

其中，*表示 beauty，N 表示它是一个名词，加上-ify 之后，变成 beautify，是一个动词(V)，再加上-ed 之后，变成 beautified，是一个形容词(A)。

同样，我们还可以得到其他单词型术语的分析结果[①]：

beautification：　　<<<*>N + ify>V + cation>N

beautifier：　　<<<*>N + ify>V + er>N

beautiful：　　<<<*>N + ful>A

unbeautified　　<un# <<<*>N + ify>V + ed>A>A

unbeautiful　　<un# <<<*>N + ful>A>A

根据前面 beautified 的例子，我们不难理解到这些分析结果的含义。

单词型术语的自动分析是对于单词型术语中的各个组成成分进行自动分析，在自然语言处理中属于自动词法分析(automatic morphological analysis)的范围。而词组型术语的自动分析，就属于自动句法分析的范畴了。下面我们讨论词组型术语的自动分析问题。

二、词组型术语的自动剖析

根据冯志伟提出的“术语形成的经济律”，在一个术语系统中，词组型术语比单词型术语多得多。因此，在计算术语学中，有必要研究词组型术语的自动分析问题。词组型术语是由若干个句法单位构成的，是有结构的。词组型术语的自动分析就是计算机自动地识别词组型术语的各个句法单位以及它们之间的相互关系的过程，这个过程，又称为“自动剖析”(automatic parsing)。

词组型术语的剖析技术是建立在自然语言的形式语法的基础之上的。所谓剖析，就是要用形式语法来分析词组型术语的结构，使之能清晰地、形式化地表示出来，因此，形式语法在词组型术语的剖析中有着极为重要的作用。

一般地说，一种好的形式语法，在语言的描述方面应该尽量地自然、明白、易懂，在数学的表达方面，应该有很强的说明力和解释力，在计算技术方面，应该具有较高的效率。

美国语言学家乔姆斯基(N. Chomsky)提出，形式语法 G 可以用下面的四元组来定义：

G = (Vn，Vt，S，P)

其中，Vn 是非终极符号的集合，这些符号是专门用来描述语法类别的，它们是范畴符号，如词类符号、词组类型符号等；Vt 是终极符号的集合，它们就是被定义语言中的具体的单词；S 是初始符号，它是集合 Vn 中的一个特殊成员；P 是重写规则的集合，其中的每一条规则都具有

$$\varphi \rightarrow \psi$$

的形式，φ称为规则的左部(Left Hand Side，LHS)，ψ称为规则的右部(Right Hand Side，RHS)，φ→ψ意味着可以用规则的右部φ来置换规则的左部φ。

给定了一个语法 G，就可以从初始符号 S 开始，应用重写规则推导出这种语法 G 所描述的语言 L(G). 具体地说，可以用重写规则 S→φ_1，从 S 推导出新的符号串φ_1，再利用重写规则$\varphi_1 \rightarrow \varphi_2$，从$\varphi_1$推导出新的符号串$\varphi_2$，…，一直到我们得到不能再重写的符号串$\varphi_n$为止。这样推导出的终极符号串$\varphi_n$，就是语言 L(G)中成立的词组型术语。

① 参见 Christian Jacquemin，Spotting and Discovering Terms through Natural Language Processing，p20，The MIT Press，2001.

“短语结构语法”是乔姆斯基形式语法中最重要的一个类型。确切地说，这种短语结构语法应该叫做上下文无关的短语结构语法(Context-Free Phrase Structure Grammar，CF-PSG)。这种语法的重写规则为

$$A \rightarrow \omega$$

其中，A 是单个的非终极符号(即范畴符号)，ω是非空的符号串，ω可以由终极符号组成，也可以由非终极符号组成，也可以由终极符号与非终极符号混合组成。

有了一个上下文无关的短语结构语法，就可以用 RHS 中的符号串来重写 LHS 中的范畴符号，RHS 的符号串中可以含有范畴符号，也可以含有具体的单词。当用上下文无关的短语结构语法把 LHS 中的范畴符号重写为具体的 RHS 的时候，不必考虑 LHS 的范畴符号所出现的上下文，规则的使用对于上下文没有任何的限制，这就是为什么这种语法称为 “上下文无关的短语结构语法”的原因。当今在程序设计语言中所使用的巴库斯—瑙尔范式(Bacus-Naur Normal Form)就是上下文无关的短语结构语法。

为了行文上的方便，在不引起混淆的情况下，在下面的叙述中，把上下文无关的短语结构语法叫做“短语结构语法”(Phrase Structure Grammar，PSG)。

我们提出如下的短语结构语法来剖析词组型术语“延迟线存储器”和“失灵区部件”：

G = (VN，VT，S，P)

VN = {NP，VP，N，V}

VT = {线，区，部件，存储器，延迟，失灵}

S = {NP}

P：

NP → N (i)

NP → V + N (ii)

NP → NP + N (iii)

VP → V + N (iv)

N → {线，区，部件，存储器} (v)

V → {延迟，失灵} (vi)

词组型术语“延迟线存储器” 和“失灵区部件”的结构是相同的，下面，以“延迟线存储器”为例来说明它们的自动过程。

我们从初始状态开始，写出词组型术语“延迟线存储器”的推导过程：

推导过程	所用规则
NP	开始
NPN	(iii)
V N N	(ii)
延迟 N N	(v)
延迟 线 N	(iv)
延迟 线 存储器	(iv)

上述推导过程，也就是这个词组型术语的生成过程。

由短语结构语法生成的词组型术语，可以用树形图 4 来表示。

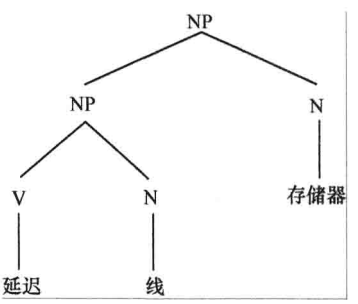

图 4 树形图

这种与短语结构语法相对应的树形图，称为“短语结构树”(Phrase Structure Tree)。

也可以把短语结构树表示为一个表(list)，表中的第一个元素是树形图的根上的标记，后面的各个元素是相应节点的直接后裔的标记，按它们在词组型术语中出现的顺序排列，在 LISP 语言中，上述的短语结构树可表示为

(NP(NP((V 延迟)(N 线))(N 存储器)))

由于表中的第一个元素是树形图中根结点的标记，而后的各个元素依次是其后裔的标记，而这些元素本身也是表。这样的表写成下面的形式更醒目：

```
(NP
   (NP
      ((V 延迟)
      (N 线))
         (N 存储器)))
```

如果一种语言可以由短语结构语法来描述，由于短语结构语法是与上下文无关的，因此，这种语言可以称之为上下文无关语言(Context Free Language，CFL)。

由于短语结构语法便于书写，便于修改，因而受到了自然语言处理研究者的普遍欢迎，推动了自然语言处理的发展，在自然语言处理中屡建奇功。短语结构语法的形式清晰，易学易记，在剖析、翻译和编译等技术中得到广泛的应用，自然语言处理早已研制出了用于剖析和识别上下文无关语言 CFL 的高效算法，可见自然语言处理学界对于短语结构语法之重视。词组型术语是自然语言的重要组成部分，当然也就可以使用短语结构语法来对词组型术语进行自动的剖析，从而揭示词组型术语的内部结构，加深我们对于词组型术语的认识。

下面，介绍两种基于短语结构语法的剖析技术——自底向上剖析(Bottom-up Parsing)和自顶向下剖析(Top-down Parsing)。

1．自底向上剖析

如果有包含三个词的汉语词组型术语“延迟线存储器”，经过自动切词(Automatic Segmentation)之后，这个词组型术语的词与词之间出现了空白，其形式变为

延迟　线　存储器

使用前述的短语结构语法 G，可知第一个词“延迟”应该属于 V 这个句法范畴，因为在语法 G 的重写规则(v)中，与规则右部 RHS“延迟”相匹配的规则左部 LHS 是范畴符号 V。这样，得到如下的剖析图：

```
V__
延迟　线　存储器
```

然后，继续剖析符号串“V 线　存储器”。我们检查在语法 G 中，有没有右部 RHS 为 V 的重写规则。例如，如果在语法 G 中有 K→V 这样的重写规则，那么，就可以把 V 置于 K 之下，让 K 来支配 V；但是，在语法 G 中没有这样的重写规则，因此，来检查所得符号串中的第二个词“线”，根据规则(v)，发现“线”的范畴符号是 N，于是，得到如下的剖析图：

```
V__　　N__
延迟　线　　存储器
```

在剖析过程中，要设法在语法 G 所容许的范围内，尽量把符号串中的范畴符号组合起来。首先，再一次检查在语法 G 中，有没有右部 RHS 只包含 N 的重写规则，我们发现重写规则(i)正是这样的规则，于是，把 N 置于 NP 的支配之下，得到如下的剖析图：

```
　　NP__
V__　　N__
延迟　线　　存储器
```

现在，NP 位于初始符号 V 之后。我们再一次检查语法 G 中有没有右部 RHS 中只包含 NP 的重写规则，检查结果是没有。我们再来检查语法 G 中有没有规则右部 RHS 为符号串 V NP 的重写规则，检查结果也没有，在这种情况下，不可能再继续处理了，我们一定是在剖析过程的什么地方误入歧途，而导

致了剖析的失败，使剖析进入了死胡同。

为了跳出这个死胡同，我们采用“回溯”(backtracking)的办法，回到剖析过程中进行多重选择的情况去。为此，首先把支配 N 的 NP 去掉，得到如下剖析图：

V__　　N__
延迟　线　　存储器

可以看出，前面的剖析过程之所以进入死胡同，是因为我们过早地把 N 置于 NP 的支配之下，而 V NP 本身又不能单独地出现在语法 G 的重写规则的右部 RHS 之中，因此，我们使用重写规则(iv)，把 V 和 N 置于 VP 的支配之下，得到如下剖析图：

____VP_____
V__　　N__
延迟　线　　存储器

在这种情况下，首先检查 VP 能否作为规则右部 RHS，发现不行。因此，我们不得不进一步回溯，抹去 VP 这一个范畴符号，于是，得到如下剖析图：

V__　　N__
延迟　线　　存储器

使用重写规则(ii)，把 V 和 N 置于 NP 的支配之下，得到如下的剖析图：

____NP_____
V__　　N__
延迟　线　　存储器

首先检查符号串 NP 能否成为语法 G 中重写规则的右部 RHS，发现不行。于是根据规则(v)把“存储器”规约为 N，得到如下的剖析图：

____NP_____
V__　　N__　　N__
延迟　线　　存储器

再来检查符号串 NP　N 能否成为语法 G 中重写规则的右部 RHS，发现重写规则(ii)正好满足这样的条件，于是，我们把符号串 NP　N 置于 NP 的支配之下，得到如下的剖析图：

_____NP_____
____NP_____
V__　　N__　　N__
延迟　线　　存储器

这个 NP 的跨度从词组型术语的头部开始，到词组型术语的尾部结束，覆盖了整个的词组型术语，因此，这个词组型术语的剖析成功。

前面的剖析过程可以归结为如图 5 所示的搜索树(search tree)。

从搜索树上可以看出，要完成一个词组型术语的剖析，其搜索过程是比较复杂的。如果搜索一开始，就能找到正确的途径而得到成功，那当然是最理想不过的。然而，在实际的剖析过程中，往往要经过多次的反复和回溯才能取得成功，有时还要不厌其烦地穷尽各种可能性，我们的程序总有那么一股顽强劲，不达目的，决不休止。在这个搜索树中可以看出，如果按照如下的顺序搜索。便可避开死胡同，直接走上成功之途。例如：

延迟　线　存储器
V　　线　存储器
V　　N　存储器

NP　　存储器

　　N

NP

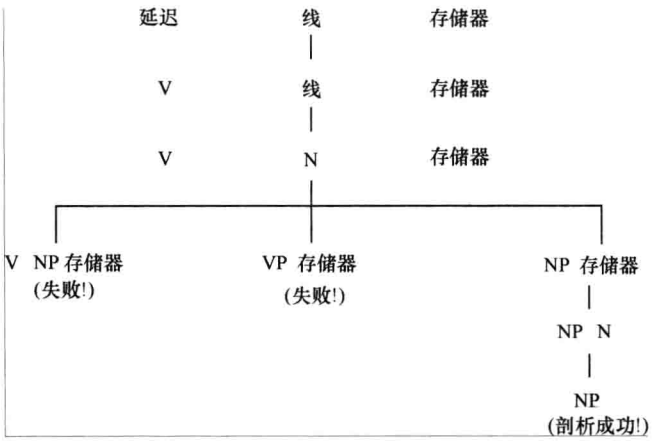

图 5　自底向上剖析的搜索树

用 LISP 语言，很容易就可以把上述的自底向上剖析过程一目了然地写出来，例如：

(延迟　线　存储器)

((V 延迟)　线　存储器)

((V 延迟)　(N 线)　存储器)

(NP((V 延迟)　(N 线))　存储器)

(NP((V 延迟)　(N 线))(N 存储器))

(NP(NP((V 延迟)(N 线))(N 存储器)))

“失灵区部件”的分析与“延迟线存储器”类似，不再赘述。

心理学家金补尔(J. P. Kimball)研究证明，人们在理解自然语言时，总是试图把新出现的词依附到前面与它紧连的组成成分上，把这个词与它前面的一个词联系起来，以便减轻记忆的负担，避免从记忆中搜索有关的组成成分或词汇。 由于使用这样的策略，人们在理解如下的英语句子时往往会感到困惑：

The man offered one thousand dollars for the conference is my uncle.

(为会议提供一千美元资助的人是我的叔父)

The horse raced past the barn fell.

(疾驰过牲口棚的那匹马跌倒了)

人们在开始时往往会把第一句中的 offered 当作它前面的词 man 的谓语，把第二句中的 raced 当作它前面的词 horse 的谓语，等到句子快结束时，才发现这样的理解是错误的，于是回过头去对句子重新进行分析，采取类似于“回溯”的方法，从而得到正确的理解。这种句子称为 “花园幽径句”(garden path sentence)，它正如花园中曲曲弯弯的路径那样，要屡次三番地重复通过。金补尔研究为剖析技术中的回溯机制提供了心理学上的根据。冯志伟深入研究过英语和汉语中的“花园幽径句”，并且使用计算机对于“花园幽径句”进行了成功的剖析。

2．自顶向下剖析(Top-down Parsing)

仍然以“延迟线存储器”这个词组型术语为例来介绍自顶向下剖析。为了便于读者了解思路，以第一人称“我”作为叙述的主体，自顶向下剖析的过程如下：

• 我来找查 NP

*** • NP 由什么组成？

- NP 可以由一个单独的 N 组成
- 所以我得首先找查 N
- N 由什么组成？
- 语法 G 中没有什么规则可以扩展 N
- 单词“线”可以作为范畴符号 N 的一个成员
- “线，区，部件，存储器”这几个单词都是这个词组型术语中开头的第一个词吗？
- 不是，回溯到***

- NP 还可以由什么组成？
  - NP 还可以由一个 V 和一个 N 组成
  - 现在我需要找查 V
    - V 由什么组成？
    - V 可以由“延迟，失灵”等单词组成，其中的“延迟”与 V 匹配
    - “延迟”是这个词组型术语中开头的第一个词吗？
    - 是的
    - 现在我发现 V 是由单词“延迟”组成的，我需要继续找查 V 后面的 N
- N 由什么组成？
- N 可以由“线，区，部件，存储器”等单词组成，其中的“线”与 N 匹配
- “线”是这个词组型术语中开头的第二个词吗？
  - 是的
  - 现在我发现这个 N 是由单词“线”组成的，因此， NP 是由包含单词“延迟” 的一个 V 以及包含单词“线”的一个 N 组成的，
  - 是不是到达词组型术语“延迟线存储器”的结尾了？
  - 没有
  - 哎呀，一定是我做错了什么事
  - 回溯到***处，用另外的办法来做
- NP 还可以由什么组成？
- NP 还可以由另一个 NP 和一个 N 组成
- 另一个 NP 可以由什么组成？
- 另一个 NP 还可以由一个 V 和一个 N 组成
- 现在我需要找查 V
  - V 由什么组成？
  - V 可以由“延迟，失灵”等单词组成，其中的“延迟”与 V 匹配
  - “延迟”是这个词组型术语中开头的第一个词吗？
  - 是的
- 现在我发现 V 是由单词“延迟”组成的，我需要继续找查 V 后面的 N
  - N 由什么组成？
  - N 可以由“线，区，部件，存储器”等单词组成，其中的“线”与 N 匹配
  - “线”是这个词组型术语中开头的第二个词吗？
  - 是的
  - 现在我发现这个 N 是由单词“线”组成的，因此， NP 是由包含单词“延迟” 的一个 V 以及包含单词“线”的一个 N 组成的