时间短语的分析与识别*

刘智颖
北京语言文化大学语言信息处理研究所，北京 100083
liuzhiying@blcu.edu.cn

摘要：时间短语是指描述时间概念的短语。在 HNC 理论中，时间概念大致可分为三种类型：基本时间概念，物化的时间概念，人化的时间概念。并依据语义将时间短语分为四种基本类型：特定时间短语，特殊时间短语，时间的序短语，时间间隔短语。本文给出了各个类型的构成模式，探讨了时间短语的处理策略。

关键词：时间短语　HNC 理论　概念语义网络

The analysis and Recognition of the time phrase

Liu Zhiying
Language Information Processing Center, Beijing Language & Culture University, Beijing 100083
liuzhiying@blcu.edu.cn

ABSTRACT: The time phrase means the phrase that describing the time concept. This paper stated the three types of the time concept and the four basic category of the time phrase, gave the construction models of the time phrase, and discussed the process strategy accordingly.

Keywords: time phrase, HNC theory, semantics network of concept

1　引言

时间短语也就是指描述时间概念的短语。此类短语以时间概念作为核心。时间概念在HNC中是基本概念的一个重要组成部分，时间短语由此也成为 HNC 定义的基本概念短语中的一个重要类型。纵观概念语义网络，时间概念大致可分为三种类型：基本时间概念，物化的时间概念，人化的时间概念。

1.基本时间概念　用符号 j1 表示。包括两个子节点：j11 和 j12。j11 是对时间的序的描述，如“过去”、“现在”、“将来”；j12 是对时间间隔的描述，如“小时”、“分”、“秒”。

以基本概念为中心，时间概念通过挂靠的表示方法，还可以生成两类扩展时间概念：

2. 物化的时间概念　用符号 wj1 表示，其中 w 是拼音 wu 的简写。如：年、月、日、旬、季节等。这些时间概念都是依据日月运转等自然现象而命名的时间概念，所以定义为物化的时间概念。

3. 是人化的时间概念　用符号 pj1 表示，其中 p 是 people 的简写。如：20 世纪 30 年代，“世纪”和“年代”这类时间概念都是与人类活动有关的时间，所以将其归为人化的时间概念，诸如此类的概念还有公元，星期等。

*本文得到国家重点基础研究发展规划（973）项目 G1998030506 和北京语言大学青年研究项目的资助。

研究时间短语对自然语言处理具有极为重要的意义。首先，时间短语是一类重要的信息，它可以服务于文本的信息抽取。把文本中的时间信息从各种纷繁的信息中提取出来，反馈给用户。第二，时间短语也是一类特殊的激活信息。黄曾阳先生指出："由基本概念构成的短语，特别是数量、时间和空间短语需要先行处理，不需要也不应该等到句类检验之后。"（黄曾阳《概念层次网络理论》第 202 页）"HNC 的先验规定是：基本概念短语的局部处理不依赖于句子的全局处理结果。"（黄曾阳《句类分析的 20 项难点》）这就是说，时间短语属于局部处理，它可以独立于句类分析之外，不依赖于对句子进行句类分析的全局处理。时间短语本身是一中自足性的语义结构。这也就为我们单独处理和分析时间短语提供了依据。

2 时间短语的构成

依据语义，时间短语可以分为四种基本类型：特定时间短语，特殊时间短语，时间的序短语，时间间隔短语。每种类型都有其各自的构成方式。下面将进行一一描述。

2.1 特定时间短语

特定时间短语是以特定时间概念为核心部分的时间短语。所谓特定时间概念，就是指时间序列中一些特定的时刻或区间，包括物化的一般时间概念 wj10（如：年、月、日、上午、下午、晚上等）、人化的时间概念 pj10（如：公元、世纪、年代等）。常见的特定时间短语如："10 月 29 号"、"1901 年"、"光绪二十七年辛丑"、"立冬"等等。

特定时间短语中以"天"或大于"天"的时间单位进行时间描述的，为特定时间宏观描述，如：2002 年 10 月 1 日；以小于"天"的时间单位进行描述的，为特定时间局部描述，如：下午 2 点 35 分；宏观时间描述和局部时间描述可以结合起来，构成混合时间描述，如：2002 年 10 月 1 日下午 2 点 35 分。

2.1.1 特定时间宏观表示式

KQ=fpj1;f30j1

KH=Σ(Σj308+wj10-)

K=(Σj308+pj12-)+((j3080+j30811)+pj12-0)

这里，我们用 K 代表时间短语，KQ 代表构成时间短语的前半部分，KH 代表构成时间短语的后半部分。.符号Σj308 是基本数连用的一般表示式。Σ（Σj308+wj10-）是数量连用的特殊表示式，专用于表示特定时间。第一个求和符号Σ是数词与"量词"的连用符号，且"量词"由高位向低位递减。同一求和序列中的量词必须是包含性概念。我们用符号"-"来表示包含关系。

特定时间宏观表示有两种基本方式，传统方式和现代方式。传统方式必须采用 KQ+KH 构成，现代方式省略 KQ。传统方式可选取任一 fpj1（如：光绪、贞观等表示朝代或帝王名号的人化时间专名）为参照系，现代方式则取唯一的 f30j1（即"公元"）为统一参照系，因而它可以省略 KQ。（著作 209 页）

"****年**月**日"的时间短语就是 KH 的展开形式的典型代表。

表示式的第三项 K 的构成是用于专门描述“**世纪**年代”这类特定宏观时间短语的。

特定宏观时间的语言实例及构成见下：

语言实例	构成公式
1999 年 12 月 31 日	KH=(Σj308+wj10-)+(Σj308+wj10-0)+(Σj308+wj10-00)
二月下旬	KH=(Σj308+wj10-0) + wj101c33
3 月 5 号下午	KH=(Σj308+wj10-0)+(Σj308+wj10-00)+wj10-00c21c31
20 世纪 90 年代	K=(Σj308+pj12-)+((j3080+j30811)+pj12-0)
清光绪 27 年辛丑	KQ+KH= fpj1+fpj1+(Σj308+wj10-)+wj10
九七年夏	KH=(Σj308+wj10-)+wj11c42
公元 5 世纪	KQ+KH=f30j1+(Σj308+pj12-)
农历正月十五	KQ+KH=f30j1+(Σj308+wj10-0)+Σj308

2.1.2 特定时间局部表示式

KQ=wj10-00c

KH=(Σj308+wj10-000)+(j308+jzz12-0)

如：下午 3 点 23 分， 上午 8 点左右

微观特定时间也就是指对一日内的时间进行计时。KQ=wj10-00c 用于表示一日内的时间段，如上午，中午，下午，晚上等。这里 KQ 实际相当于宏观特定概念向微观特定概念的过渡。KH 是对时分秒的时间概念进行概括。j12-本身包含“时、分、秒”等概念，它们是一般时间间隔基元，不包含自然现象或人类活动的激活因子。

特定微观时间的语言实例及构成见下：

语言实例	构成公式
7 点 42 分 30 秒	Σj308+j12-+Σj308+j12-0+Σj308+j12-00
五点	Σj308+j12-
五点整	Σj308+j12-+hfK
六点半	Σj308+j12-+jzu41
十一点三刻	Σj308+j12-+Σj308+j12-0
七点过五分	Σj308+过+Σj308+j12-0
差十分钟八点	差+Σj308+j12-0+Σj308+j12

2.1.3 混合表示式

K=K1 + K2

混合表示式是宏观特定时间表示式和微观特定时间表示式的集合。宏观时间表示在前，局部时间表示在后。混合时间表示构成了“年月日点分秒”完整的时间表示链。在这个链中，可以截取其中的一段表示时间，在进行时间表示时中间应当没有跳跃。混合表示式的例子如：明天下午 3 点 20 分。

2.2 特殊时间短语

特殊时间主要是联系于人类活动的时间概念，如节日和纪念日等。如：千禧年，斋月，国庆节，春节，生日等。特殊时间描述在知识表示时应注意要与其所属的时间范围对应起来，

这样可为时间的判定分界提供方便。如："千禧年"就应对应于特定时间概念中的"年"；伊斯兰教的"斋月"这一特殊时间，就应定位到"月"这一层次的宏观时间概念；"国庆节"这一特殊时间，就应定位到"日"这一层次的宏观概念。其隐含的信息是指"10 月 1 日"。如此就能解释今年国庆节是一个时间概念短语，而"今天国庆节"则是一个简明状态句。关于此项知识还可见文章后面关于时间的序描述规则。

2.3 时间的序短语

时间的序短语的核心时间概念可以由 j11(表示时间的序)充当，表示的是一种相对时间概念。它可以与特定时间表示式组合使用。表达式为：j11+K1+K2。K1 为宏观概念表示式，K2 为微观概念表示式。

时间的序短语的语言实例及构成见下：

语言实例	构成公式
今年二月	j11+K1
明天下午两点	j11+K2
周末	pj11
这个星期	l91+pj11
下星期六	j014+pj11c76

另外，时间的序描述短语也可以由表示序的概念词或指代词来决定。如果特定时间中的时间概念前的数词用指量词（由指示代词＋数量词构成）或序（序数词）来替换。即Σj308 也可替换为 l9 类、j0 类概念，且 l9、j0 类概念后可以有 zz 的概念（如"个"），那么这时的宏观时间概念短语就转变为序时间概念短语。语言实例如：

语言实例	构成公式
这个春天	l91+K1
本月 20 日下午	l91+K1+K2
那段日子	l91+j1

值得一提的是，有一类词语，单独构成表示序的时间短语，前后不加其他任何附加修饰成分。这类词如：

从此　不时　近来　目前　从前　如今　新近　以前　以往　眼前　而今　刚才　那时　面前　往常　往年　下回　现时　每常　平昔　此时　此刻　平时　从古至今

2.4 时间间隔短语

时间间隔短语，类似特定时间描述短语，也可分为宏观和微观两类。

K1=Σ(Σj308+wj1)

K2=Σ(Σj308+jzz12-0)

K1 为宏观时间间隔表达式。例如：2300 年；365 天；三个月零四天；两个季度。K1 中的时间概念可以是：世纪、年、月、旬、日（天）、年度、季度等表示时间间隔的概念。

K2 为微观时间间隔表达式。例如：两小时三十分四十秒。K2 中的时间概念可以是小时、

分、秒、毫秒、微秒等表示时间间隔的概念。

在时间间隔短语中，数与时间概念词语之间可插入量词“个”(用 zz 表示)、表约数的“来，多，几”(用表示量的基本概念 j4 表示）等，数量间的组合连接处可插入“零”。

语言实例	构成公式
十多天	j3080+j4+wj1
半个来月	j4+zz+j4+wj10-0
十多个小时	j3080+j4+zz+j12-
二十几分钟	Σj308+j4+j12-0
三个月零四天	j308+zz+wj10-0+零+j3080+wj10-00

时间间隔短语后也可出现表示约数的词语（概念类别符号为 jz4)。常见的有：多，左右。如：半年多；两小时左右

此外，一些词语可单独做时间间隔短语，如：半日，半天，长期，全年，世代，四季等。

3 识别时间短语的资源和策略

3.1 资源配置

对时间短语进行识别所需的资源之一是必须有一个描述时间概念的词语库。并用特定的符号来表示它们，以便计算机能够激活这些符号，进而达到对时间短语的判定、分析和处理。HNC 已经初步建立了时间词语库，并用一套符号对其进行了形式化的描述。如前所述，在 HNC 的概念语义网络里，时间概念大致可分为三种类型：基本时间概念，物化的时间概念，人化的时间概念。基本时间概念下又设了三个二级节点：一般时间（用 j10 表示)、时间的序（用 j11 表示）和时间的间隔（用 j12 表示)；物化的时间概念下也设了三个二级节点，用以表示年代（wj10)、季节（wj11）和年代间隔（wj12)；人化的时间概念下设置了公元（pj10)、周（pj11)、世纪（pj12）三个二级节点。这样，就形成了具有关联性的 HNC 时间概念词语库，用于服务时间短语分析和处理。

资源之二是时间短语的构成模式。文章第二部分已经给出了不同类的时间短语的相应的构成模式。

资源之三是时间短语的前后边界词集合。除了依靠时间短语的构成模式对其进行识别外，对时间短语进行前后边界的确认则是识别时间短语的一大捷径。能够给出时间短语边界信息的可以是充当时间短语前缀或后缀的时间词语，也可以是由逻辑符充当的时间短语外边界。

时间短语可加后缀（用 hfK 表示)，表示模糊特定时间。这类后缀词本身也是时间词。如：初　初期　初叶　早期　中期　中叶　晚期　末　末叶　末期　时分　年间　初年　末年　底　的时候

形成的时间短语如：本世纪初叶　清顺治年间　傍晚时分

在时间短语前后有时会出现一些逻辑符，用以限定时间的表示范围，我们称之为外边界。外边界可以帮助我们判断时间短语的起止位置。

充当前边界的这类词如：

从　自从　从打　自打　早在　在　于　早于　到　至　截至　截止　直到

形成的时间短语如：自从20世纪70年代　于1997年12月　截至去年9月

充当后边界的这类词如：

前　以前　前夕　后　之后　过后　以远　为止　以来　来　起　前后　左右

形成的时间短语如：1月15日前1998年1月1日起　近四十年来

前后边界还可以同时出现，搭配使用。这类词如：

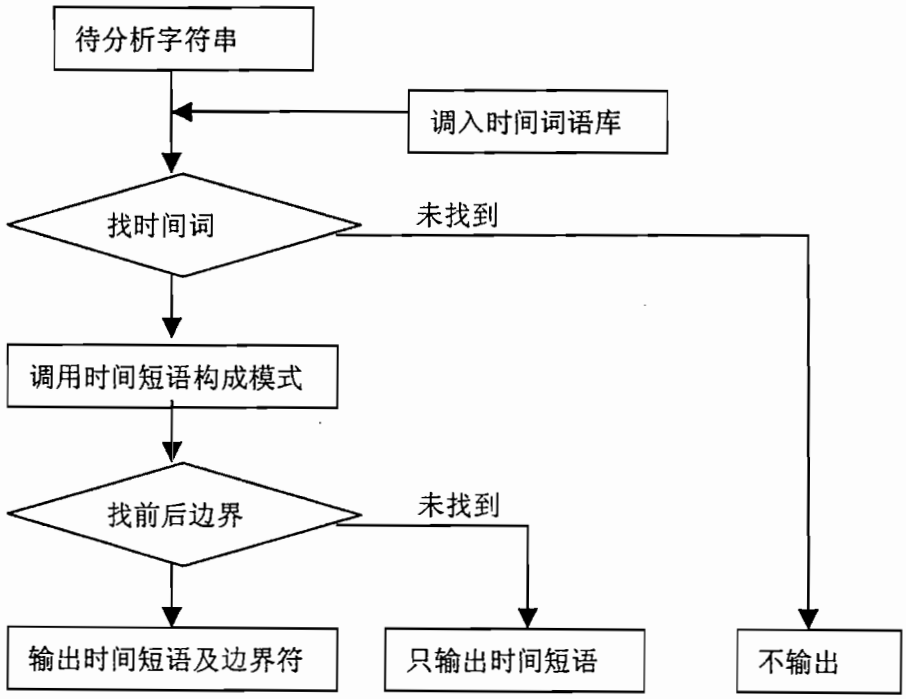
从……起　自……起　自从……以后　自（从）……以来　在……后　在……前　在……之前　早在……之前　在……之际　在……中　在……里　在……同时　在……时　在……时期　从……开始　当……（的）时候　自……至今

形成的时间短语如：　从1998年开始　　在1月15日前　在这一年里

3.2 处理策略

拥有了描述时间概念的词语库，得出了时间短语的构成模式，就可以对时间短语进行识别处理。

处理流程如下：



4 结语

时间短语的识别属于HNC局部性难点处理的一部分。正如黄曾阳先生所说："局部性难点处理需要细致的、技巧高明的、精益求精的长期朴实积累，需要对繁杂语言现象与重复性脑力劳动具有坚韧的耐心。"时间短语的处理还需要做大量的工作，诸如完备时间概念知识库，明晰时间短语的边界信息，由句子充当的时间短语的识别处理等等。笔者将就这些未完成的工作进行更加深入细致的后续研究。

参考文献

[1] 黄曾阳：HNC（概念层次网络）理论，清华大学出版社，1998年11月第一版

[2] 黄曾阳："句类分析的20项难点"，内部资料。

[3] 黄伯荣等主编：《现代汉语》，高等教育出版社，1999年版。

[4] 张全："时间概念短语初探"，《HNC与语言学研究》论文集，2001年10月第一版。

[5] 吕叔湘："时间、地点、数目"，《吕叔湘语文论集》，1963年2月6日。

[6] 陆俭明："说'年. 月. 日'"，世界汉语教学，北京语言学院出版社，1987年。

[7] 陆俭明："现代汉语时间词说略"，第三届国际汉语教学讨论会论文选，北京语言学院出版社。

[8] 陈平："论现代汉语时间系统的三元结构"，现代语言学研究：理论·方法与事实，重庆出版社，1991年。

[9] 赵恩芳："浅谈时间名词"，语文月刊，1989年7月。