面向计算机的二重复句层次划分研究

李晋霞　　刘　云
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
E-mail:lijx666@sohu.com　　E-mail:liuyun@pku.edu.cn

摘　要：多重复句结构层次的自动分析是篇章计算语言学需要解决的一个重要问题。本文依托复句本体研究的现有成果，重点考察复句关系词语的包孕机制对二重复句结构层次自动分析的辅助作用。这种考察对于二重以上复句结构层次的自动分析也有一定的借鉴意义。

关键词：二重复句，复句关系词语，层次划分，自然语言处理

The Computer-oriented Study of Construction Analysis of Two-level Complex Sentences

LI Jinxia　　LIU Yun
Institute of Applied Linguistics　　Institute of Computational Linguitics
Ministry of Education, Beijing 100010　　Peking University, Beijing 100871
E-mail:lijx666@sohu.com　　E-mail:liuyun@pku.edu.cn

Abstract: Based on the existing ontological research of complex sentences, this paper mainly discusses the assistant function of relationship words to the automatic construction analysis of two-level complex sentences in natural language processing. This study is also helpful to the automatic construction analysis of multi-level complex sentences.

Keywords: two-level complex sentences, relationship words, construction analysis, natural language processing

一、引言

多重复句是包含不止一个结构层次的复句，多重复句的层次划分可以看作是两个以上分句之间相互选择、匹配构成不同层级复句模块的过程。如：

（1）由于$_1$工作的需要，我虽然$_2$读过一些语言学的书，但$_3$自知在语言学家跟前仍是一个门外汉，所以$_4$不敢妄评。

这个多重复句有四个分句，引领它们的复句关系词语分别是："由于$_1$"、"虽然$_2$"、"但$_3$"、"所以$_4$"（下标数字表示关系词语出现的先后顺序，下同）。在正确理解多重复句语义内容

的基础上，人们可以准确地对其进行层次划分："虽然 $_2$" 引领的分句与"但 $_3$" 引领的分句首先匹配构成一个低层次的复句模块，这个复句模块又被包孕在"由于 $_1$" 引领的分句中，这个包孕有低层复句模块的"由于 $_1$" 分句又与"所以 $_4$" 引领的分句匹配构成一个更高层次的复句模块。

如果对句义的处理不过关，计算机要完成上述多重复句的层次划分是比较困难的，比如计算机不知道"由于 $_1$" 引领的分句应该管到哪里，从形式上看有三种可能：a 管到"工作的需要"；b 管到"我虽然读过一些语言学的书"；c 管到"但自知在语言学家跟前仍是一个门外汉"。

复句关系词语，是复句中用来联结分句标明关系的词语。在多重复句里，关系词语作为关系标志在不同结构层次上使用（邢福义 2001：26-28）。复句关系词语（以下有时简称关系词语）对多重复句结构层次的自动划分有一定的辅助作用。本文以二重复句为主要研究对象，依托复句本体研究的现有成果，从关系词语的包孕机制出发考察关系词语对二重复句结构层次自动分析的辅助作用。

二、复句关系词语的包孕机制

2.1 二重复句结构层次自动分析的步骤

对于关系词语没有省略的二重复句来说，根据关系词语的常规组配形式以及出现的前后位置有望实现二重复句结构层次的自动分析。具体包括三个步骤：

第一，让计算机掌握关系词语之间的常规组配形式。

复句关系词语数量有限，而且有比较固定的组配模式。让计算机掌握关系词语的常规组配形式是计算机自动完成二重复句层次划分的首要条件。这个条件不难实现，就是通常所说的关系词语之间的配对用法，如：因为/由于……所以……，虽然……但是……，如果……那么/就……，既然……那么……，即使……也……，只要……就……，只有……才……，无论……都……，一方面……另一方面……，首先……然后……，不但……而且……，或者……或者……，要么……要么……，不是……就是……，既……又……，等等。

第二，把紧挨着的能够配对使用的关系词语引领的分句首先匹配构成一个复句模块，这个步骤可称为"最临近配对分句首先构成复句模块"的原则。仍以例（1）为例：

（1）由于 $_1$ 工作的需要，我虽然 $_2$ 读过一些语言学的书，但 $_3$ 自知在语言学家跟前仍是一个门外汉，所以 $_4$ 不敢妄评。

根据第一步，"由于 $_1$" 应与"所以 $_4$" 匹配，"虽然 $_2$" 应与"但 $_3$" 匹配。根据第二步，"虽然 $_2$" 引领的分句与"但 $_3$" 引领的分句之间没有被其他关系词语引领的分句隔开，是紧挨着的关系词语能够配对使用的两个分句，因此"虽然 $_2$……但 $_3$……" 首先匹配构成一个复句模块。

第三，根据关系词语出现的前后位置，确定由第二步得来的复句模块的层次归属。如对于例（1），即要确认"虽然 $_2$……但 $_3$……" 这个复句模块的被包孕情况。根据关系词语的相

对位置，复句模块应被包孕在出现于其前的关系词语所引领的分句中。如例（1）中的“虽然$_2$……但$_3$……”复句模块应被包孕在于其前出现的“由于$_1$”引领的分句中，即“由于$_1$”引领的分句应该管到“门外汉”。至此，再次根据“最临近配对分句首先构成复句模块”的原则，管到“门外汉”的“由于$_1$”分句与“所以$_4$”分句匹配构成更高层次的复句模块。

上述二重复句结构层次自动分析的三个步骤对于关系词语没有省略的二重以上复句结构层次的自动分析也有较为普遍的适用性，如：

（2）这次试验虽然$_1$成功和失败均有可能，但$_2$由于$_3$无论$_4$是成功还是失败，都$_5$会产生不小影响，因此$_6$要及早制定对策。

根据第一步，例（2）中的“虽然$_1$”与“但$_2$”配对，“由于$_3$”与“因此$_6$”配对，“无论$_4$”与“都$_5$”配对。“是成功还是失败”这种紧缩复句由于在形式上内部没有分隔，自动构成一个整体，这里不予讨论。第二步，根据“最临近配对分句首先构成复句模块”的原则，“无论$_4$”引领的分句与“都$_5$”引领的分句之间没有被其他关系词语引领的分句隔开，因此首先匹配构成一个复句模块；同时，由于“无论$_4$”是紧挨着“由于$_3$”出现的，因此“无论$_4$”所在的复句模块应被包孕在“由于$_3$”引领的分句中，即“由于$_3$”管到“因此$_6$”之前。再次根据“最临近配对分句首先构成复句模块”的原则，“由于$_3$”引领的分句与“因此$_6$”引领的分句匹配构成一个复句模块。同时，又因为“由于$_3$”紧挨着“但$_2$”出现，所以“由于$_3$”所在的复句模块应被包孕在“但$_2$”引领的分句中，即“但$_2$”在层次上应该管到句末。仍根据“最临近配对分句首先构成复句模块”的原则，“虽然$_1$”引领的分句与“但$_2$”引领的分句匹配构成最高层次的复句模块。需要说明的是，对于例（2），根据“最临近配对分句首先构成复句模块”的原则，“虽然$_1$”引领的分句与“但$_2$”之间也没有被其他分句隔开，因此在自动分析结构层次时，“虽然$_1$……但$_2$……”这个实质上最高层次的复句模块会被当做最低层次的复句模块首先匹配起来，即：

[这次试验虽然$_1$成功和失败均有可能，但$_2$由于$_3$无论$_4$是成功还是失败]，都$_5$会产生不小影响，因此$_6$要及早制定对策。

“[ ]”表示匹配而成的复句模块。但是，这种分析方法势必割裂“由于$_3$”与“因此$_6$”、“无论$_4$”与“都$_5$”之间相互匹配构成复句模块的关系。即不同层级的复句模块之间通常应该是包容与被包容的关系，而很少是交叉关系。因此虽然根据“最临近配对分句首先构成复句模块”原则，例（2）中的“虽然$_1$……但$_2$……”有可能被分析为最低层次的复句模块，但是这种分析将在接下来的层次划分中受阻，因而将被证明是失败的。

总之，在关系词语没有省略的情况下，上述三个步骤对于多重复句结构层次的自动分析具有较为普遍的适用性。但是，自然语言中多重复句的关系词语常有省略，这时由于标志分句与分句之间相对独立性及匹配关系的固定形式标记——关系词语的丧失，就给多重复句结构层次的自动分析带来了较大困难。如：

（3）如果天下雨，因为路滑，你不来上课，我会原谅你。

（4）如果天下雨，因为路滑，你可以不来上课。

例（3）、例（4）中，复句关系词语有省略，如果补上分别应该是：

（3）如果天下雨，因为路滑，（所以）你不来上课，（那么）我会原谅你。

（4）如果天下雨，（那么）因为路滑，（所以）你可以不来上课。

由于假设复句“如果……那么……”的前后分句都有可能包孕因果复句模块“因为……所

以……”，即“如果（因为……所以……），那么……”、“如果……那么（因为……所以……）”这两种包孕形式的多重复句在现代汉语中都存在，前者如例（3），后者如例（4）。因此，对于例（3）、例（4）而言，依据关系词语进行结构层次的自动分析首先面临的问题就是“因为”代表的因果复句模块是被包孕在“如果”引领的假设复句的前分句中，还是被包孕在关系词语“那么”省略的假设复句的后分句中。可见，在关系词语有省略的情况下，依据关系词语进行多重复句结构层次自动分析的有效性受损。因此，考察多重复句关系词语有所省略的不同情况以及这些情况下关系词语对多重复句结构层次自动分析所能起到的辅助作用，将是进一步需要解决的问题。

2.2 复句关系词语的包孕机制

下面仍以二重复句为例，讨论关系词语有省略时关系词语的包孕机制对二重复句结构层次自动分析的辅助作用。从复句本体研究的现有成果可以看出（金立鑫2000，周刚2001），关系词语的音节形式对关系词语所在分句的包孕能力有一定影响。因此，根据音节形式，可把关系词语的配对形式首先大致分为两类：包含单音节关系词语的配对形式和不包含单音节关系词语的配对形式。下面分别予以讨论。

（一）对于“只要……就……”、“只有……才……”、“即使……也……”、“无论……都……”这些包含有单音节关系词语的配对形式而言，由于单音节关系词语所在的分句在包孕其他复句模块时，单音节关系词语很有可能不再继续使用。因此，对于上述这些复句形式而言，可以根据单音节关系词语的出现与否判断其前后分句是否包孕有低层复句模块，具体包括两种情况：

1. 当单音节关系词语出现时，一般情况下是上述几种复句的前分句包孕了一个低层复句模块，如：

（5）只要$_1$你去，或者$_2$他去，我就$_3$不去。

（6）只有$_1$你去，或者$_2$他去，我才$_3$去。

（7）即使$_1$你去，或者$_2$他去，我也$_3$不去。

（8）无论$_1$你去，还是$_2$他去，我都$_3$不去。

以上二重复句中，“或者$_2$”、“还是$_2$”等关系词语在运用上都省略了与其匹配的引领前分句的关系词语，如例（5）若把未出现的关系词语补上应是“只要（或者）你去，或者他去，我就不去。”又例（8）若把未出现的关系词语补上应是“无论（是）你去，还是他去，我都不去。”上述这些二重复句在进行层次自动分析时，根据第一步，得知例（5）中的“只要$_1$”与“就$_3$”匹配，例（6）中的“只有$_1$”与“才$_3$”匹配，例（7）中的“即使$_1$”与“也$_3$”匹配，例（8）中的“无论$_1$”与“都$_3$”匹配。虽然，上述二重复句中根据关系词语的相对位置无法完成“最临近配对分句首先构成复句模块”这一步，但是根据这些二重复句中的单音节关系词语都出现了的情况可以判定上述“只要$_1$……就$_3$……”、“只有$_1$……才$_3$……”、“即使$_1$……也$_3$……”、“无论$_1$……都$_3$……”这些复句的后分句没有包孕低层复句模块，从而断定“或者$_2$”、“还是$_2$”等所在的复句模块是被包孕在上述复句的前分句中。实际情况也是如此，如例（5）的结构层次可以表示为：

只要$_1$[你去，或者$_2$他去]，我就$_3$不去。

2. 当单音节关系词语没有出现时，一般情况下是上述几种复句的后分句包孕了一个低层复句模块，如：

（9）只要$_1$你去，即使$_2$他不去，也$_3$不要紧。

（10）只有$_1$你去，因为$_2$他听你的话，所以$_3$这件事肯定能办成。

（11）即使$_1$你去，只要$_2$他不听你的话，这件事就$_3$办不成。

（12）无论$_1$谁去，如果$_2$他一意孤行，这件事就$_3$办不成。

上述二重复句中，与“只要$_1$”、“只有$_1$”、“即使$_1$”、“无论$_1$”相匹配的单音节关系词语“就”、“才”、“也”、“都”没有出现。在自动分析结构层次时，如例（9），根据“最临近配对分句首先构成复句模块”的原则，“即使$_2$”引领的分句与“也$_3$”所在的分句首先匹配构成一个复句模块。同时，根据与“只要$_1$”匹配的单音节关系词语“就”没有出现，可以断定是“就”所在的后分句包孕了一个低层复句模块“即使$_2$……也$_3$……”，即：

只要$_1$你去，[即使$_2$他不去，也$_3$不要紧]。

其他几例与例（9）相同。

（二）对于不包含单音节关系词语的配对形式而言，关系词语出现与否与其引领的分句是否包孕有低层复句模块之间没有必然的对应规律。因此，下面着重从一般意义上根据关系词语出现的实际数量讨论不同情况下关系词语对二重复句结构层次自动分析的辅助作用。

1. 四个复句关系词语出现三个，如：

（13）因为如果你不去，事情就办不成，所以你还是去。

对于例（13），四个关系词语出现三个的情况有以下几种：

（13a）因为$_1$如果$_2$你不去，事情就$_3$办不成，你还是去。

（13b）因为$_1$如果$_2$你不去，事情办不成，所以$_3$你还是去。

（13c）如果$_1$你不去，事情就$_2$办不成，所以$_3$你还是去。

（13d）因为$_1$你不去，事情就$_2$办不成，所以$_3$你还是去。

例（13a）中，“如果$_2$”与“就$_3$”首先匹配构成一个复句模块，同时由于“如果$_2$”紧挨着“因为$_1$”出现，所以“如果$_2$……就$_3$……”这个复句模块应被包孕在“因为$_1$”引领的分句中。同时，根据复句关系词语的常规组配形式，可以判断与“因为$_1$”匹配的表示结果的分句应该是没有关系词语引领的“你还是去”分句。例（13c）中，“如果$_1$”与“就$_2$”首先匹配构成一个复句模块，与“所以$_3$”匹配的关系词语“因为”没有出现，但是根据“因为……所以……”复句的前分句可以包孕假设复句“如果……就……”，即“因为（如果……就……），所以……”这种复句包孕形式是存在的，因此可以推断例（13c）中的“如果$_1$……就$_2$……”复句模块是被包孕在与“所以$_3$”匹配的关系词语“因为”没有出现的原因分句中。例（13c）、例（13d）与例（13a）、例（13b）类似。

通过上述分析可以看出，当四个关系词语出现三个时，二重复句仍可借助关系词语完成结构层次的自动分析。由此也可进一步推测，当“n”重复句（n≥2）实有“2n”个关系词语而实际出现“2n－1”个时，凭借关系词语通常仍可完成多重复句结构层次的自动分析。

2. 四个复句关系词语出现两个，如：

（14）因为$_1$如果$_2$你不去，事情就$_3$办不成，所以$_4$你还是去。

对于例（14），四个关系词语出现两个的情况有六种：

（14a）因为$_1$你不去，事情办不成，所以$_2$你还是去。

（14b）如果$_1$你不去，事情就$_2$办不成，你还是去。

（14c）因为$_1$如果$_2$你不去，事情办不成，你还是去。

（14d）你不去，事情就$_1$办不成，所以$_2$你还是去。

（14e）因为$_1$你不去，事情就$_2$办不成，你还是去。

（14f）如果$_1$你不去，事情办不成，所以$_2$你还是去。

根据关系词语的出现情况可将上述（14a）—（14f）分为三种类型：

（1）所出现的两个关系词语可以配对使用，如（14a）、（14b）；

（2）所出现的两个关系词语分属不同层次的复句模块，但二者在所常规组配形式中相对位置一致。如例（14c）中出现的关系词语有两个："因为"和"如果"，"因为"在所常规组配形式"因为……所以……"中位于前置关系词语的位置，"如果"在所常规组配形式"如果……那么……"中也位于前置关系词语的位置。例（14d）中所出现的两个关系词语"就"、"所以"在所常规组配形式中都是后置关系词语。

（3）所出现的两个关系词语分属不同的层次，并且二者在所常规组配形式中相对位置不同。如例（14e）中的两个关系词语"因为"和"就"，前者在所常规配对形式"因为……所以……"中是前置关系词语，后者在所常规配对形式"如果……就……"中是后置关系词语。例（14f）与之相同。

经检验，上述三种情况中，只有第（2）种情况可以依据关系词语实现二重复句结构层次的自动分析。根据经验，对于第（2）种情况，当出现的两个关系词语都是前置关系词语时，从位置靠后的关系词语开始匹配复句模块，当出现的两个关系词语都是后置关系词语时，从位置靠前的关系词语开始匹配复句模块。这样做是为了从管界范围确定性最高的关系词语出发进行二重复句内部复句模块的组配。如例（14c）中，"因为$_1$"与"如果$_2$"都是前置关系词语，因此从位置靠后的关系词语"如果$_2$"开始匹配复句模块。前置关系词语所指示的与之匹配的分句常规情况下位于其引领的分句之后。因此，例（14c）中，"如果$_2$"引领的分句"你不去"首先与"事情办不成"匹配构成一个复句模块，这个复句模块被包孕在"因为1"引领的分句中，进而可以判断句末分句"你还是去"是与"因为$_1$"引领的分句相匹配的结果分句。例（14d）与之同理。

对于上述第（3）种情况，即所出现的关系词语在所配对使用的关系词语中相对位置不同，这时由于不同复句间有可能存在互相包孕的能力，因此仅靠关系词语进行结构层次的自动分析往往会出现多种可能。如例（14f）"如果$_1$你不去，事情办不成，所以$_2$你还是去。"中关系词语出现的情况可用"如果……，……，所以……"表示，这种形式的二重复句，其内部层次有两种可能，如果把未出现的关系词语补上（以"（）"表示），分别是：

如果……，（那么）（因为）……，所以……。

（因为）如果……，（就）……，所以……。

可见，对于在配对使用形式中相对位置不同的两个关系词语参与构成的二重复句来说，由于不同关系类型复句之间有可能存在彼此包孕的能力，因此仅仅依靠关系词语往往无法完成二重复句结构层次的自动分析。

对于上述第（1）种情况，即所出现的两个关系词语能够配对使用的二重复句来说，依靠关系词语无法揭示其"二重"复句的性质，因此也无从谈起二重复句结构层次的自动分析。

总之，对于四个关系词语出现两个的二重复句来说，当这两个关系词语在所常规配对形

式中相对位置一致，那么仍有望凭借关系词语实现二重复句结构层次的自动分析。可以进一步推测，当“n”重复句（n≥2）实有“2n”个关系词语而实际出现“n”个，并且这“n”个关系词语分别代表“n”重复句内部的“n”个不同层次时，这时如果这“n”个关系词语在所常规配对形式中相对位置一致，那么仍有望根据关系词语实现多重复句结构层次的自动分析。

3. 四个关系词语出现一个，如：

你不去，事情就办不成，你还是去。

如果你不去，事情办不成，你还是去。

因为你不去，事情办不成，你还是去。

你不去，事情办不成，所以你还是去。

以上二重复句中，凭借唯一一个关系词语无法识别其“二重”复句的性质，因此也无从谈起二重复句结构层次的自动分析。因此，对于四个关系词语只出现一个的二重复句来说，依据关系词语进行结构层次自动分析的可能性很小。可以进一步推测，当“n”重复句（n≥2）实有“2n”个关系词语而实际出现“n－1”个时，这时关系词语对多重复句结构层次自动分析的作用是有限的。

三、小结

邢福义先生（2001：14-15）指出：在实际运用中，二重复句使用频率很高，三重复句也比较常见；四重复句不多，五重复句更少，六重、七重复句就很罕见了。本文重点讨论复句关系词语对于二重复句结构层次自动分析的辅助作用，可以看出，在关系词语没有省略的情况下，凭借关系词语按照一定的步骤可以实现结构层次的自动分析。在关系词语有省略的情况下，根据关系词语的包孕机制，对于包含单音节关系词语的配对形式而言，通常可以根据单音节关系词语的出现与否判断其前后分句是否包孕有低层复句模块。对于不包含单音节关系词语的配对形式而言，关系词语的不同省略情况将影响其对二重复句结构层次自动分析的辅助效果。多重复句结构层次的划分，是语篇内部不同层级意义相对完整性的体现。寻找语篇内部的不同层级的意义相对完整体，对于语篇的理解无疑是重要的。需要说明的是，虽然在有些情况下依据关系词语可以解决二重乃至多重复句结构层次的自动分析，但是在更多情况下多重复句结构层次的自动分析是仅凭关系词语所无法解决的。本文的价值也许就在于有助于更清楚地看到什么情况下关系词语能对多重复句结构层次的自动分析起到一定的辅助作用，从而也对其局限性有更为清楚的认识。

参 考 文 献

[1] 金立鑫，《语法的多视角研究》，上海外语教育出版社，2000 年，127-141。

[2] 邢福义，《汉语复句研究》，商务印书馆，2001 年。

[3] 张仕仁，“汉语复句的结构分析”，《中文信息学报》，1994 年第 8 卷第 4 期，43-54。

[4] 周　刚，“关联成分的套用及其省略机制”，《汉语学习》，2001 年第 6 期，29-40。