

研究生课程作业

课程名称： 翻译技术原理与实践

小组选题： 第四题

小组成员： 历珵 张欣童 刘茹

陆韵 董永佳 裴佳闻

二〇一八年十月

**This page intentionally left blank**

# 目录

目录 III

1. 翻译单元概述 4

1.1翻译单元的定义 4

1.2翻译复用单元的分类 4

2.亚句级翻译记忆复用单元——单词 5

2.1 词的定义 5

2.2 词作为翻译单位 5

2.3缺点 6

3. 亚句级翻译记忆复用单元——语块 6

3.1语块在语言中的重要性 6

3.2语块的定义 7

3.3语块的分类 7

3.4语块作为翻译单位 8

3.5 优点 8

3.6 缺点 8

3.7 技术难关 9

4. 亚句级翻译记忆复用单元——语块 9

4.1子句在语言中的重要性 9

4.2子句的定义 10

4.3子句的分类 10

4.4 子句作为翻译复用单位 11

4.5 优点 11

4.6 缺点 12

4.7 子句的句法分析框架——X-阶标理论 12

4.8 技术难关 15

参考文献 17

# 1. 翻译单元概述

## 1.1翻译单元的定义

在翻译过程中，译者需要将原文切分成片段，然后逐个片段为单位进行翻译，因此我们就有了“翻译单位”这一概念。设立翻译单元的目的是使译文与源文取得语意对等、句法对等和语用对等的最佳平衡。

对于翻译单元这一定义，国内外学者给出了不同的答案，法国学者Jean-Paul Vinay和Jean Darbelnet将翻译单位定义为“思维单位”，它在语言上表现为“最小的、无需分开翻译的言语单位”。（罗，1984）

前苏联翻译学理论家巴尔胡达罗夫(1969)在《Levels of language hierarchy and translation》一书中，将翻译单元定义为“译出语言作品中存在，而在译入语言作品中有对应物的最小语言单位。”且提出，任何形式的语言学单位，无论大小，都有可能成为翻译单元。这些单元可以像音素或词素一样小，也可以像整个文本一样大。

Shuttleworth and Cowie将翻译单元定义为，在源文本（Source Text）向目标语（Target Language，TL）转化的过程中，一个语言学层面的概念 。

德国语言学家Werner Koller将巴尔胡达罗夫关于翻译单元大小的想法与所涉及的两种语言之间的差异联系起来。认为源语和目标与的差异越大，翻译单元就越大。

理论翻译学家Irma Sorvali将翻译单元定义为：译者在进行后续内容翻译之前，一次性集中注意力翻译的部分（Thunes，2017）。

## 1.2翻译复用单元的分类

Halliday的系统功能语言学对英语进行了系统的分析，将语言结构分为：Morpheme（词素）

Word（单词）

Group（意群）

Clause（子句）

Sentence（句子）

在《翻译教程》一书中，Peter Newmark参考Halliday的理论，提出句子是天然的翻译单元（1988，p.65）,他认为，在句子中，还有五种其他的亚句级翻译子单元，其中词素是最小的翻译单元，除此以外，从语法上来看，翻译单元可以是还有子句（clause），意群（group），从词汇学上看，翻译单元还可以是短语（collocation）和单词（word）（包括习语和复合词等等）（1988，p.66）。

然而，国内在这方面的研究并没有跟上国际的步伐，进二十年来，随着翻译理论的不断引入，罗国林（1986）主张四个层次划分单位：词素、词、短语和句子；罗选民主张把翻译单位定位为小句（1992）；也有学者秉承着翻译单元的动态观，认为翻译单元在同一篇文章里也是可以不断变化的。

通过阅读相关文献，我们发现，翻译单元大体上分为三类，即“亚句级”，“句子级”和“超句级”。本组的研究以亚句级为立足点，总结国内外研究，认为亚句级翻译复用单元可以是音素（phonemes），词素（morphemes），单词（words），短语（phrases），语块（chunk）和子句（clause），并针对单词、语块、子句作为翻译复用单元的可行性、可靠性、优缺点进行了重点调研。

# 2.亚句级翻译记忆复用单元——单词

## 2.1 词的定义

词是指具有一定的整体意义和固定的语音形式、能够独立运用的最小的语言单位。词是语言的基本符号，它是语言中最小的自由的内容和形式相结合的实体单位。词不仅是基本的词汇单位，而且也成为语法的基本单位，语法中无论句法和词法，都以词为必要的基础。语言系统内的各种单位里面，词是居于中心地位的，作用最为突出。

## 2.2 词作为翻译单位

1.第一代机器翻译系统

第一代机器翻译系统使用的是word-level的翻译单元。这些系统是使用所谓的直接翻译策略设计的，将源文本中的单词直接映射到目标语言中的单词。根据特定的语法规则重新排序以适应源语言和目标语言的结构差别。

2.统计翻译系统

1980年代末IBM首次开展统计机器翻译研究，IBM 的统计机器翻译方法的基本思想是，把机器翻译看成一个信息传输的过程，用一种信源信道模型对机器翻译进行解释。假设一个目标语言句子 e，经过某一噪声信道后变成了源语言句子 f，也就是假设源语言句子 f通过某种编码后得到了目标语言句子 e，而翻译的目的就是要将 f 还原成 e。并提出了统计机器翻译的基本公式。

3.1993年IBM提出五种基于词的统计翻译模型IBM Model 1-5

Mode 1将句子切分成词然后进行统计，只考虑单词之间的对译概率，不考虑语序。

Mode 2 考虑了翻译过程单词的位置变化

Mode 3 考虑了一个单词翻译成多个单词的情况，引入了产生概率。

Mode 4 在对齐时不仅仅考虑词的位置变化，同时考虑了该位置上单词的词类

Mode 5 修正模型4，消除了模型4中的缺陷，避免了对不可能出现对齐给出非零的概率

## 2.3缺点

1.第一代翻译系统

从大量翻译实践和理论可知把词作为翻译单位的逐词翻译 ,其操作的层次偏低 , 或歪曲原意 ,或行文不畅 ,往往造成死译或形式主义的翻译。

在不同语言的词汇中，意义能一一对应的词主要是一些专业术语。在大多数情况下，某语言中一个多义词甲的意义与另一语言中相应的乙词的意义往往是互相错综交叉的，它们可能在某个或某些义项上是相同的，而在其他几个义项上却不同，如汉语的“运动”和英语的“movement”，它们有相同的义项，如“表物体运动、社会运动等”，但是也有不同的义项，如汉语可指“体育运动”，英语可指“粪便”等。也可能甲词的全部义项包含于乙词之中，只和乙词的部分义项相同，如汉语的“猫”和英语的“cat”，后者包含前者的全部义项，而“心地恶毒的女人”这个义项是前者所不具有的；也可能某词或某个义项在其他语言的词汇中根本找不到对应的词或义项，例如一部分音译词就是在这种情况下产生的；另外，词语的附加意义或附属色彩，语言之间往往也难以直接对应。

因此，词对词的直接翻译策略很难提供高质量的译文。

2.IBM 词对齐模型

IBM提出的统计机器翻译基本方程式对于机器翻译领域具有开创性的意义，但从理论上讲，IBM 的几个模型只考虑了词与词之间的线性关系，没有考虑句子的结构。对于语序差别比较大的语言对而言，效果可能会不大好。而后基于短语的统计机器翻译方法对此进行了改进。

# 3. 亚句级翻译记忆复用单元——语块

## 3.1语块在语言中的重要性

语块作为一类比较稳定、明确的信息存储单位，在语言中占据重要的位置，研究者们从很早起就开始对语块进行研究。Becker（1975）发现语言的记忆和存储、输出和使用不是以单个词为单位，而是以固定或半固定模式化的语言板块结构来进行，这些板块结构构成人类交际的最小单位。而Sinclair（1991）则通过语料库研究，揭示那些出现频率高、不同程度词化的语块是英语的基本语言单位。由此可见，将语块作为翻译单位具有一定的理论基础，符合前人的语言研究成果。

## 3.2语块的定义

Wray（2002: 9）把语块定义为“一串预制的连贯或者不连贯的词或其他意义单位，它以整体形式存储在记忆中，使用时无需语法生成和分析，可直接整体提取使用”。

常宝宝（2002）所提出的多词组合单位（multi-word unit），是指源语言或目标语言文本中，稳定共现、并且具有合理句法结构的多个单词的组合，在某种程度上即指语块。

姜柄圭等人（2007）在”面向机器辅助翻译的汉语语块自动抽取研究“中，将文本中比较固定的、重复出现的短语，包括名词短语、动词短语等，作为多词组合的语块（multi-word chunk）。这些语块不局限于词语的长度，包含从2-gram到10-gram的短语，只要在句子中稳定共现、语义比较完整，就视为一个合法的语块。

王立非和张岩（2006）在一项基于语料库的语块使用模式研究中，将语块定义为计算机可自动检索出的以相同形式反复出现的两词或多词的有意义连续词组单位。

虽然不同的研究者对语块的定义有所不同，但结合以上定义，可总结出语块的三大特点：结构相对固定，反复出现；语义比较完整，句法合理；易于提取。

在自然语言处理方面，语块的定义有所差异。

国际会议CONLL-2000将语块作为共享任务，其采用的语块描述框架定义了英语中的11中基本语块，即常见的基本短语，包括NP，VP，PP，ADJP，ADVP，SS等（Sang & Buchholz，2000）。这些基本语块的特点是不相交且无嵌套，也就是说一个词只能属于一个语块并且每一个语块内部都不包含其他的语块。

在此基础上，李素建等人（2002）提出汉语语块是一种语法结构，是符合一定语法功能的半递归短语。每个语块都有一个中心词，并围绕该中心展开，以中心词作为语块的开始或结束，并且任何一种类型的语块内部不包含其他类型的语块。

## 3.3语块的分类

Nattinger和Decarrico（1992）从形式结构角度出发，将语块分为四类：

（1）多元词语块（polywords）：由几个词构成，形式固定而充当一个词组来使用的短语，例如“by the way”；

（2）习俗语语块（institutionalized expressions）：作为语言单位存储在记忆中的句子语块，包括谚语、格言以及社交套语等约定俗成的表达形式，例如“long time no see”；

（3）短语架构语块（phrasal constraints）：其空格可由有聚合关系的同类词语填充，具有很强的语言生成性，例如“as I was saying/mentioning”;

（4）句子构建语块（sentence builders）：给整个句子提供结构框架，语块中可以插入其他成分以表达完整的语义，例如“My point is that…”。

Lewis（1993）从结构和功能角度出发，将语块分为复合词和短语（complex words and polywords）、搭配（collocations）、惯用话语（institutionalized utterances），以及句子框架和引语（sentence frames and heads），共计四类，与Nattinger和Decarrico（1992）的分类具有一定共通性。

## 3.4语块作为翻译单位

相关研究发现，把语块作为翻译单位，在人工翻译中有着很大的影响。

翻译单位的大小对于不同程度的译者而言具有不同的认知难度，Loerscher（1991）在对译者操控翻译单位能力的研究中发现：普通学生翻译时使用的翻译单位往往较小，大部分以词为单位；而职业译者的翻译单位均大于词，包括短语、句子和段落。

郑冰寒和谭慧敏（2007）在对英译汉过程中翻译单位进行研究时，同样指出翻译初学者实际操作的翻译单位较小，而且关注的语言单位长度也较为有限，因此过大的翻译单位会对其造成较重认知负担。在同一研究中，他们还发现绝大部分译者的翻译单位为词、词组、小句和句子。

袁卓喜（2009）在“试论语块在翻译过程中的作用与启示“中，提出翻译的输出（presentation）过程是一个由记忆把单个语言信息组成更大单位信息的加工过程，即美国心理学家Miller和Selfridge（1950）所说的组块（chunking）过程。而在外语学习中积累预制语块，并在翻译过程中提高语块意识，将有效避免大脑复杂运算，使更多的翻译问题通过语言浅层处理，自动化地得以解决，即翻译信息的“无标记处理”。

基于语块在人工翻译实际过程中的影响，将语块作为翻译记忆的划分单元，符合译员的工作需求和认知限制，可以大大减轻译员信息处理的压力，并提高译员语言处理的效率。

## 3.5 优点

语块较单词级别的单位而言，对上下文的依赖性更低，出现歧义的现象更少。在翻译过程中，单词可能有很多义项，很难在缺乏相关语境的情况下找到合适的对译词。但是当多个单词组成的语块作为一个整体来进行翻译时，歧义现象相对较少，也更易得出译文（姜柄圭 et. al，2007），例如“打个电话“和“打个乒乓球”能够解决“打”的义项选择。

语块较句子级别的单位而言，结构形式更加灵活、重复出现频率更大，因此在使用翻译记忆的过程中，能够进行较为精确匹配的可能性更高。此外，语块所包含的信息较少，可以有效避免大脑复杂运算，减轻译员的处理压力和认知负担。

## 3.6 缺点

语块边界的划分问题：相较于句子划分，语块的边界较为模糊，更加难以划分和提取。目前的技术还不能准确高效地彻底分析所有的语块。

语块内部结构的合法性问题

语块存储和处理问题：相较于句子划分，语块划分之后数据量剧增，难以存储和处理。

语块复用问题：语块作为翻译记忆提取后，应该如何指导译员有效创作出通顺、准确的译文。

## 3.7 技术难关

N-gram串频统计算法 + 统计关联度法

统计关联度分为：

内部统计值：互信息统计值，X2，对数可能性分值

外部边界法：最大熵

（姜柄圭，2007）

除重复的语块外，还要抽取重复的语言结构，包括词性序列的组合、特殊句式等等。这些工作会有助于建造翻译的模板，提高语句匹配的效率。

# 4. 亚句级翻译记忆复用单元——语块

## 4.1子句在语言中的重要性

子句是复合句的一个部分。与语块和句子相比，子句有其独特的优点：与语块相比，子句更为复杂，内含更多的信息，处理难度更大，不管是从复用单元的长度还是翻译难度上而言，子句都比语块更具整体性。尽管句子比子句更为完整，且句子匹配是似乎是最为有用的匹配方式，但是以句子作为翻译复用单元，我们难以得到相似度高的匹配结果，对于句法难度高的句子来说，在记忆库中匹配就显得更为困难。与句子相比，子句的结构更为简单，更适合放在翻译记忆库中以备译员进行复用。如何对子句进行预处理（pre-processing），会极大的影响翻译记忆库里的语料检索结果。预处理进行的全面，会帮助系统匹配出更多的搜索结果。Katerina和Ruslan（2015）认为，以亚句级翻译单元中的子句作为翻译单元，可以显著提高语句在翻译记忆库中的匹配度。这使得译者能够充分利用更多的之前做过的翻译成果，并且形成一种统一的翻译风格。在现代翻译市场中，翻译需求量巨大，翻译时效性高，且对翻译质量和翻译效率的要求也颇高。因此，以子句作为翻译复用单元，可以极大提高翻译生产力，提升翻译效率。由此可见，以子句作为翻译复用单元，好处多多，且有自己独特的优点存在，是值得研究的课题。

## 4.2子句的定义

子句是从句的旧称，在英语中，子句是复合句的一个成分，有自己的主语和谓语结构，其中谓语一般是主谓短语，由动词和宾语等其他修饰成分构成。子句与词组有明显的不同主要就在于词组一般不包含主谓关系。子句是构成一个完整命题的最小语法单位。简单的来说，子句就是句子里面的小句子。比如说：I learn that you are going to Peking University. 其中，Peking University就是一个子句。

## 4.3子句的分类

子句主要可以分为两大类，分别是独立子句(independent clauses)和从属子句(dependent clauses)，从属子句，根据在句子中充当的不同成分，又可以划分为：名词子句、形容词子句、副词子句。其中形容词子句，又可以划分为限定子句(restrictive clauses)和非限定子句(nonrestrictive clauses)：

（1）独立子句：独立主句可作为单独的句子存在，当它们与其他子句分离后，通常可以视为句子。例如：I don’t feel like eating this apple because I am full. 在这个句子中，有两个子句，分别是I don’t feel like eating this apple和because I am full. 其中I don’t feel like eating this apple，当拆分开来，都可以视作独立句子。

（2）从属子句：从属子句不能单独存在，他们必须与独立子句结合。根据从属主句在句子中担任的成分，以及他们在句子中的功能，我们可以将从属子句分为三类，分别是：名词子句、形容词子句、副词子句。

• 名次子句：名次子句在句子中的作用等同于一个名词，具有名词所有的功能。例如：That he didn’t pass the driving test made him disappointed. 在这个句子中，That he didn’t pass the driving test是一个名次子句，在句中充当名次作用。

• 形容词子句：形容词子句在句子中的作用等同于一个形容词，具有形容词所有的功能。比如，I like the girl who is beautiful.在这个句子中，who is beautiful是一个子句，作用等同于在girl的前面放形容词beautiful.

a. 限定子句：也被称必要子句，它的存在对句意表达很重要，对主句起到限定的作用。

b. 非限定子句：也被称为非必要子句，他可以从句子中移除而不会改变句子的基本含义，对主句起到补充的作用。非限定子句通常会以一个逗号与句子的其他部分隔开，或者是前后各用一个逗号，与句子的其他部分隔开。The girl, who used to dance every day, is now in love with a handsome man. 其中who used to dance every day是非限定子句，用逗号与其他部分隔开。且该剧对主要语意没太大影响，只是起到对the girl进行补充说明的作用。

• 副词子句：形容词子句在句子中的作用等同于一个形容词，像时间子句、条件子句等都是副词子句，用来修饰整个句子。

## 4.4 子句作为翻译复用单位

翻译单元的划分，主要看其形式、语义和语篇上的作用（苏&丁，2009）。对于翻译单元该如何划分，学者们意见不一，大体分为三大类划分单位“亚句级”，“句子级”和“超句级”。亚句级单位又分为“单词”，“词组”，“语块”和“子句”等等。通过阅读文献，我们发现现阶段，以“子句”作为翻译记忆复用单元的相关论文较少，国内文献多以“语块”作为翻译记忆复用单元，还有学者秉承着翻译单元的动态观，认为“应该以复用程度为标准对转换单位进行提取、存储和检索”。Timonera和Mitkov认为，将句子拆分为子句，可以极大地提高相似度匹配的效率。实际上，通过阅读文献，我们组发现，子句作为翻译复用单位，有其独特的优势，但是若以子句作为复用单位，也有很多技术难关需要攻克。

## 4.5 优点

作为一个亚句级翻译复用单元，子句可以克服很多“句子级”翻译复用单元在使用中会出现的问题，比如整句模糊匹配会过滤掉很多我们可能会复用的内容，Macklovich和Russel在2000年提出，句子级翻译复用单元在进行相似度匹配时，会出现因无法识别出句子中部分结构的相似而导致整句的相似度匹配过低的情况，并举了一个例子，如下：

（a） w1 w2 w3 w4 w5 w6…w20

（b） w1 w2 w3 w4 w5 w21…w35

假设我们在翻译记忆库里存在句子（b），句子（b）包含了35个单词。这两句中，有5个共同的单词，而这5个单词只占句子总数的1/4。当我们需要翻译（a）小句时，对（a）进行相似度匹配时，匹配率可能仅为25%，低于模糊匹配的翻译单元复用下限。而采用子句作为翻译复用单元，每个子句都会单独匹配，子句的匹配度就会显著提升（2015）。

与单词、短语以及语块相比，以子句作为翻译复用单元也具有其独特优势，相比于其他亚句级翻译复用单元，子句的形态更加完整，更接近一个句子。将词作为默认的基本翻译单位，在消除歧义和语境联系方面仍然存在不足（常宝宝，2002）。译者在使用时一次匹配的相似译文长度更长，语意更完成。会极大地提高译者翻译的速度和效率，节约成本，增加翻译产出，提高译文质量。

## 4.6 缺点

句子形态学分析问题：将子句作为翻译复用单元，保存在翻译记忆库中，最先面临的问题是句法结构分析的问题。Planas和Furuse（1999）认为，如果翻译记忆库无法做形态学分析（morphological analysis），那么在句子相似度识别上就会出现问题。他举了一个例子：

（1） The wild child is destroying his new toy.

（2） The wild chief is destroying his new tool.

（3） The wild children are destroying their new toy.

在这个例子中，翻译记忆系统如若缺少形态学分析，便会认为（3）与（1）的句子相似度要小于（3）与（2）的句子相似度。

子句因为不是独立存在的，子句与它的标志词关系十分密切，而进行子句划分时往往会将子句与它的标志词分开。同时，某些含有标句词的子句，其标句词往往具有两种或两种以上的含义，而仅仅从子句本身无法看出标句词的具体含义。而且对于定语从句来说，就其在文中的位置来说，汉语与英语差异过大，如何灵活的分配定语从句的位置也是困难。因此，如果单独将子句提出来进行复用，可能会对翻译造成许多困难。

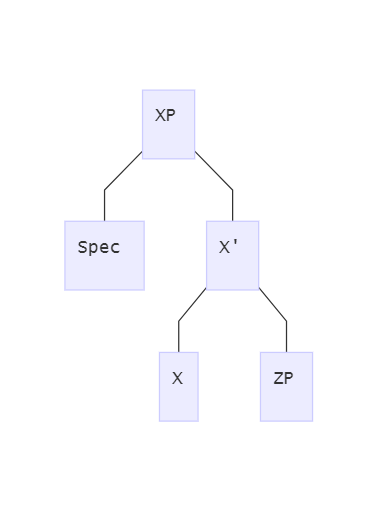
数据储存上的困难，不同于句子级别的翻译记忆，将句子划分为子句会将翻译记忆库的体积增大数倍。

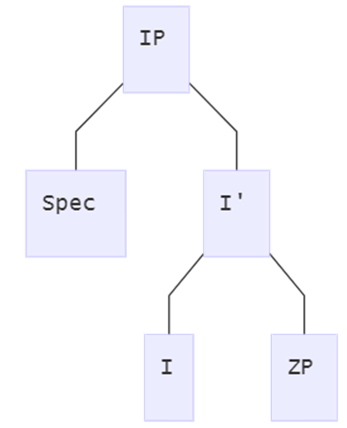
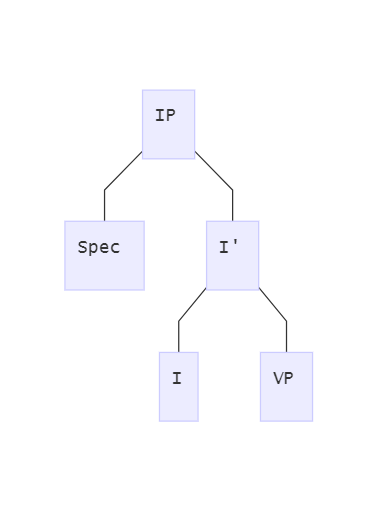
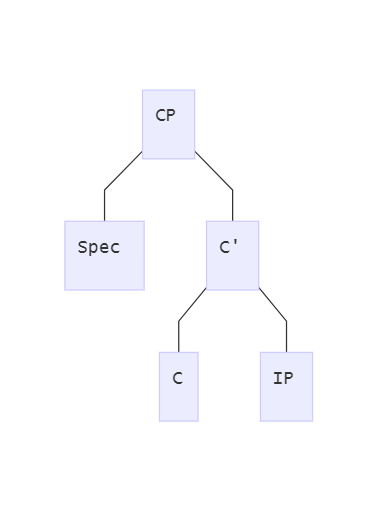
不同于语块的逐字逐句匹配，因为子句仍然具有较为“完整”的句子结构，在与待翻译文本进行匹配时，如何确定匹配的相似度，所以需要对待翻译文本也进行划分这样才可以有效的匹配。

除此以外，从译员的认知角度来看，子句比短语，语块结构更为复杂，长度更长，因此如果翻译记忆系统匹配出一个相似度较低的译文，译员需要自主判断该译文是否可用或如何修改之后可用，翻译复用单元长度的增加，结构的更复杂，可能会增加译员的认知负担。

## 4.7 子句的句法分析框架——X-阶标理论

X=N,V,A,P或Adv

****分句结构也符合X-阶标图示，有的分句是标句词短语CP，有的是屈折语短语IP：

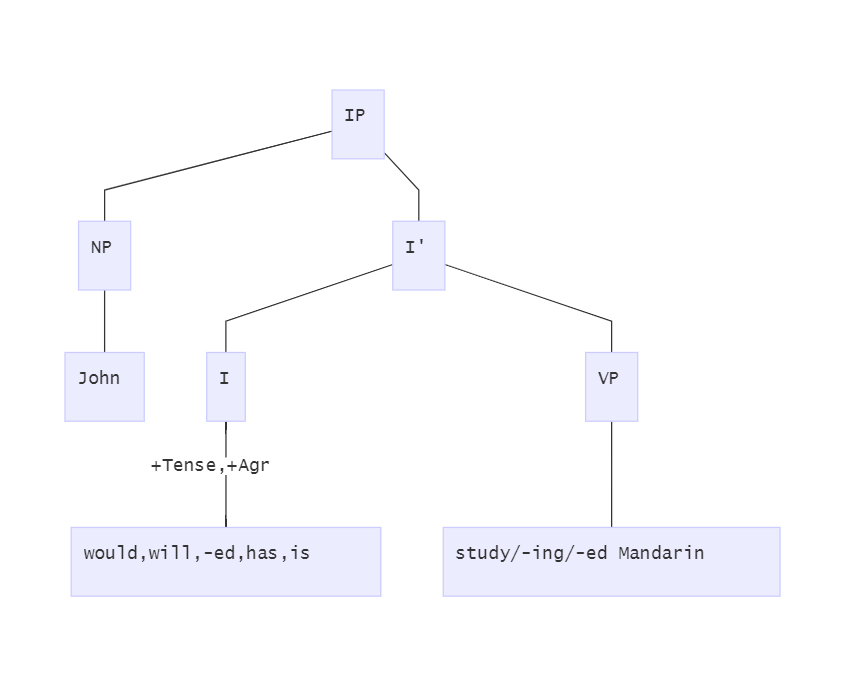


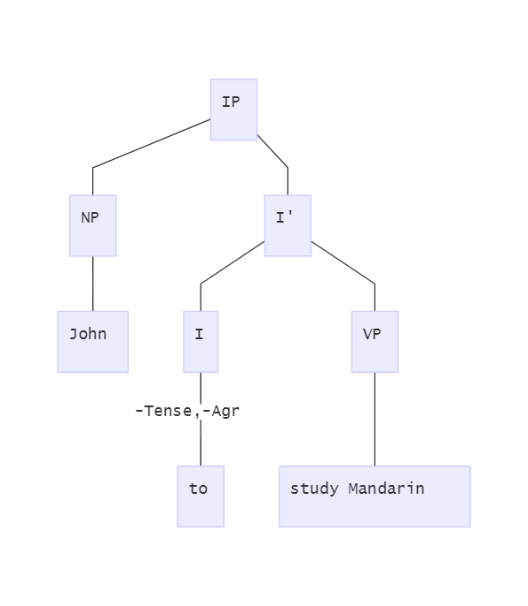
ZP=NP,AP或PP

五种实义短语结构NP,VP,AP,PP和AdvP以实义语类为中心语，称为实义性投射；分句结构CP和IP以功能语类为中心语，称为功能性投射。

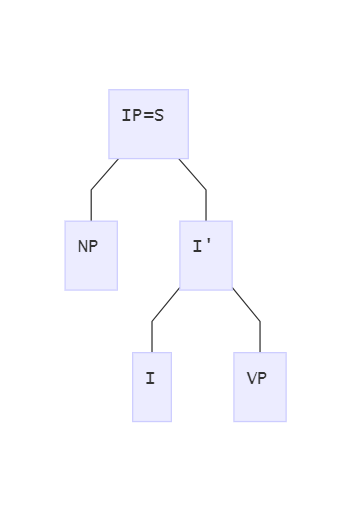
（1）非标句词分句

would、will、和-ed都可以视为表示时态的语素，区别在于前两者是自由语素，后者是粘着语素。因此将这些语速进行抽象概括，都看作动词的屈折变化形式，并成为屈折词，记作INFL或I。is和has既有体态意义也具有时态意义，因此也具有[+Tense]的特征，此外这两个词还表示主语和动词在数和人称上的一致关系，具有[+Agr]的特征。所有的屈折词都具有一致意义。





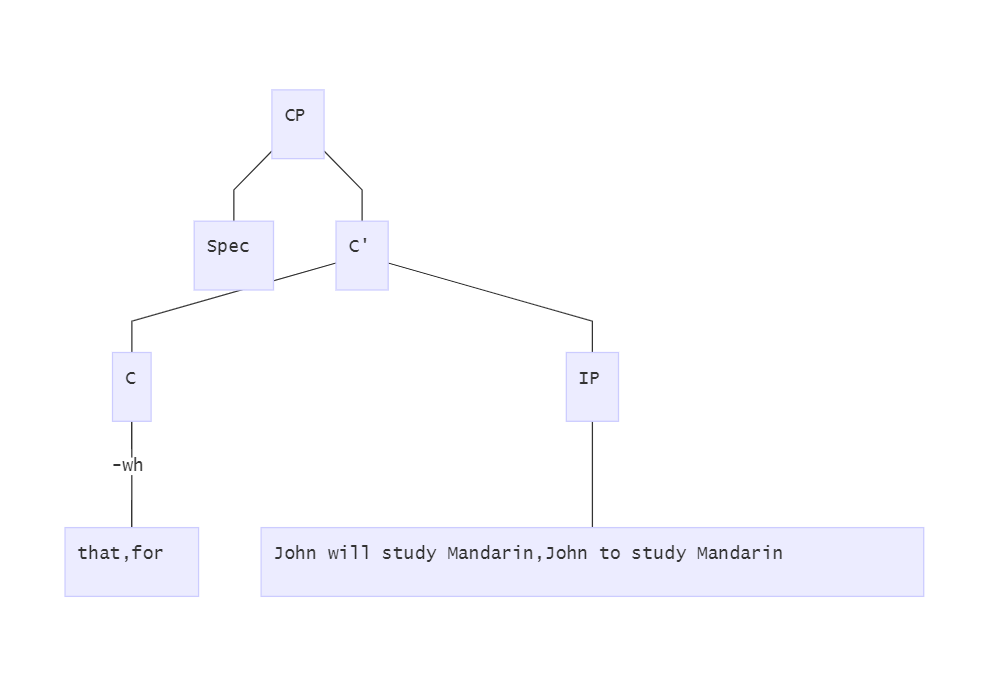
而对于不定式小品词to，其与助动词有同样的分布。因此也可以视其为一种屈折词。不定时分局与时态分句的唯一区别就是前者没有时态变化和人称、数上的变化；也就是说，时态分句的中心于有[+Tense]和[+Agr]的特征，而不定式分句的中心语to缺少这些特征，或者说具有[-Tense]和[-Agr]的特征。

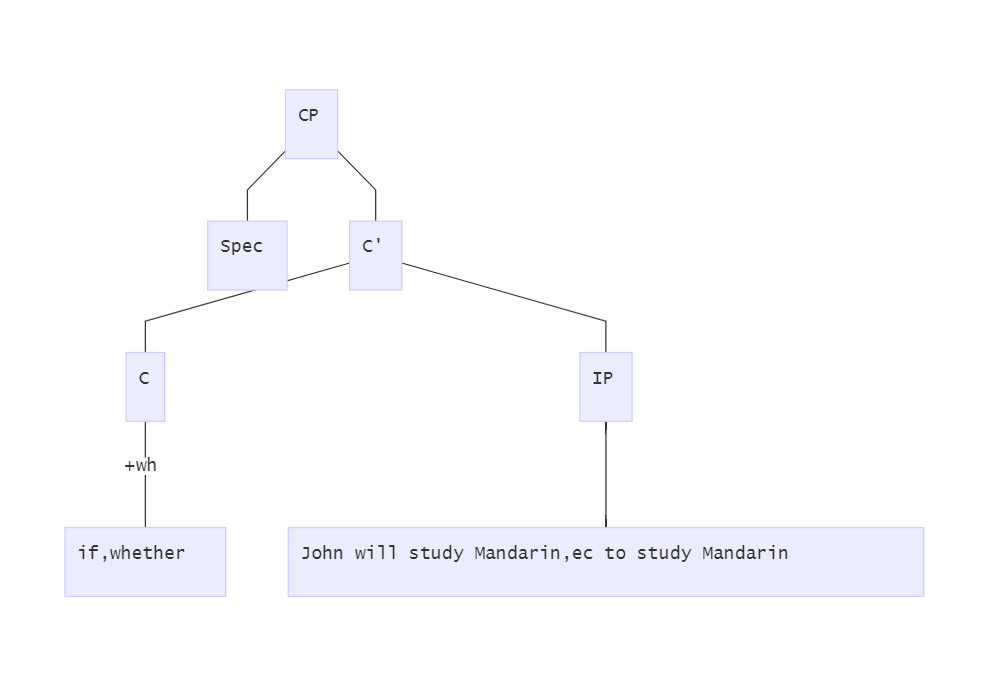
 综上所述，一个非标句词分句，无论用作独立分句还是嵌入分句，其结构都可以用X-阶标图示来描写：这种分句实际上是一个IP。中心语I与补语VP一起构成I’；主语NP处于标志语位置，与I’一起构成最大投射IP 非标句分句结构

在生成语法中，主语被定义为IP的标志语，即在[Spec,Ip]位置上出现的成分，或者说在I’的姐妹节位置上的成分。

（2） 标句词分句

上面把非标句词分句分析为IP。现在分析标句词分句。按照我们的定义，标句词分句由标句词与后接分句构成。(a)They say [that[IPJohn will study Mandarin]] (b)It is difficult [for[IPJohn to study Mandarin]] (c) I wonder [if/whether[IPJohn will study Mandarin]] (d)John has not decided [whether[IP ec to study Mandarin]] 从时态上看，that，if，wehther选择时态分句，for选择不定式分句。另外，如（d）所示，whether也选择以空语类（empty category；ec）充当主语的不定式分句。从意义上看，that和for引导的分句具有陈述意义，if和whether引导的分句具有疑问意义。之所以有这种差别，是因为标句词本身具有不同的语义特征。这说明，标句词在整个标句词分句中起着决定性作用。因此，标句词分句可以视为以标句词为中心与的短语，即标句词短语（CP）





(3)小句

小句可以分析为隐性的屈折词I为中心语的IP。构成小句的两个成分必须在性和数上保持一致，以此为根据，可以把小句的中心语视为一个隐性的屈折词I，该屈折词支配性[gender]和数[number]两个特征（Haegeman and Gueron）

## 4.8 技术难关

由于子句与其他亚句级翻译复用单元相比，结构较为复杂，信息较完整，长度较长，因此，对子句的处理的过程也更为复杂，与机器翻译所需的技术有很多重叠的部分。如果利用子句作为翻译复用单元，我们需要利用一系列技术来提升子句的相似度匹配水平。这些技术包括：句子形态学分析、字符串编辑距离的算法、命名实体的识别等等。

**• 子句的识别：句子形态学分析**

**• 单词间的编辑距离（Edit Distance）**，目前编辑距离最常用的是字符串相似度测量算法（刁兴春等，2010），通过计算单词间的编辑距离可以帮助我们提高相似度匹配效率。在单词间编辑距离的计算过程中，会利用到基于N-Gram的聚类算法，此算法能适应单词的常见拼写错误，从而更好地聚类相似重复记录。Pair-wise的比较算法，以单词间的编辑距离为基础，通过计算量记录中，单词见的编辑距离来判断记录相似与否（邱越峰等，2001）

• **命名实体的识别（named entity）**：命名实体是文本中的固有名称，包括时间，日期，机构，人名，机构名的识别。命名实体的识别对于信息抽取、机器翻译、问答系统、句法分析等具有重要作用。对于命名实体的识别，主要有两种：分别是机遇统计的方法和基于规则的方法。基于统计的方法主要有隐马尔可夫模型（Hidden Markov Models, HMM）、最大熵模型（Maximum Entropy Models）、条件随机场（Conditional Random Fields）等等：（向晓雯等，2005）

# 参考文献

Becker J. The Phrasal Lexicon [M]. Cambridge Mass: Bolt and Newman, 1975.

Erik F. Tjong Kim Sang & Sabine Buchholz, (2000) Introduction to the CoNLL-2000 Shared Task: Chunking [P] Proceedings of CoNLL-2000 and LLL-2000, p127-132, Pottugal, 2000.

Martha Thunes. (2017).The concept of ‘translation unit’ revisited. Bergen Language and Linguistics Studies,8(1),241-259

Macklovitch E., Russell G. (2000) What’s Been Forgotten in Translation Memory. In: White J.S. (eds) Envisioning Machine Translation in the Information Future. AMTA 2000. Lecture Notes in Computer Science, vol 1934. Springer, Berlin, Heidelberg

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, Robert L. Mercer (1993) The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational linguistics - Association for Computational Linguistics. 19(2),263-311.

Timonera, K. R., & Mitkov, R. (2015). Improving translation memory matching through clause splitting. In Proceedings of the Workshop Natural Language Processing for Translation Memories (pp. 17-23).

常宝宝（2002）. 翻译等价单位自动获取研究[J]，*语言信息处理*，24-29.

刁兴春, 谭明超, & 曹建军. (2010). 一种融合多种编辑距离的字符串相似度计算方法倡. *计算机应用研究*, 27(12).

姜柄圭, 张秦龙, 谌贻荣, & 常宝宝. (2007). 面向机器辅助翻译的汉语语块自动抽取研究. *中文信息学报,* *21*(1), 9-16.

苏明阳, & 丁山. (2009). 翻译单位研究对计算机辅助翻译的启示. *外语研究*, 6, 84-89.

王立非, & 张岩. (2006). 基于语料库的大学生英语议论文中的语块使用模式研究. *外语电化教学*(4), 36-41.

向晓雯, 史晓东, & 曾华琳. (2005). 一个统计与规则相结合的中文命名实体识别系统. *计算机应用*, 25(10), 2404-2406.

袁卓喜. (2009). 试论语块在翻译过程中的作用与启示. *外语界*(5), 85-91.

邱越峰,田增平,季文贇,周傲英.一种高效的检测相似重复记录的方法[J].*计算机学报*,2001(01):69-77.

张卫萍&杨萍.(2002).翻译单位略论.*华北水利水电学院学报,社会科学版*,(2). 93 – 94

郑冰寒，谭慧敏，英译汉过程中翻译单位的实证研究[J]. *外语教学与研究*，2007（2）：145-154.