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。

**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。

**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)。

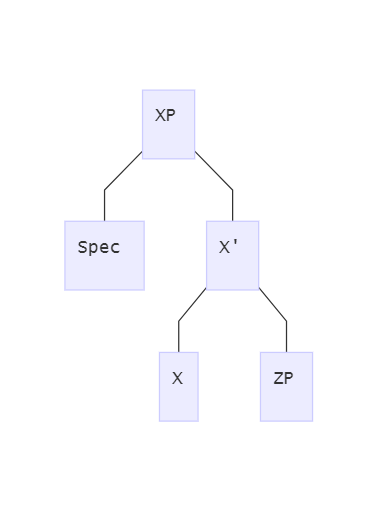
（1）独立子句：独立主句可作为单独的句子存在，当它们与其他子句分离后，通常可以视为句子。例如：I don’t feel like eating this apple because I am full. 在这个句子中，有两个子句，分别是I don’t feel like eating this apple和because I am full. 其中I don’t feel like eating this apple，当拆分开来，都可以视作独立句子。

（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.

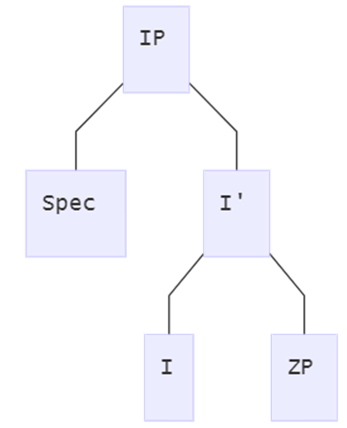
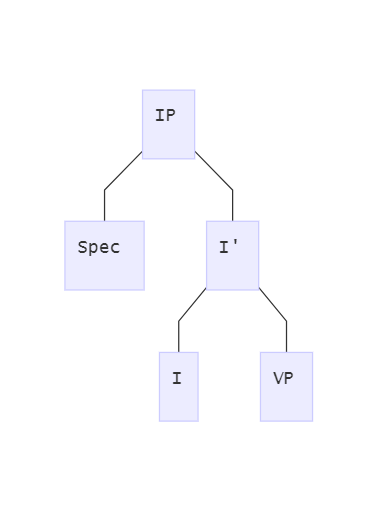
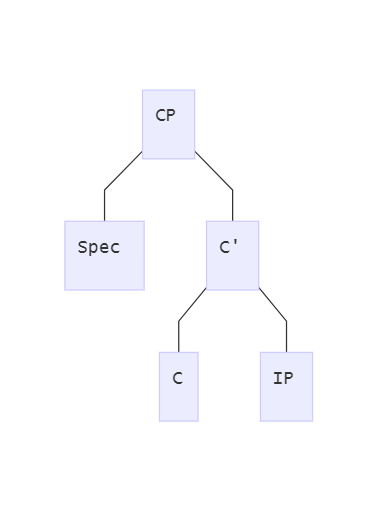
1. 限定子句：也被称必要子句，它的存在对句意表达很重要，对主句起到限定的作用。
2. 非限定子句：也被称为非必要子句，他可以从句子中移除而不会改变句子的基本含义，对主句起到补充的作用。非限定子句通常会以一个逗号与句子的其他部分隔开，或者是前后各用一个逗号，与句子的其他部分隔开。The girl, who used to dance every day, is now in love with a handsome man. 其中who used to dance every day是非限定子句，用逗号与其他部分隔开。且该剧对主要语意没太大影响，只是起到对the girl进行补充说明的作用。

* 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。

**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：

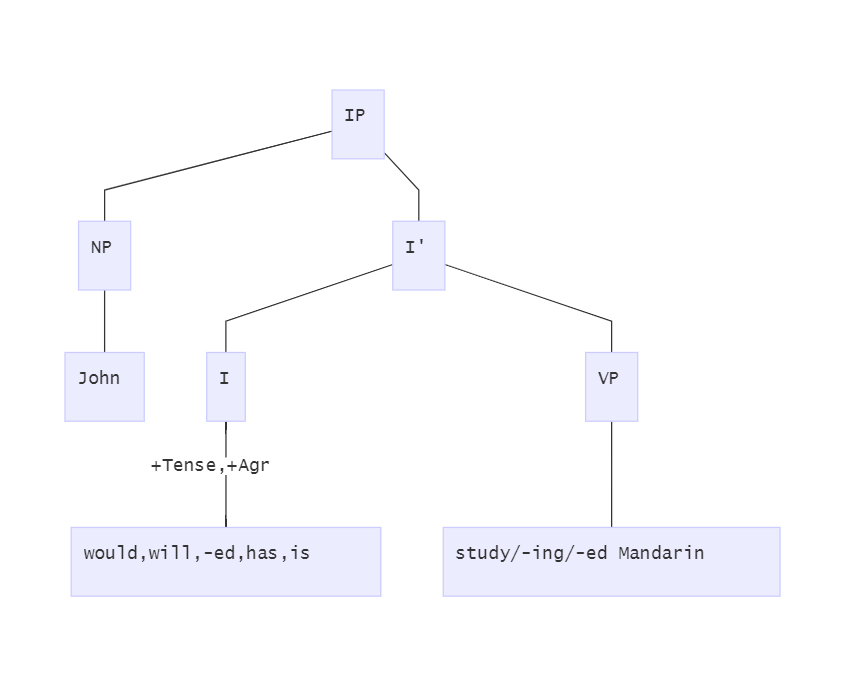
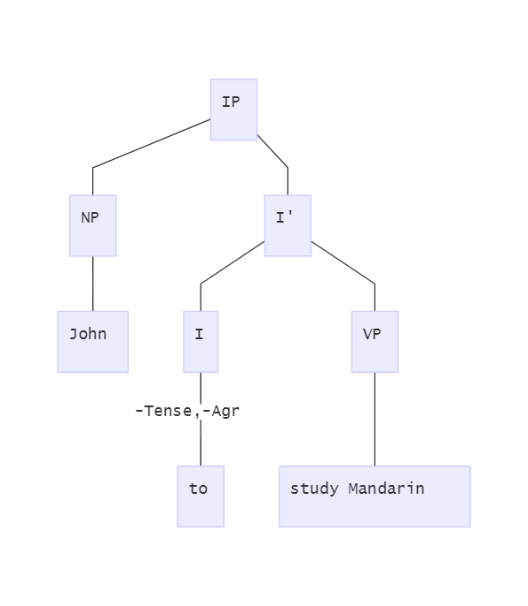


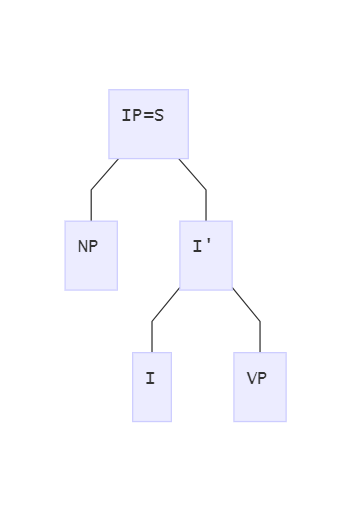
ZP=NP,AP或PP

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。

**1.1 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位置上出现的成分，或者说在I’的姐妹节位置上的成分。

1.2 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。

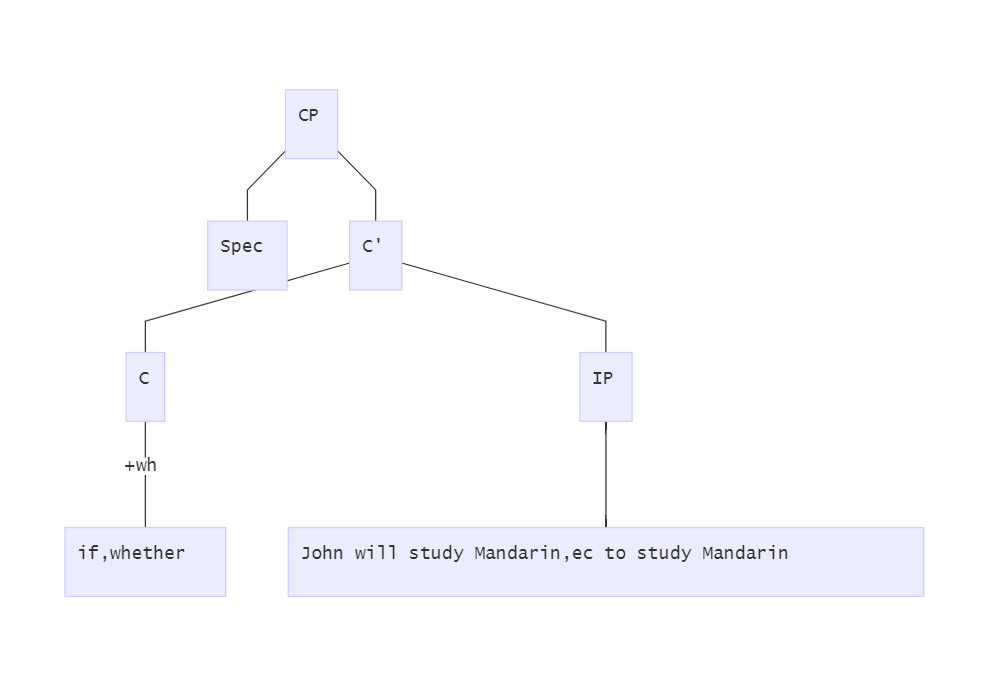
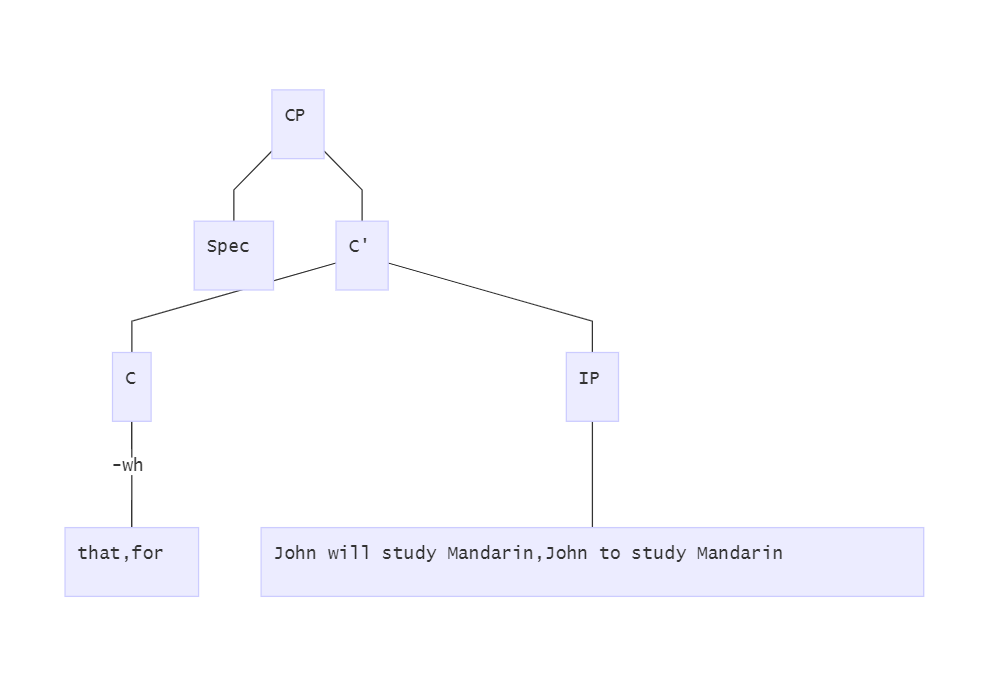
A. They say [that[*IP* John will study Mandarin]]

B. It is difficult [for[*IP* John to study Mandarin]]

C. I wonder [if/whether[*IP* John 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）



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)

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。

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：

1. w1 w2 w3 w4 w5 w6…w20
2. 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。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。

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。

初次以外，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，

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）

Citation

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).

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

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.

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).

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.

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.