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。语块存储信息明确、结构相对稳定、出现频率较高，因此，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。

语块的定义

研究者对于语块的定义不尽相同。

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”。

在自然语言处理方面，常宝宝（2002）提出多词组合单位（multi-word unit）的定义，即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。多词组合单位，在某种程度上，即是指语块。与此相近，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。在该研究中，姜柄圭等人还提出“以多少的语块作为自然语言处理系统的对象至少要应用目标、技术与理论的发展水平以及语言类型”。

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。这一定义更为宽泛，更加具有可操作性，目前大多数基于语料库的语块研究采用这一定义。

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。

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）则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。这种分类结果本质上与Nattinger和Decarrico（1992）所提出的分类具有一定共通性。

语块作为翻译单位

相关研究发现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。

在人工翻译方面，把语块作为翻译单位具有广泛的应用性和操作性。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践过程中的影响，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。

优点

**1. 语块 Vs. 单词**

相较于单词级别的单位，语块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）。

**（1）减少语境缺失导致的翻译歧义**

“打”在《新华字典》中有21个释义，在《新英汉大辞典》中也有超过20个释义。如果以单词级别作为单位，则很难在缺乏上下文的情况下找出准确的翻译。而以语块级别作为单位，例如“打个电话”、“打乒乓球”，则提供了一个较为完整的语境以供参考，从而可以减少歧义现象。

**示例：**

打个电话 make a phone call

打乒乓球 play table tennis

**（2）避免单词对等翻译造成的错译**

在英汉翻译的过程中，有很多表达是以比较固定的形式重复出现的，其对应的翻译也是比较固定的，因此可以将这些表达直接以语块级别进行存储和调用。如果以单词级别作为单位，通过单词对等翻译的方式对这些表达进行重新翻译，反而容易产生错译现象，尤其是在专业性较高的领域，如法律。

**示例：**

提高人民健康水平 to improve people’s health ~~to raise people’s health level~~

criminal lawyer 刑事律师 ~~犯罪的律师~~

substance abuse 滥用毒品 ~~物质滥用~~

**（3）地道翻译习俗语，保留其传达的意义**

习俗语，如谚语、格言等，往往除其字面意义之外，还传达其他的意义。如果以单词级别作为单位，则很难保证其传达的意义能够得到完整的保留，同样也很难保证目标语是否能准确理解。

**示例：**

对牛弹琴 cast pearls before swine ~~play the lute to a cow~~

**2. 语块 Vs. 句子**

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。

**（1）增加翻译记忆的匹配概率**

如果以句子级别作为单位，在进行相似度匹配时，会出现因为无法匹配句子中部分内容而导致整句相似度匹配过低且未能超过设定阈值的情况，从而使得很多可能进行复用的内容被过滤掉，这种现象在长难句中尤为突出。在示例中，若以句子作为记忆单位，则没有翻译记忆条目相似度达到70%以上，从而调用结果为零，然而待翻译语句的部分内容在翻译记忆库中其实是有条目可以参考的。

如果采用语块级别作为单位，则能够尽量多地匹配相似内容，为译者提供更加充分的翻译参考。在示例中若以语块作为记忆单位，则共计至少四个语块达到100%相似度匹配，分别是“multi-stage dynamic game model - 多阶段动态博弈模型”、“production inflexibility - 生产刚性”、“stochastic price - 价格随机”和“expecting profit - 预期利润”，从而帮助译员在获取这些翻译记忆后进行翻译处理。

**示例：**

**【翻译记忆库】**

1）The paper puts forward a **multi-stage dynamic game model** to study the mutual influence between ship owners and home countries.

本文建立了**多阶段动态博弈模型**来分析母国政府和船东之间在船舶移籍问题上的互动关系。

2）The system is making produced circle short and decreasing **production inflexibility** and enhancing the competitive capability of business.

实践表明，该系统有利于缩短产品生产周期，降低产品的**生产刚性**，增强企业的竞争能力

3）This paper puts a model for the renting-or-buying problem under the **stochastic price** of real estate and solves the problem with a series of conclusions.

在房地产**价格随机**性条件下对房产投资者是租房还是买房的投资时机决策问题进行建模研究，给出最优停止决策的最优阀值及相关结论

4）**Expecting profit** is usually employed by engineering corporations to make bidding decision.

工程公司通常把**预期利润**作为投标决策的依据。

**【待翻译语句】**

Using **multi-stage dynamic game model**, the paper studies the strategic behaviors of members on the supply chain with **production inflexibility** and **stochastic price** and gets the probability and supply chain's **expecting profit** of the three strategies that manufacturer purchase stock.

**【译文输出】**

通过运用**多阶段动态博弈模型**，本文对**生产刚性**、**价格随机**条件下的一类供应链上各成员的行为策略进行了研究，得出了制造商购买原料的三种行为的概率和供应链的**预期利润**。

**（2）减轻译员的信息处理负担**

在专业性较强的文本翻译中，以句子级别作为单位可能会对译者造成较重的认知负担，尤其是在提取出的翻译记忆相似度偏低的情况下。在这种情况下，译者需要分别理解待翻译文本、翻译记忆原文以及译文，并且分析待翻译文本与翻译记忆原文的异同，这对于译员而言可能是颇为艰巨的挑战。如果采用语块级别作为单位，由于语块所包含的信息较少，可以有效避免大脑复杂运算，从而减轻译员的信息处理负担。

缺点

使用语块作为翻译复用单元具有诸多优点，较于句子和字词，以语块为翻译单位重复出现的概率更高，复用率也比较高。但基于语块的翻译记忆也存在一定的缺陷。

**1. 语块记忆库的搭建难度较大**

一方面，语块自动抽取等技术尚不完善，语块的边界较为模糊且难以划分和提取，目前的技术还不能准确高效地彻底分析所有的语块。此外，许多语块并不是连续的多词字符串，而是断开的、灵活富有变化的短语，这对语块的自动识别和抽取技术造成进一步的困难。

另一方面，构建双语的语块库难度更大。因为，通常的翻译记忆来源于译员在翻译的同时，CAT软件将译文存储起来，通过句对齐等方法，构建出平行的双语语料库，进而搭建出翻译记忆库。而语块的对齐比较难，因为源语言与目标语言之间存在差别，语块在两个语言之间所处的位置和形态可能有所不同，导致记忆库的搭建更加复杂。相比较句子作为翻译单元，语块的翻译记忆库的搭建更加困难。

**示例：**

短语“take … into consideration”，在识别出该语块后，对应的中文译文应该是“考虑…”，因此，句子的语序发生了变化。

**2. 语块复用问题**

在语块记忆库搭建完成之后，语块作为翻译记忆的提取和呈现比较困难。如果译者在翻译过程中仍以句子为翻译单元，语块作为翻译记忆的复用单元调取出来之后，该语块的译文应该处于句子中的何处，因此可能需要译者手动地调整句子地语序。为了指导译员高效地创作出正确且通顺的译文，并且创造出用户友好型地翻译界面，语块应该如何呈现给译者这一问题需要考虑。

**3. 缺乏上下文导致的歧义问题**

语块是重复出现、稳定共现的字符串，在作为翻译记忆单元时，虽然大多数情况下意义都比较明确，但是仍有可能因为缺乏上下文，在语言色彩和语用环境上导致歧义。

**示例1：**

【语境：在对一位中方经济专家的访谈中】  
英：Professor Wang, do you agree with the prediction, as reported, that China will keep a sound and steady GDP increase this year?   
译：王教授，有报道预测中国今年将会保持健康和稳步的GDP增长，您同意吗？   
中：**是的，当然我同意。**我同意报道中所提及的有关中 国今年将会保持健康和稳步的 GDP 增长的预测。   
译：**Yes, of course I agree.** I agree with the prediction that reports China will keep a sound and steady GDP increase this year.

在汉译英过程中，“当然”一词通常译为“of course”，但在比较严肃的场合，如果将“of course”译为“当然”，显然是非常不妥当的。“of course”的语气色彩在该语境中显得语气十分生硬，带有藐视对方的语用预设，即“这是众所周知的事情，你连这也不知道？” 译为“I couldn’t agree more”更为恰当。

**示例2：**

【语境：译员在中方迎接外方的商务考察中】   
中：感谢各位专程远道来我公司考察，**恐怕**您们一定一路辛苦了。   
译：Thank you for your coming all the way to our company. **I’m afraid** you must have a tiring journey.

在这里，“恐怕”译为“I’m afraid”就将问候的语气变成了担忧的语气，这样的翻译是不对的，可以使用聚合词“might as well”，译为“You might as well have a break after having a tiring journey.”

技术难关

语块的高重复率和强稳定性，反映了语块作为翻译记忆单元的较高可行性与必要性。但是在使用语块作为翻译记忆单元时，仍然面临的技术难点包含以下几点：

一、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提取所有的语块。此外，抽取前的准备工作，包括分词与词性标注等，是目前自然语言处理中的难点；而抽取后的对齐工作，无疑也是目前需要攻克的技术难关。

二、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。

三、语块复用问题。语块作为翻译记忆复用单元，但译者在翻译过程中大都以句子为翻译单位（以trados为例，源文本被划分为句子，译者逐句翻译），而语块作为翻译记忆单元在检索复用后，如何呈现匹配结果，如何营造用户友好型界面，以及如何高效地指导译员创作出正确且通顺的译文，都是语块复用的难点所在。

-------------------------------------------------------------------------------

参考文献：

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.

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.

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.

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.

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.

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.

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

王立非，张岩，基于语料库的大学生英语议论文中的语块使用模式研究[J]. 外语电化教学，2006（4）：6-11.