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。

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。

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。

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。

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。

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。

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。

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

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.