一种面向汉英口语翻译的双语语块处理方法

程葳　赵军　徐波　刘非凡

（中国科学院自动化研究所模式识别国家重点实验室　北京　100080）

摘要：基于语块的处理方法是近年来自然语言处理领域兴起的一条新思路。但是，要将其应用于口语翻译当中，还需按照口语特点对涉及双语的语块概念做出合理界定。本文在已有单语语块定义的基础上，根据中、英文差异和口语翻译特性，从句法和语义两个层次提出了一种汉英双语语块概念，并对其特点进行了分析。同时，针对中、英文并行语料库，建立了一套计算机自动划分与人工校对相结合的双语语块加工方法。应用该方法，对汉英句子级对齐的口语语料进行双语语块划分和对整，并以此为基础进行了基于双语语块的口语统计机器翻译实验。结果表明，本文提出的双语语块定义符合口语翻译的实际需要，使用基于双语语块的语料处理方法，能有效地提高口语系统的翻译性能。

关键词：统计机器翻译；口语翻译；语料库；语块

中图分类号：TP391

Bilingual Chunking for Chinese-English Spoken-language Translation

CHENG Wei, ZHAO Jun, XU Bo and LIU Fei-Fan

(National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences　Beijing　100080)

Abstract：Chunking is a useful step for natural language processing. The paper puts forward a definition of bilingual chunks for Chinese-English spoken-language translation, based on both the characteristics of spoken-language and the differences between Chinese and English. Some special features of these chunks are also analyzed. Based on the definition and analysis, a method is proposed to segment the chunks in bilingual corpora. This method includes two steps of the automatic chunking and the manually modification. Using this method we got a chunk-aligned Chinese-English bilingual corpus. A series of chunk-based statistical machine translation experiments are then conducted which shows that the proposed definition and the bi-chunking method can lead to great improvement to the quality of the Chinese-English spoken-language translation.

Keywords：statistical machine translation; spoken-language translation; corpora; chunk

---

本文受国家“973”项目 G1998030501A-06、国家自然科学基金重点项目 69835003 和国家自然科学基金项目 60272041 资助；作者：程葳，博士研究生，主要研究方向为口语翻译；赵军，博士，主要研究方向为自然语言处理、信息检索和机器翻译；徐波，研究员，博士生导师，主要研究方向为语音识别和语音翻译；刘非凡，博士研究生，主要研究方向为中文信息处理。

一、 引言

口语自动翻译（以下简称为口语翻译）是机器翻译的一项研究热点。基于语料库的统计方法（Statistical Machine Translation，以下简称为 SMT），因对不规范语言现象具有较强的鲁棒性，而被广泛应用于口语翻译当中[1]。但是，目前的 SMT 方法主要以“单词”作为基本处理单元，其粒度较小，限制了系统翻译精度的提高[2]。因此，探索合理的、适合口语特点的双语结构单元，对口语翻译的进一步发展具有重要意义。

近年来，“基于语块（chunk）的处理方法”为人们提供了一条新思路。它根据分治（Divide-Conquer）原则，把原先粒度较精细的处理单元——单词——扩大为具有结构上稳定性和功能上无歧义性的语块，从而达到加大信息处理粒度、简化源语句型和化解机器翻译歧义的目的。目前，有关单语语块的研究工作已取得大量成果。例如，英文方面，[1991，Abney]提出了一个完整的语块描述体系。[2000，Erik 和 Sabine]介绍了自然语言学习国际会议（CONLL-2000）提出的语块共享任务（Chunking Shared Task），该任务采用 Abney 的语块描述框架和 Penn 树库的华尔街日报（WSJ）部分，旨在开发出一个大规模的英语语块库，为基于统计的各种不同的部分句法分析方法提供统一的训练和测试库。中文方面，[1999，周强，孙茂松，黄昌宁]提出了词界块和成分组等概念，以建立一种汉语句子的组块分析体系。[2000，刘芳，赵铁军等]将“包含一层或两层的符合一定句法功能和反映组成意义的短语”定义为“汉语组块”，并应用统计与错误驱动相结合的方法实现语块的自动分析；[2001，周强，詹卫东等]介绍了一种“侧重于自顶向下地描述句子基本骨架”的语块定义，并以此构建了一个大规模汉语语块库。[2001，鲁川]将汉语语块分为“中枢语块”和“周边语块”两种类型，并在此基础上对汉语句子的语序进行了研究。

不过，要想将基于语块的处理方法应用到口语翻译当中，还需解决以下两方面问题。

1) 对口语灵活的表达方式、以及它包含的诸如省略等不规范的语言现象予以考虑；
2) 建立双语相关的语块描述框架。

有关后者，[2002，李沐，吕学强，姚天顺]提出过一种扩展语块（Extended Chunk, E-Chunk）的概念。它是针对基于实例的翻译方法，完全根据语义定义，将具有语义自足性和转换充分性的无歧义、可嵌套翻译单元定义为双语 E-Chunk。并提出了一种基于 E-Chunk 的机器翻译模型（E-Chunk based machine translation, ECBMT）。

本文从另一角度，以汉英口语为研究对象，在我们已建立的基于单词的 SMT 系统[14]基础上，根据中、英文双语间差异和口语翻译的实际需要，提出了一种汉英对应的双语语块描述框架，并以此建立了一套针对口语并行语料库的双语语块加工方法。以下是文章的基本内容：第二部分对面向口语翻译的汉英双语语块定义和特征进行了描述。第三部分则介绍了基于双语语块的汉英并行语料加工方法。第四部分将该方法应用于统计口语翻译，并对处理前后的语料进行对比试验，结果表明，采用双语语块处理过的语料库训练 SMT 模型参数，可以使系统的翻译精度提高约 20%。最后，对方法中尚待研究的一些问题予以讨论。

二、 面向口语翻译的汉英双语语块基本概念

2.1 双语语块的定义

按照 Abney 的说法[3]，单语语块是具有单一语义核心和严格非递归句法结构的单词或连续词串，句子中的每一个单词都将划到某个语块当中。但是，对于翻译系统来说，建立源语言和目标语言间的对整关系是需要遵循的一条重要原则。在此基础上，保障具有与单语语块相类似的句法结构，可使已有研究成果得到继承。因此，本文将满足下述句法、语义特征的成分单元定义为双语语块。

- 具有不相交、非嵌套的句法结构。其句法解析树（parse-tree）是句子解析树的某一连通子图，树根（root）代表了语块的结构类型。应用上下文无关的短语结构语法（CF-PSG）即可对其进行描述。
- 具有无歧义的语义核心和翻译上的转换充分性。语义核心（s-head）由语块内部的一个或多个实词组成，它们之间满足语义关联的局部性假设[10]，因此可以消解词

汇歧义；而翻译上的转换充分性是指，每个源语语块（如中文语块）都可对应地转换成为一个目标语语块（如英文语块）或空语块（这里，我们将空语块定义为一种特殊的双语语块），反之亦然。

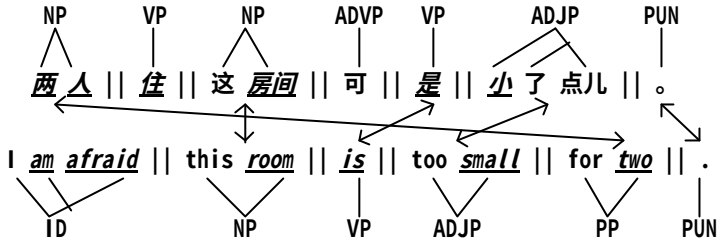

图 1 双语语块定义举例

（图中：双线隔开的词串构成语块；NP、VP、ADJP 等是双语语块标记，具体含义请参见表 1 中的注释；带下划线的单词为所在语块的语义核心。）

图 1 给出了双语语块的句法分析和翻译转换实例。从中可以看出，双语语块与单语语块相类似，亦由单词（orphan nodes）、短语和子句（clause）三个层次[11]组成，而且中、英文双语句对中的每一个单词都将划到某个语块当中。

2.2 双语语块的特点

由上述定义可以看出，双语语块概念具有以下特点：

1. ***双语相关的语义框架。***定义中“语义核心的无歧义性”和“翻译上的转换充分性”一方面建立了中、英文语块间的语义联系，另一方面也使双语语块具有较好的单元对整性，从而有利于提高翻译的精度。

2. ***严整的句法结构。***定义中基本保留了 Abney 对单语语块的句法结构要求，这就使得现有的单语语块分析方法（如浅层分析方法）可直接应用于双语语块的辅助划分，同时也有利于双语语块的进一步归类、分析。

表 1 英文语块和双语语块英文部分在句法方面的比较

功能分类	句法规范上的主要差别	
	CoNLL-2000 语块共享任务[4]	汉英双语语块中的英文部分
NP(名词语块)	1、 所有格 NP 在所有格符号（’和’s）前分开（例：[NP Eastern Air-lines] [NP ’ creditors]）。 2、 名词前修饰名词的形容词或 ADJP 划入 NP；名词后的修饰成分则不包括在内。	1、 语块中可包括名词所有格和物主代词修饰成分（例：[NP John’s hat]）。 2、 名词前后修饰名词的 ADJP 均可划入 NP，也均可与名词划开，这主要取决于对应的中文翻译。
VP(动词语块)	1、 VP 中不包含代词,代词作为单词语块独立处理。 2、 动词前修饰动词的副词或 ADVP 划入 VP；动词后的修饰成分则不包括在内。	1、 代词将划入与其最相关的相邻语块。例如：动词后紧跟的代词宾语一般划入 VP 中；系动词一般与代词主语划在一起等。 2、 动词前后修饰动词的 ADVP 均可划入 VP，也均可与动词划开，这主要取决于对应的中文翻译。
ADJP、ADVP(形容词、副词语块)	基本一致。	
PP(介词语块)	一般仅包含介词本身。	介词将与其最相关的相邻语块划在一起，构成 PP。
SBAR(小句语块)	基本一致。	
ID(习惯用语和固定用法语块)	无	包括口语中的习惯用语和插入语，以及中文固定用法语块所对应的英文翻译。
PUN(标点语块)	无	由标点和位置上与标点相邻的语气词组成。
其它语块	基本一致	

3. ***兼顾双语口语特性的句法规范。***尽管双语语块的句法结构要求与单语语块基本一致，但两者在具体的句法规范上却存在着很大差异。这主要是因为双语语块的句法规范需要

同时兼顾中、英文双语的口语特性，以便保障该句法结构可以适应双语相关的语义框架。以英文为例，表 1 列举了“ CoNLL-2000 语块共享任务中的英文语块 ”与本文“ 汉英双语语块中的英文部分 ”在各不同功能类型中的主要差异，而造成这些规则变动的原因大致包括：

- 中、英文语言差异：中、英文是两种差别较大的语言。例如在介词应用方面，中文中的介词数目相对较少，用法不及英文灵活，而且经常出现类似“ 在…上面 ”的形式，因此，尽管英文单语中的介词语块只包含一个介词，但对于双语语块，需要将介词划入与其最相关的相邻语块当中（见表 1 介词语块部分）。

例句 1 ：“ 看 录像带 的 ‖ 费用 ” <——>“ the charge ‖ *for* the VCR ”；

例句 2：“ 在 酒吧 || 喝 的 || ， || 由 || 323 号 房 的 || 田中 || 付 了 ”

<——>“ Mr. Tanaka ‖ *in* room 323 ‖ paid ‖ *for* the drinks ‖ *at* the bar ‖ . ”

- 口语中的不规范语言现象：中文口语中包含大量不规范语言现象，例如：省略、重复、次序颠倒和冗余等。其中，省略现象的出现几率约为 33%[12]，但在相应英文翻译中，被省略的部分往往由代词指代，这种情况在主语和宾语的位置尤为常见。因此，尽管英文单语语块将代词作为孤立的单词语块处理，但在双语语块中，代词将与其最相关的相邻语块划在一起（见表 1 动词语块部分）。

例句 3 ：“ 是 ‖ 玛丽蓝 大 饭店 ‖ 。” <——>“ *This* is ‖ the Marina Hotel ‖ . ”

例句 4：“ 请 你 ‖ 再 查查 ‖ 好吗 ？” <——>“ Would you ‖ check *it* again ‖ ? ”

- 口语中大量习惯用语和固定结构的使用：考虑到口语中存在着大量习惯说法，双语语块中将习惯用语和插入语作为特殊的语块类型进行处理。另外，中英文间的差异也使得一些中文中的固定结构与其英文翻译间存在着意译关系，例如中文的“ 把 ”字结构：“ 把/将 NP VP NP ”在英文中既可翻译成“ 动词带双宾语 ”，又可翻译成“ have sth. done PP ”。因此，双语语块中将这样的结构定义为一个语块，其相应的英文翻译也划为同一语块（见表 1 习惯用语和固定用法语块部分）。

例句 5：“ ‖ 把 行李 搬到 房间 ‖ ”<——>“ ‖ have my baggage delivered to my room ‖ ”；

例句 6：“ ‖ 将 把 您 送到 房间 去 ‖ ” <——>“ ‖ will show you your room ‖ ”。

三、 基于双语语块的并行语料加工方法

3.1 双语语块的划分原则

从功能的角度来看，中文语块包括名词语块、动词语块、疑问词语块、形容词语块、介词语块、副词语块、语气词和标点语块、插入语和习惯用法语块等八种类型。英文语块包括 NP（名词语块）、VP（动词语块）、ADJP（形容词语块）、PP（介词语块）、ADVP（副词语块）、SBAR（小句语块）、INTJ（感叹词语块）、ID（习惯用语和固定用法语块）、PUN（标点语块）等九种类型。在对中、英文参照语句进行双语语块划分时，我们主要考虑以下的基本原则。

1) ***双语对整原则。*** 即在保证语块句法结构的基础上，尽量满足双语语块的对整要求。这其中包括：尽量减少对空的单元；按照一一对应原则划分语块，尽量避免出现单元的一对多、多对一或多对多情况。

2) ***扩大语块优先原则。*** 即在需要对单元进行合并或拆分时，应优先考虑通过合并的方式扩大语块，实现双语语块的转换充分性。

3) ***以中文为基准原则。*** 由于本文的研究对象是汉英翻译，因此双语语块的划分将遵循以中文为基准原则。包括两种情况：一是在需要拆分或合并的情况下，尽量保证中文语块的结构完整性，而首先考虑拆分或合并英文语块；另一是尽量依据中文语块对英文语块进行划分。

3.2 语料加工流程

图 2 中 I 部分给出了对中、英文并行生语料库的加工流程。在对生语料进行中文分词和中英文词性标注等预处理工作之后，基于双语语块的语料加工过程主要包括以下两个步骤：

- *双语语块的自动预划分。*该步骤借鉴了单语语块处理中的浅层分析方法[13]，应用自底向上的概率上下文无关文法分析器分别对中、英文句子进行自动语块划分。由于分析器所使用的规则是根据中、英文双语语块的句法规范编写的，因此，预划分过程可以保障语块满足句法结构要求。同时，自动的划分过程也提高了语料加工的效率，减轻了下一步人工校对的工作量。不过，自动预划分的语块不一定满足翻译上的转换充分性，因此我们称其为“准双语语块”，而由自动预划分过程处理过的并行语料，被称为“基于准双语语块的语料”。

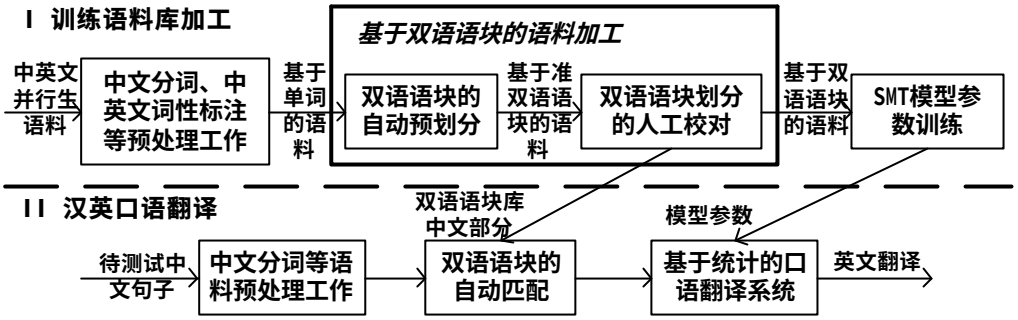

图 2 双语语块在汉英口语翻译中的应用

- *双语语块划分的人工校对。*该步骤目前主要由人工完成。即，根据双语语块的划分原则和句法规范，参照中英文并行语句对，对“基于准双语语块的语料”进行校对，使之完全满足双语语块定义。同时，对于部分由于翻译不好而无法进行双语语块划分的垃圾语料给予改造或剔除。经过人工校对后的语料，就是我们所需要的“基于双语语块的语料”。

四、 实验及结果分析

文献[14]介绍了我们构建的一个基于单词的统计口语翻译系统。其基本原理是将翻译过程视为噪声通道模型，根据 Bayes 公式，对给定的源语言句子 S，求取其相应目标语句子 T，

$$T = \arg\max\{P(T \mid S)\} = \arg\max\{P(T) \times P(S \mid T)\} \qquad (1)。$$

在旅馆预定领域的汉英语音翻译应用中，我们发现，尽管系统对口语不规范语言现象和语音识别错误具有一定鲁棒性，但其翻译精度尚不能令人满意。因此，我们对原有的“基于单词的语料”进行了双语语块加工，分别生成了“基于准双语语块的语料”和“基于双语语块的语料”，然后应用这些语料重新训练模型参数，并在相同的测试集上，对系统性能进行评测，以检验基于双语语块的语料处理方法对口语翻译的有效性。

4.1 实验过程

实验由三个子部分组成。其所用训练语料是旅馆预定领域对话语料，包括 2665 对中英文并行口语语句，共 23578 个中文单词和 27392 个英文单词。

Exp1：应用“基于单词的语料”训练系统模型参数。然后，在对“待测试中文语句”进行分词等预处理工作后，由基于统计的口语翻译系统进行翻译。

Exp2：应用“基于准双语语块的语料”训练系统模型参数。然后，在对“待测试中文语句”进行分词等预处理工作后，应用自底向上的概率上下文无关文法分析器对语句进行“准双语语块”的自动划分。其算法和所用中文规则集同语料加工中“双语语块的自动预划分”过程基本一致。最后，由基于统计的口语翻译系统对划分后的语句进行翻译。

Exp3：应用“基于双语语块的语料”训练系统模型参数。然后，在对“待测试中文语句”进行分词等预处理工作后，需加入双语语块的自动划分过程（如图 2 中 II 部分所示）。由于基于双语语块的语料加工方法中包含有人工因素，所以测试集的处理方法将与训练集的有所不同。这里，我们采用的是自动匹配方法。即，将测试语句中的单词串，同双语语块划分时提取的中文语块进行最长匹配，以实现语块的自动划分。该方法虽然简单，但由于语块

的平均词长较长（一般为 3~5 个单词，最长可达到 8 个单词），因此划分的正确率可以达到 98%以上。最后，应用基于统计的口语翻译系统，对由双语语块组成的中文句子进行翻译。

实验的测试集由 1000 句中文口语语料组成，平均句长约为 10 个单词。系统评价将采用自动和手工两种方法：

1) 自动评测。根据口语翻译的实际需要，我们研究了一种自动评价算法，其打分（score-F）的高低可以近似反映出系统输出译文质量的优劣。具体公式为：

$$score-F=\sum_{i=1}^{N}\frac{(b^2+1)\times precision_i\times recall_i}{b^2\times precision_i+recall_i}/N$$ （在本文实验中β＝1） （2）。

其中，N 为测试集大小（在本文实验中 N=1000）；$precision_i$ 为第 i 个测试语句的广义精确度；$recall_i$ 为第 i 个测试语句的广义召回率。各参数的计算方法请参见文献[15]。

2) 人工评测。由于人工评测的工作量较大，这里，我们只对测试集中的 50 句，从“可懂度”和“忠实度”两方面进行评价。具体为：

- 可懂度。即英文翻译可被人理解的程度。共分四个等级：

A：完全正确； B：存在部分语法错误，但不影响全句理解；
C：错误过多，只能读懂部分片断； D：完全错误。

- 忠实度。即翻译对原有中文语义的忠实程度。共分四个等级：

A：完全正确； B：反映了原文中的主要或关键信息；
C：只保留了原文中的部分信息； D：与原文意思完全不同。

4.2 结果分析

表 2 给出了三组实验的结果对比。从中可以看出：

表 2 三个口语翻译系统的测试结果对比

训练语料	自动评测 score-F(%)	人工评测							
		可懂度(%)				忠实度(%)			
		A	B	C	D	A	B	C	D
基于单词的语料	58.94	43.75	25.00	20.83	10.42	29.17	22.92	33.33	14.58
基于准双语语块的语料	71.50	50.00	22.92	18.75	8.33	45.83	22.92	25.00	6.25
基于双语语块的语料	79.41	60.42	20.83	12.50	6.25	66.67	22.92	10.40	0.01

- 与原有“基于单词的语料”相比，“基于双语语块的语料”使系统的自动评测打分提高了约 20%，这说明本文所定义的“双语语块”适合汉英口语翻译的特点，可以有效地改善双语单元间的对整性，因而提高了口语翻译的质量。
- 准双语语块虽然只具有双语语块的句法特性，但“基于准双语语块的语料”也使系统的自动评测打分比原有“基于单词的语料”高出了约 12%，这说明语块的句法结构合理扩大了统计翻译模型的分析单元，从而使系统的翻译性能得到改善。
- “基于双语语块的语料”使系统的自动评测打分比“基于准双语语块的语料”提高了近 8%。这说明，保障双语语块在翻译上的可转换性具有重要的意义。不过，从另一方面来讲，准双语语块的自动加工过程大大降低了双语语块划分的难度和工作量，并为并行语料库中部分无法进行双语语块处理的口语特例，提供了次优的选择，从而使扩大语料处理规模成为可能。
- 本实验还从另一个角度向人们说明，口语句子的“造句单元”是“语块”，而不是“单词”，因此，基于统计的口语翻译方法应以“语块”作为其基本处理单元。

五、 结束语

本文所提出的双语语块处理方法，在保障分析单元具有一定句法结构的基础上，有效地改善了双语间的对整关系，从而使口语翻译系统的性能得到了提高。不过，面向口语的双语

语块研究只是刚刚起步。方法中还存在着一些问题，需要进一步的研究和探索。

- *扩大语料处理规模*。从统计的角度来讲，分析单元的扩大将可能加重模型的数据稀疏问题。也就是说，利用本文提出的方法时，也许需要更多的双语语块对齐语料。因此，我们将在已加工的近三千句语料基础上，尝试应用该方法处理一个六万多句的大规模并行口语语料库，并借此完善双语语块的句法规范和划分原则。
- *研究自动划分方法*。目前，双语语块加工还需要通过人工校对来保证翻译上的可转换性。但当语料处理规模扩大时，该方法将会耗费大量的人力物力。因此，我们希望能在今后的研究工作中更深入地发掘双语间对整规律，从而建立完全自动的语料加工方法。
- *拓展双语语块的应用*。除了本文所介绍的对 SMT 系统训练语料进行加工以外，双语语块还可用于提取翻译模板和建立双语语块库。同单语语块库一样，大规模的双语语块库将成为机器翻译中的一种重要资源。因此，双语语块处理方法具有十分广阔的应用前景。

参考文献

[1] 宗成庆，黄泰翼，徐波．口语自动翻译系统技术评析．中文信息学报，1999, 13(2): 56-64

[2] M.Carl. A model of competence for corpus-based machine translation, In: Proceedings of COLING'2000. Saarbrücken, Germany, 2000

[3] Steven Abney. Parsing by Chunks. In: Robert Berwick, Steven Abney and Carol Tenny (eds.). Principle-Based Parsing. Kluwer Academic Publishers. 1991

[4] Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to CoNLL-2000 Shared Task: Chunking In: Proceedings of CoNLL-2000. Lisbon, Portugal, 2000, 127-132

[5] 周强，孙茂松，黄昌宁．汉语句子的组块分析体系．计算机学报，1999, 22(11): 1158-1165

[6] 刘芳，赵铁军，于浩等．基于统计的汉语组块分析．中文信息学报，2000, 14(6): 28-32

[7] 周强，詹卫东，任海波．构建大规模的汉语语块库．见：全国第六届计算语言学联合学术会议(JSCL-2001)．山西太原，2001

[8] 鲁川．信息处理用汉语句子语序的认知研究．见：中国中文信息学会二十周年学术会议．北京：清华大学出版社，2001, 186-197

[9] 李沐，吕学强，姚天顺．一种基于 E-Chunk 的机器翻译模型．软件学报，2002, 13(04): 669-675

[10] Yael K. and Edelman S. Learning similarity-based word sense disambiguation. Computational Linguistics, 1998, 24(1): 41-60

[11] Steven Abney. Chunks and Dependencies: Bring processing evidence to bear stntax. In: Computational Linguistics and the Foundation of Linguistic Theory. CSLI. 1995

[12] 宗成庆，吴华，黄泰翼等．限定领域汉语口语对话语料分析．见：计算语言学文集（全国第五届计算语言学联合学术会议论文集），1999, 115-122

[13] 孙宏林，俞士汶．浅层句法分析方法概述．当代语言学，2000，2(2):74-83

[14] Cheng Wei and Xu Bo. Statistical Approach to Chinese-English Spoken-language Translation in Hotel Reservation Domain. In: ISCSLP'00. Beijing, 2000, 271-274

[15] 程葳，徐波．一种面向口语的译文质量自动评价方法．中文信息学报，2002, 16(2): 47-53