“把”字句的自动释义与句式变换研究[1]

王璐璐[1,2]，孙薇薇[1]，袁毓林[1]

（1.北京大学，北京市 100871；2. 中国传媒大学，北京市 100024）

摘要：本文针对“把”字句在机器翻译中的困难，探索一种规则和统计相结合的“把”字句的自动释义和句式变换的方法。具体的计算步骤为：（1）根据“把”字句与其他句式的变换关系，我们将“把”字句分为不同的小类，并总结出每一小类的句法语义特征，得到“把”字句的语言模型；（2）我们选取北大中文树库中的“把”字句作为语料，并标注上每一小类句式的句法语义特征，从而得到富含句法语义信息的标注文本；（3）在此基础上，我们分别用组块分析的方法和完全句法分析的方法来对“把”字句进行自动识别；（4）再利用判别式机器学习的方法来对“把”字句进行自动分类。在识别结果和分类结果的基础上，我们根据释义模板和变换模板得到了一个“把”字句的自动释义与句式变换程序。

关键词：“把”字句；变换分析；框架识别；自动分类；自动释义

On the Automatic Interpretation and Pattern Alternation of Chinese *ba* Constructions

Lulu Wang[1,2], Weiwei Sun[1], Yulin Yuan[1]

(1.Peking University, Beijing 100871, China; 2. Communication University of China, Beijing, 100024 ,China)

Abstract： In order to enhance the accuracy of the automatic interpretation of the *ba* construction, this paper adopts a computational oriented and cognitive based schema, and attempts to have an automatic analysis of the semantic interpretations and syntactic alternations of the *ba* constructions. We firstly classify the *ba* construction into different subtypes based on the alternations of the *ba* constructions with other constructions and summarize the syntactic and semantic features of each subtype. Then, we build up a language model of the *ba* constructions, with an annotated text filled with the syntactic and semantic features of the *ba* constructions. Further, we automatically identify the *ba* constructions based on the chunking and parsing methods, and automatically classify them based on machine learning. Finally, we design an automatic interpretation and alternation system of the *ba* constructions.

Key Words: the *ba* construction; Pattern Alternations; Frame Identification; Automatic Classification; Automatic Interpretation;

1. 引言

“把”字句是现代汉语的一种常用句式。在语言学本体研究领域，有关“把”字句的句法结构和语义特点的研究数量众多且成果显著。但在计算应用方面，单就“把”字句的分析并不多见。而且目前主流的机器翻译系统对“把”字句的翻译并不十分理想。我们曾对 google 在线翻译系统进行中英互译测试[2]，测试发现：对于不带宾语的“把”字句，翻译结果较好，44%的句子在中译英和英译中后翻译为“把”字句；但是对带宾语的“把”字句，翻译结果较差，

[1] 本课题的研究得到国家社科基金重大招标项目《汉语国际教育背景下的汉语意合特征研究与大型知识库和语料库建设》（批准号：12&ZD175）的资助，谨此致以诚挚的谢意。

[2] 在线测试时间：2011 年 12 月 28 日。

58%的句子的中译英可以理解，42%的句子翻译完全错误。英译中则没有一句翻译出“把”字。例如，“他们把粮食装上汽车”译为“They loaded grain cars”（“他们装粮车”）。我们认为，google 在线翻译系统在对“把”字句的翻译中有两点不足：一是“把”字句基本句式意义的缺失。如处置义及影响义都没有在译文中得到反映。二是带宾语的“把”字句中，由于动词后宾语的出现，提高了句式的复杂度，更提高了计算的困难度。

针对“把”字句在机器翻译中的困难，我们认为有必要对“把”字句的计算分析进行改进。现有的应用系统，无法将“把”字句这种具有复杂的句法语义信息的句式进行精细化的自动分析。对于“把”字句的研究，语言学界已经发现了不少语言事实和相关规律，也有一些成熟的理论，在论元结构和句式语义方面的研究成果尤为突出。如果能把这些理论借鉴到形式化的语法研究之中，应该有利于提高计算分析的准确度。

由此，本文采取一种基于认知假设并面向计算分析的技术路线（袁毓林 2008），尝试将语言学的学理性研究与计算方面的实证性分析结合起来，探索一种规则和统计相结合的“把”字句的自动释义和句式变换的方法，为机器翻译等应用系统提供可供复述（Paraphrases）的资料。

2. 基于变换的“把”字句的计算建模

2.1 基于变换的“把”字句的语义类

我们将机器理解“把”字句的过程处理为一个分类问题，即将无限的语言实例（token）对应到具体的语言类型（type）上面。由此，我们需要多级标注的语料来训练机器，让它自动学习“把”字句的句法语义信息。那么，机器要学习哪些句法语义信息，需要借鉴语言学领域的研究成果。

在汉语学界，关于“把”字句的语法意义是语言本体研究中的一个难点。主流的观点有处置说（王力 1943），致使说（郭锐 2003、叶向阳 2004），影响和结果说（邵敬敏 1986、薛凤生 1987、崔希亮 1995、张伯江 2000、张旺熹 2001 等）。这些不同的观点正说明了“把”字句语义构成的复杂性。对于计算机而言，区分“把”字句内部的语义差异是十分必要的。王璐璐（2013）提出，可以通过不同句式之间的变换关系分析为手段，来揭示“把”字句内部在结构形式和语义表达方面的差异，并为后面自动获取“把”字句的语义解释做准备。这种分析技术主要参考了 Levin（1993）提出的动词词汇语义类与句式变换之间有内在关系的理论假设，即句式之间不同的变换关系反映了其中动词的不同意义差别。据此，我们推广到句式层面，假定不同结构形式的“把”字句有着不同的语义解释，也有着不同的变换式系列。詹卫东 （2004）指出，对于计算机而言，“理解‘意思’的过程，可以表示为对符号进行‘变换’的过程”。所以说，通过“把”字句与其他句式的变换关系，我们可以将复杂的语法意义具体化为每一小类 “把”字句的句法语义信息，并力求每一小类“把”字句的句法结构与语义关系的相对单一性，从而达到对“把”字句精细化理解的目的。王璐璐（2013）根据对真是文本中“把”字句的考察，总结出了 26 类细分类和 8 类粗分类的“把”字句的语义类。考虑到信息的粒度，太粗或者太细的分类标准都不太合适。过粗分类对于后面的释义来说意义不大，因为无法区分出各小类“把”字句的意义差异。过细的分类对实验的结果具有很大的干扰性，不容易有效地分出各个小类。有鉴于此，我们的实验采用了八小类的类别标准，如下所示：

表 1 粗分类的“把”字句语义类

序号	语义类	释义模板	类例数量[3]
1	处置-位移	{X+V.+Y} 使得 {Y+DV/到/在(+SP)+DV/去}	1137
2	处置-结果	{X+V.+Y} 使得 {Y/{V.+Y}+AP/VP}	483
3	处置-结果（隐含）	{X+V.+Y} 使得 {Y 发生了某种变化}	382
4	处置-转移	{X+V.+Y} 使得 {Y+属于+NP3}	222
5	处置-关系	{X 认定}， {Y+相当于+NP3}	150
6	处置-变成	{X+V.+Y} 使得 {Y+变成+NP3}	124
7	方式-结果	{X+将+Y+VP}	11
8	主观-结果	{X} 经历（造成并遭受）了{Y+ V.-了}这件事	2

2.2 语料的来源与标注

本文的数据选取了北大中文树库中的 2441 句“把”字句，并在此基础上进行句法语义信息的深加工工作。

在现有的树库资源中，对“把”字句的句法结构有两种不同认识：一种是以宾大中文树库为代表的“IP”说，他们将“把”看作是动词，其后的成分是一个小句。另一种是以北大汉语树库为代表的“pp”说，他们将“把”看作是介词，它与“把”后名词构成一个介宾短语。这种区别实际上可对应于这两大类树库资源背后不同的语法观，即对汉语语法结构的不同认识。前者的“动词说”在国外的“把”字句研究中占据主导地位，如 Zou（1995）、Bender（2000），Gao（2001），Lipenkova（2010，2011）等都支持这一观点。这种观点的好处在于，它可以很好地解释带宾语的“把”字句，并将部分“把”字句与无标记的被动句联系起来。在汉语学界，部分支持这个观点是朱德熙先生提出的受事主语说，即删除“把”字，后面的部分是受事作主语的句子。但是，在实际语料中，这个观点仍不能解释所有的“把”字句。例如，“作为”类动词所在的“把”字句是不能删除“把”字的。如“*吸引外资作为缓解就业压力的一个有效办法”是不成句的，它要么作“是”字句的主语，要么需要加上“把”字。

相比较而言，北京大学汉语树库在描述汉语时采用的是朱德熙先生的功能分类的思想，而且语素、词和短语之间构建了较好的功能对应关系（周强等，1997：44）。在对“把”字句分析中，北大树库沿袭了王力先生提出的“提宾说”，将“把”后的名词看作是从主动宾句中动词后的宾语位置提前所致。虽然并不是所有的“把”字句都能用提宾说来解释，如“把老伴儿死了”，但是，这种观点是最为符合母语者的语感，这与“把”字句所具有的处置义是密不可分的。根据我们的统计，北大树库对“把”字句标注的准确率高达 99.6%。由此，我们倾向于在北大中文树库的基础上深加工“把”字句的句法语义信息。

由于本文是面向于大规模的语言工程实现，标注资源的构造也应考虑到大规模数据的特点。这就要求我们标注那些对计算分析最为重要的信息，而不能如语言学分析那样做到面面俱到。据此，我们的标注工作主要落实在句法和语义两个层面：句法层面标注组块边界[4]和论元成分，如 NP1、NP2、NP3；语义层面标注论元的语义角色[5]，如施事 A、受事 P、与事 D 等。除了这两个重要层面以外，我们还需要标注谓语部分的形式类型，如根动词 VROOT、趋向动词 DV 和形容词 A 等。最后，我们标注出具有区别性意义的谓词语义类（如“当作义”、“给予义”和“成为义”等）以及有关成分的语义特征（如“有生”）。例如：

[3] 数据统计来自我们对北大中文树库中“把”字句的标注文本，共有 2441 句。

[4] 我们将“把”字句分为三个组块部分，包括“把”字前的组块 $X_{1...n}$，“把”字后的组块 Y，以及谓语部分 $Z_{1...n}$。我们采用 IOB2 序列表示法（Sang and Veenstra 1999）[4]来对组块进行序列标注。其中，“B”表示当前词是一个组块的开始，“I”表示当前词在一个组块中，而“O”表示当前词不在任意一个组块中。

[5] 语义角色的标注标准主要借鉴了袁毓林（2002）提出的语义角色标注体系。

老干部	!n	B-X1#NP1:A
把	!pba	B-BA*1
经验	!n	B-Y#NP2:P
传授	!v	B-Z1#VP=V_ROOT@AlterKnowledge
给	v	B-Z2#VP=GEI
新	!a	B-Z3#NP3:D
干部	!n	I-Z3#NP3:D

2.3 “把”字句的计算步骤

在标注语料的基础上，我们将机器对“把”字句的理解具体化为框架识别与自动分类这两个任务。基于识别和分类的结果，我们设计一个自动释义和句式变换程序来对归入相应类别的“把”字句实例生成人工语言释义和可变换的句式实例。如下所示：

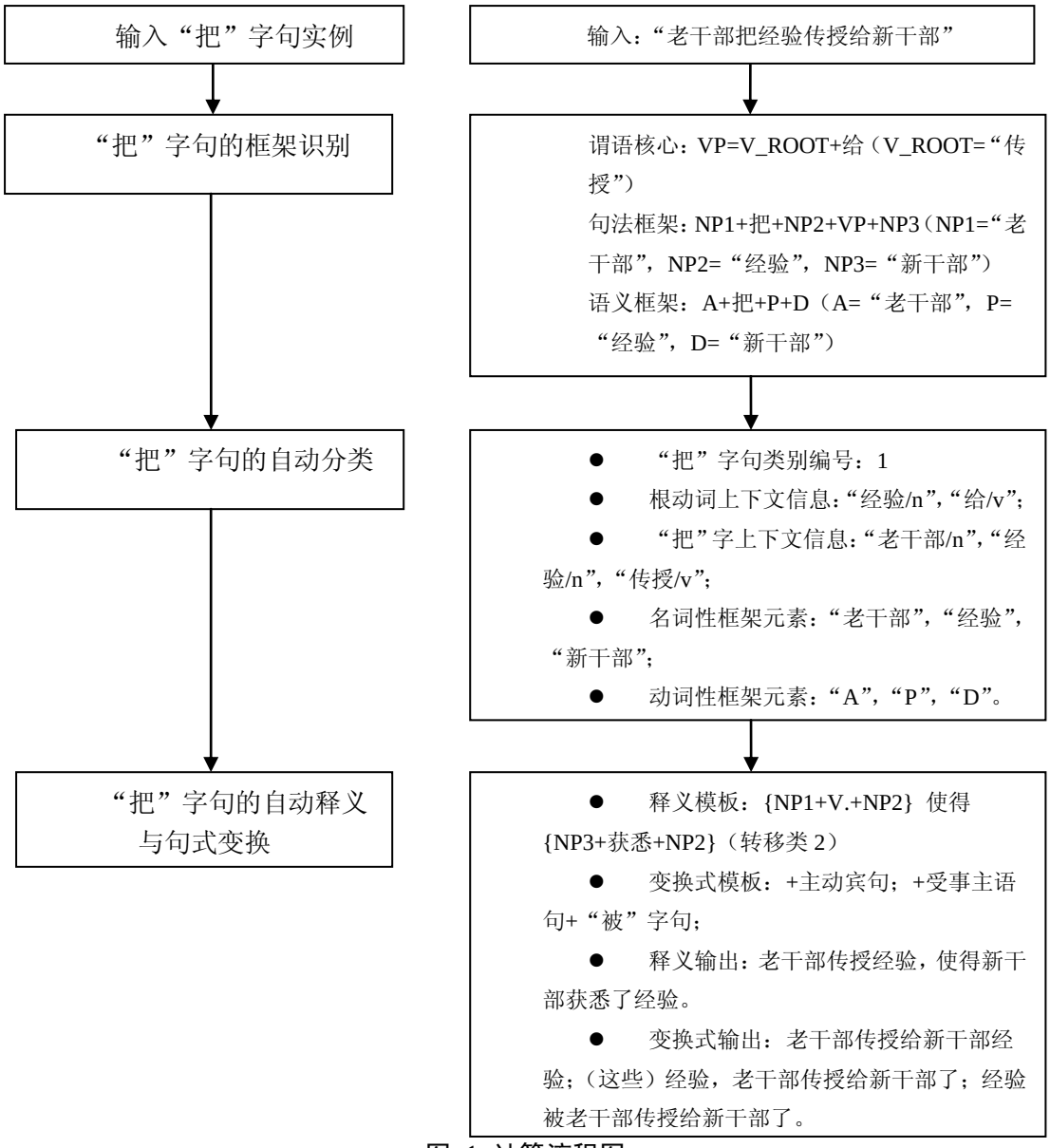

图 1 计算流程图

3. “把”字句的自动识别

关于 “把”字句的自动识别，我们分别采用浅层分析与深层分析这两种技术。浅层分析一般是分析句子的局部，适用于分词、命名实体识别等任务。近年来，组块（chunks）分析为浅层句法分析带来新的思路。Abney（1991）提出组块是介于单词与语句之间的单位，它将整句划分为不同的部分，可以避免完全句法分析中的诸多难题，如歧义问题。如果不考虑语法单位之间的语法关系，我们可以将其句法框架和语义框架看作是一个线性序列结构。这样，我们就可以用组块分析的方法来识别“把”字句。另一方面，我们还对“把”字句进行深层的句法分析。虽然深层句法分析在鲁棒性和效率等方面不如浅层句法分析，但是复杂语法对复杂的语言现象有更好的把握能力。下面，我们将分别介绍组块分析与深层分析的实验方法和实验结果，并从二者的比较中得到最优的方案。

3.1 基于组块分析的“把”字句识别

3.1.1 组块分析及算法

组块分析是将句子进行部分的句法分析，从而降低语法分析的难度。在汉语的浅层句法分析中，组块分析主要用于短语边界的识别，尤其是名词短语的自动获取。目前，基于成熟的机器学习算法的组块分析，对序列性数据（sequential data）的处理性能优势明显而被普遍采用。

“把”字句的句法框架与语义框架都是一个线性的序列结构，即框架成分间互不相交、没有叠加。换句话说，任意一个框架元素与其它框架元素都不会共享同一个词。考虑到“把”字句的这种线性结构，我们采用组块分析的方法来对“把”字句的框架成分进行识别。那么，“把”字句的框架识别可以具体化为对给定句子中的词进行序列标注的问题。考虑到“把”字句中框架成分的线性特点，我们采用 IOB2 表达法（Sang and Veenstra 1999）来进行序列标注。通过这种 IOB2 表示法，任意给定一个“把”字句，我们都可以通过这样的表示法来表示其所有的关涉成分；而只要能够对句子的词进行正确的 IOB2 标签分类，就可以实现“把”字句框架的自动识别。

接下来，本文使用了一种结构化的学习算法——条件随机场（Conditional Random Fields, 简称 CRFs）。条件随机场是由 Lafferty 等人（2001）提出的一种统计模型方法，它兼具判别模型和无向图模型的优点：特征设计灵活、无需考虑特征独立性、避免了标记偏执（Label bias）问题[6]。它常用于标注或分析序列性数据，适用于分词、词性标注和命名实体识别等任务。

根据该学习算法，我们依次对给定句子中的词进行分类。分类的依据是该词的上下文特征及已经完成了的前一词的标签分析。本文中使用的特征包括：

（1）给定词的前两个词及后两个词窗口内的一元词特征，含词与词性两种；

（2）给定词的前两个词即后两个词窗口内的二元词特征，含词与词性两种。

3.1.2 实验结果及分析

对于框架识别任务，我们采用准确的词性标注结果作为输入。在本节的实验中，我们

[6] 参考常宝宝《计算语言学》课程讲义。

采用 wapiti 工具包[7]作为框架识别的学习器。[8]鉴于语料包括句法与语义两个层级的标注，我们分别设计句法框架和语义框架的识别实验，并比较二者在召回率和准确率上的区别。

在第一个实验中，我们对“把”字句的句法框架进行识别，即采纳标注为“B-NPx”一类的标签作为预测内容。实验结果如下所示：

表 2 基于句法框架的识别结果

	召回率	准确率	F 值
NP1	68.09	84.33	75.34
NP2	90.14	90.93	90.53
NP3	58.71	84.29	69.21
VP	79.12	90.16	84.28
V_ROOT	86.40	93.37	89.75

在第二个实验中，我们对“把”字句的语义框架进行识别。我们主要将谓词论元的语义信息（标签形如“B-A”）抽象为预测内容，以下是实验结果。

表 3 基于语义框架的识别结果

	召回率	准确率	F 值
A	68.70	83.26	75.28
D	63.16	85.71	72.73
P	93.14	92.23	92.68
L	59.92	79.14	68.20
VP	78.80	90.37	84.19
V_ROOT	86.22	94.37	90.27

以上两组实验结果表明：（1）对“把”字句的句法识别与语义识别的结果相差不大，准确率和召回率都非常接近；（2）“把”的宾语（通常为 NP2/P）较容易识别、准确率高。这说明“把”作为一个功能词，有很强的句法语义标示性；（3）组块分析算法的准确率达到一定的精度，但相比之下，召回率很低，很多正确的关涉成分没有被找到。如果考虑到完全句法分析的信息，有可能会改进框架元素识别的召回率。

3.2 基于完全句法分析的“把”字句识别

3.2.1 完全句法分析及算法

相较于组块分析，完全句法分析将整句剖析（parse）成一棵完整的句法树。也就是说，我们将对自然语言的理解具体化为生成句法树的过程。句法树的生成依赖于背后的形式语法理论。粗略地说，基于上下文无关文法的语言模型在计算句子语义方面不如基于约束的语法系统和基于依存语法的语言模型。在对汉语的形式语法研究中，基于上下文无关语法和依存语法的句法模型占据主流地位，已经开发有较为成熟的树库资源，如宾大中文树库、北京大学的现代汉语树库、清华汉语树库、台北中研院的Sinica汉语依存树库，以及哈尔滨工业大学信息检索研究室从短语结构树库转化来的依存树库等。这些资源的建设为面向大规模真实文本的内容计算的语言知识的挖掘和形式表示等方面的研究提供了真实有效的语料支持。

相比较而言，基于约束的形式语法方面的研究并不多见，大规模的语法资源也只限于雏形。在基于HPSG理论的多国语法开发平台中，美国华盛顿大学的Bender教授主持开发的

[7] 详见 http://wapiti.limsi.fr/。

[8] 鉴于该语料规模有限，为了更加准确地衡量我们的算法，我们报告的实验结果均是十折交叉验证的结果。

矩阵语法（Matrix）包含了一部分汉语语法资源。德国柏林自由大学的Müller教授主持的汉语语法资源正在建设当中。德国萨尔兰大学在矩阵语法的基础上进一步开发汉语语法（MCG）资源。此外，日本东京大学的Miyao教授生成了由宾州汉语树库自动转换而来的HPSG语法树库。

在这些语法资源中，我们选取了较有代表性的宾大中文树库、北大汉语树库、哈工大依存树库以及柏林自由大学的汉语语法库。研究发现，他们对“把”字句的分析主要有三点较大的分歧。

（一）“把”字和“把”后成分之间的关系

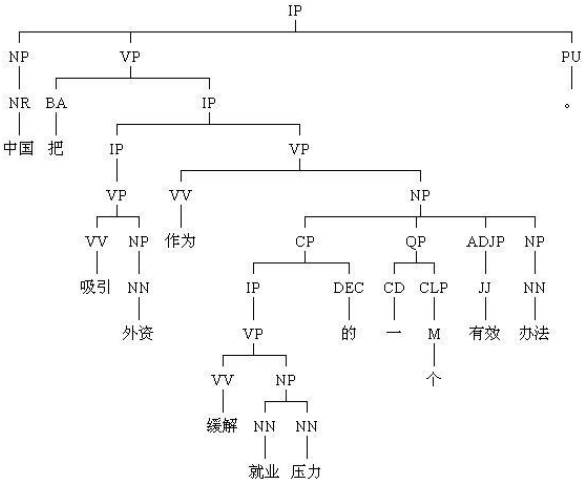

图 2 宾大中文树库例示

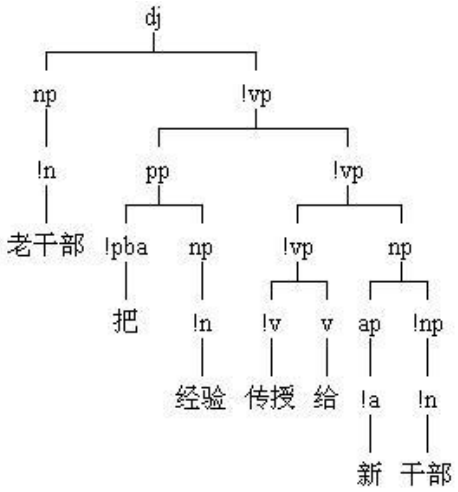

图 3 北大中文树库例示

如上图所示，宾大中文树库将“把”字后面的成分整体看作是“把”所带的小句“IP”。而北大汉语树库将“把”及其后成分看作是一个介宾短语“pp”。前文也提出，考虑到“把”字句的处置义，我们也认同后一种分析，将“把”字与“把”字后名词性成分先分析为一个单位，只不过需要根据“把”字句的类型将“把”分别看作是宾语标记（S1 类“把”字句，如“妈妈把衣服洗干净了”）与主语标记（S2 类“把”字句，如“农活儿把爷爷累病了”）。

（二）论旨角色的配置

“把”字句中，论元成分的语义角色配置是非常复杂的。如果能够在语料中得到正确的论旨角色关系，无疑对理解“把”字句有着至关重要的作用。目前，宾大树库和哈工大依存树库都能给出简单的语义角色标注信息，但是实践证明，目前的标注工作还存在一些问题。

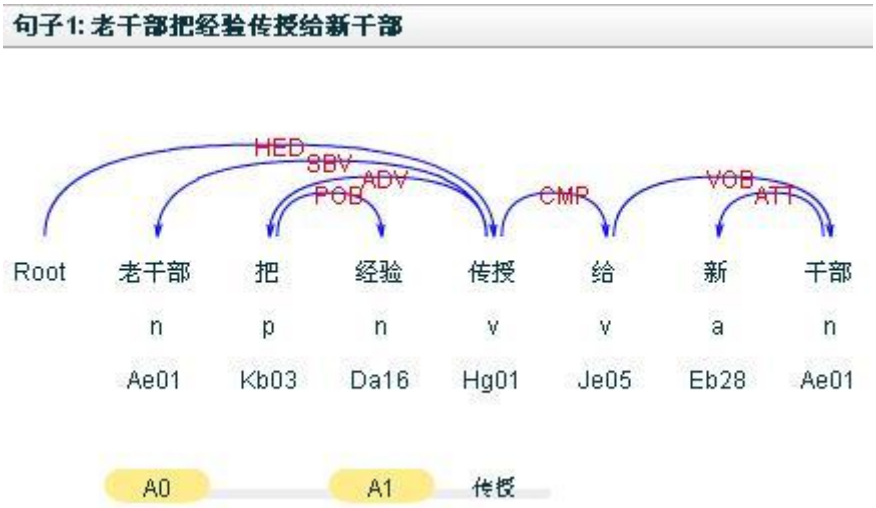

图 4 哈工大依存树库例示

在上面依存树图中，分析器只给出了“老干部”和“经验”这两个论元成分的语义角色信息。但是，“传授”类动词在“把”字句中实际上关涉三个论元成分，图中并没有给出与事的标注。同样，我们在依存树库中还测试了一系列具有复杂语义角色配置关系的句子，如“她把女儿打哭了”，“打哭”分析出来是个连谓结构，而语义角色标注中“哭”是错误的，把“她”标成了 A0（施事），实际上应该是“女儿”哭了。

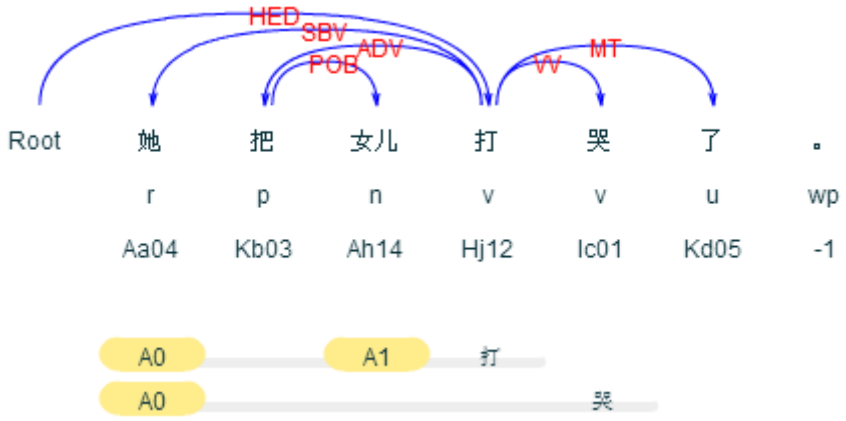

图 5 “她把女儿打哭了”的依存分析

对于这个句子的分析，陈鑫（2011：11-12）将“打哭”类动词看作是连谓结构。他认为，“连谓结构是同属一个主语的多个谓语，这些谓语成分地位相等，只是在时间或空间上不一样，中间可以被逗号分隔，多数谓词都有自己的宾语。”这个定义显然不适合解释“把”字句，因为“打哭”并不是连谓结构，而是述补结构。句子的语义可以还原为“她打女儿”使得“女儿哭了”，对应到我们建立的分类体系中的述补结构类。所以说，我们对“把”字句分而治之的思想是非常必要的，这样才能使得复杂的论旨角色配置对应到相应类别的句子中。

（三）谓词核心关涉成分的约束关系

在基于短语结构语法的宾大树库和北大中文树库中，“把”字句谓词核心所关涉的论元成分通过自底向上的规则组成更大的语法结构。但是，这些语法无法描述所谓的“提宾说”，即动词后名词性成分提前到“把”字后的位置上。对此，基于约束的形式语法可以提供这种长距离的依存范式。在 HPSG 语法中，这种语法现象可以用长距离依存原则来描写，移动后所留下的空位可以用 GAP 特征来表示（Sag and Wasow 1999），具体如下图所示：

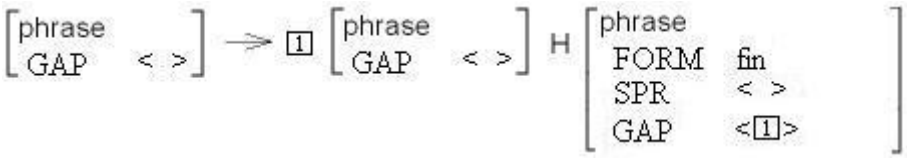

图 6 长距离依存原则

Gao（2000）应用这一原则描述“把”字句中论元成分的长距离依存关系，分析结果证明是有效的。例如，对“我把他抢了”这个句子的分析如下所示：

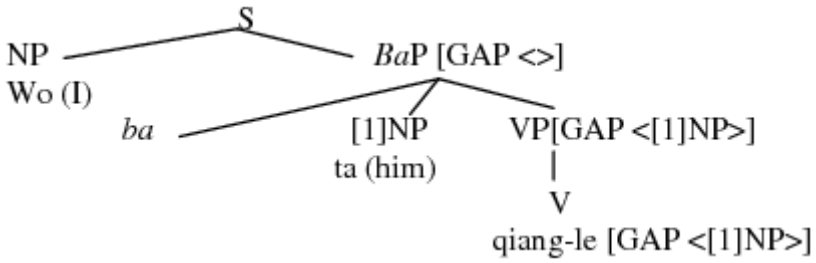

图 7 Gao（2000）的句法分析

进而，柏林自由大学 Müller 教授主持开发的汉语语法系统中还可以对这种句法空位进行自动分析，从而可以自动分析出合格的“把”字句，如下图所示：

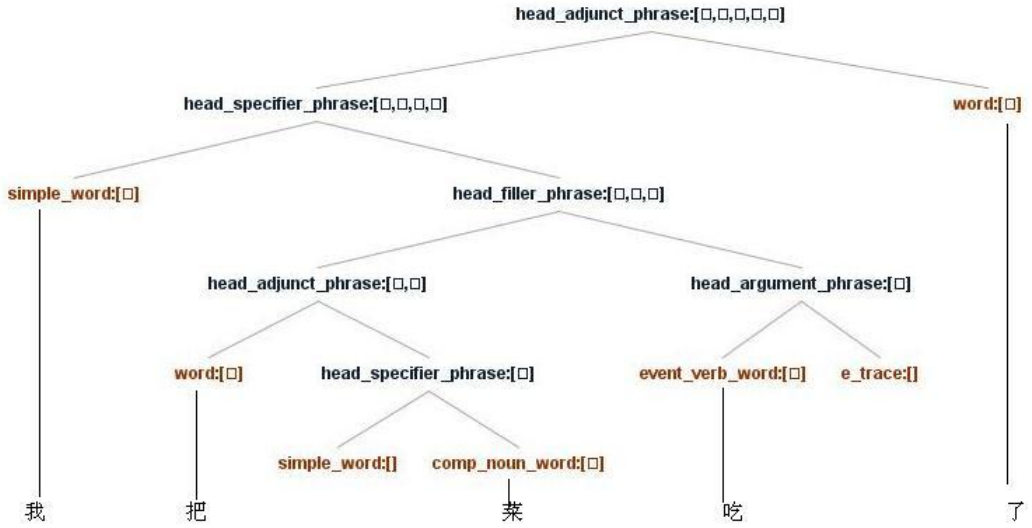

图 8 柏林自由大学 TRALE 语法例示

据此，我们将 HPSG 理论的这种长距离依存的思想结合到我们的短语结构文法中，并设计程序来自动找到“把”字句中的框架成分“NP1”、“NP2”和“NP3”。如下图所示：

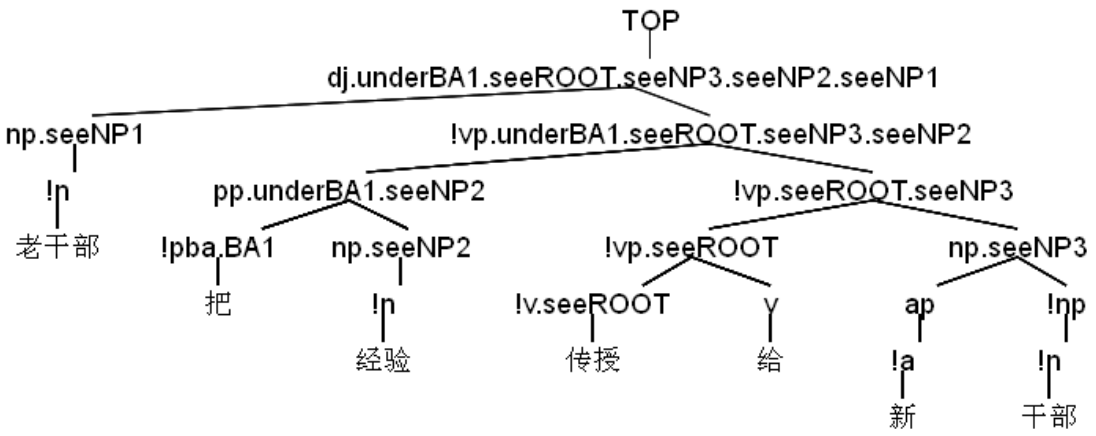

图 5 “老干部把经验传授给新干部”的句法树

如图所示，我们采取一种自底向上、逐层遍历的扫描方法。先从子节点开始寻找“NP3”，如果找到了，就在该节点的上层节点标注“seeNP3”；接着寻找根动词“ROOT”，如果找到，在根动词所在节点标注“seeROOT”，这些标注信息会逐级向上承继；再继续找“NP2”，找

到后在其上层节点标注“seeNP2”；最后寻找“NP1”，找到后在其上层节点标注“seeNP1”。这样，在这句话的父节点上会累积标注出我们找到这些框架成分的信息，也就意味着我们识别出了句子中的框架成分，达到了自动识别的目的。

4.2.2 实验结果及分析

我们在标注文本中随机抽取了 240 句“把”字句作为测试集来进行开放测试。首先来看根据句法框架的识别结果，如下所示：

表 4 基于句法框架的完全句法分析识别结果

标签	召回率	准确率	F 值
xp.seeROOT	80.55	80.71	80.63
xp.seeNP2	96.03	92.44	94.20
xp.seeNP1	78.60	55.21	64.86
xp.seeNP3	78.39	63.85	70.38
xp.seeSP	80.77	65.12	72.10

从上面的结果可以看出，基于深层思想的框架识别结果优于组块分析的结果。其中，召回率的结果要明显优于组块分析的召回率，NP2 达到 96.03%，根动词，NP1 和 NP3 在 80%左右。这说明，完全句法信息的系统的预测要明显优于组块分析的系统。但是，因为找到的框架成分多了，也会对系统的准确率造成一定的影响。目前，该系统对“NP2”的识别准确率最高，达到 92.03%。但是，对根动词、NP1、NP3 和 SP 的识别不如 NP2，分别为 80.71%、55.21%、63.85%和 65.12%。NP1 的识别效果最差，这是因为“把”字句的 NP1 可以不出现，也可以出现在“把”字结构所在小句的前一小句中，所以难以准确地识别。

再来看根据语义角色的识别结果：

表 5 基于语义框架的完全句法分析识别结果

标签	召回率	准确率	F 值
xp.seeROOT	77.01	77.60	77.30
xp.seeP	93.87	92.41	93.13
xp.seeA	77.30	55.45	64.58
xp.seeL	69.79	68.10	68.94
xp.seeD	75.00	52.69	61.89
xp.seeR	32.05	15.43	20.83

总体来看，基于语义角色的识别结果与框架成分的识别结果基本一致。但是，由于语义角色的数量要多于框架成分的数量，在结果的集中度上会受到一些影响。

最后，我们详细考察了识别错误的结果，发现没有得到正确识别的原因主要集中在词性标注错误（25%）、句法成分的识别错误（15%）与句法结构关系的识别错误（20%）上。这与我们的词性标注器和句法分析器的效果有关。可见，我们的句法分析器与北大中文树库的分析结果存在一定的差距，这也是以后需要改进的地方。

5. “把”字句的自动分类

5.1 句式分类算法

对“把”字句的句式分类一般可以转化为一个关于“把”字的词义消歧问题，但传统的词

义消歧方法在解决我们的问题上有很大局限性。最主要的原因是，词义消歧所利用的信息一般是和目标词搭配的词，通常属于词汇语义的范畴；而我们所关心的“把”字句的语义，则是“把”字句的框架分类，属于句法语义的范畴。这样，仅仅通过搭配词的信息，我们无法准确完成对“把”字句的分类。我们的实验也充分的说明了这一点。

虽然具体的词义消歧算法不支持“把”字句的分类，但词义消歧的思想仍然有很大的借鉴意义。在本文中，和词义消歧算法相似，我们采用判别式机器学习算法来对“把”字句进行类别的自动分析，只是在特征提取方面，我们采用的不是词搭配的信息，而是采用我们已经识别出来的“框架元素”。下面是对特征的具体说明：

（1）根动词的上下文信息：根动词的前一个词及后一个词窗口内的一元词特征，含词与词性两种；

（2）“把”字的上下文信息：“把”字的前两个词及后两个词窗口内的一元词特征，含词与词性两种；

（3）名词性框架元素信息：名词性框架元素的尾词及其词性；

（4）动词性框架元素信息：动词性框架元素的首词及其词性。

在自动识别了“把”字句框架信息的基础上，我们可以很方便地提取以上特征，并根据这些特征训练一个分类器，从而实施对“把”字句进行分类。在分类器的选择方面，有很多算法可以考虑，我们采用了支持向量机的算法。[9]另外，用于分类的学习器，我们使用的是 liblinear 线性分类器。[10]

5.2 实验结果及分析

对于语义分类任务，我们将前文自动识别实验的结果作为输入，并选取根动词和“把”字的上下文信息，以及名词性和动词性框架元素的信息作为分类的重要特征。由于前文分别采用了组块分析和完全句法分析的方法进行了识别实验，解析来的分类实验也要基于这两组不同的数据。

首先是通过组块分析得到识别成分的语义分类的结果：

表 6 基于组块分析的语义分类结果

特征	精度
自动句法框架信息	67.21
自动语义框架信息	67.21

通过实验，我们发现，基于自动的句法框架信息和自动的语义框架信息对“把”字句框架分析没有影响，系统精度一致[11]。

接下来是基于完全句法分析得到识别成分的语义分类结果：

表 7 基于完全句法分析的语义分类结果

特征	精度
自动句法框架信息	61.79
自动语义框架信息	60.53

从上表的数据可以看出，在基于完全句法分析得到的框架信息的基础上的语义分类结果

[9] 支持向量机（Support Vector Machine，简称为 SVM）是一种监督式学习的方法，可广泛地应用于统计分类以及回归分析。
http://zh.wikipedia.org/wiki/%E6%94%AF%E6%8C%81%E5%90%91%E9%87%8F%E6%9C%BA

[10] Liblinear 是一个用于大规模数据分类的开源库，支持逻辑回归和向量机。（Fan et al., 2008）

[11] 由于分类问题的判断标准在于分类的准确度，并没有召回的文本。

（准确率为 61.79%）并没有大幅度的改进，反而还不如基于组块分析得到的框架信息上的语义分类结果（准确率为 67.2%）。

这两组实验的结果都不够理想。通过对分类错误结果的分析[12]，我们发现造成“把”字句的分类错误的原因主要有以下几点：

（1）分词结果不一致。比如“流言把他击倒”，树库中倾向于分析为“击/v 倒/v”，但是我们采用的分词系统将“击倒”切分为一个动词。这样就既影响识别结果，又影响分类结果。在 41 句错误结果中，有 4 句属于这类错误。

（2）识别结果不一致。例如下例中，测试文本中的根动词和 NP2 的识别都与标准答案中的不一致。在 41 句错误结果中，有 27 句都属于这类错误。

（3）分类结果错误。这类句子中，句子成分与句子结构关系的识别结果都是正确的，只有分类结果是错误的。在 41 句错误结果中，有 8 句属于这类错误。

（4）标准答案的分析结果错误。在我们的语料中，包括一部分“将”字句，我把这些“将”字句也标作“把”字句，但是有些“将”字句在标准答案中没有标注上。在 41 句错误结果中，有 2 句都属于这类错误。

综合来看，真正由本系统造成的分类错误只有 8 例，占 20%左右。大部分的错误结果都是与识别错误直接相关。由此，我们需要在以后重点改进识别结果的召回率与准确率。

6. “把”字句的自动释义与句式变换程序

在前面的自动识别与自动分类的基础上，我们设计一套“把”字句自动释义与句式变换程序。该程序按照如下的计算步骤来实现对“把”字句的自动释义与句式变换：

【1】系统在文本框中输入一句“把”字句；

【2】系统后台对该“把”字句进行自动识别与句式分类，并将识别结果和分类结果保存到临时文件中；

【3】根据分类结果，系统后台找到相应类别的释义模板和句式变换模板，将识别出的框架成分分别填入到相关的释义模板中，得到一个释义结果和一组变换式；

【4】系统将该释义结果和变换式输出到文本框中；

【5】系统再给出句法分析的结果，并以一棵树的形式显示出来。

根据以上计算步骤，我们可以对“把”字句实现自动释义。下面是“老干部把经验传授给新干部”的自动释义和句式变换过程：

第一步，经过组块识别或完全句法分析，得到这句话的句法框架信息和语义框架信息。例如，句法框架信息有：NP1=“老干部”，NP2=“经验”，NP3=“新干部”，ROOT=“传授”；语义框架信息有：A=“老干部”，P=“经验”，D=“新干部”，ROOT=“传授”。

第二步，根据框架信息和分类类别之间的对应关系（即分类模型），机器自动将这句话归入到第 1 类，在“把”字的标记上标为“BA1”。在完全句法分析中，这个类别信息会跟着“把”字向上传递到根节点。

第三步，根据第 1 类“把”字句的释义模板和句式变换模板，将框架成分代入相应的类别中。例如，释义模板是“NP1+VP+NP2，使得+NP3+获悉+NP2”，输出结果是“老干部传授经验，使得新干部获悉经验”；句式变换模板是“NP1+VP+NP2+GEI+NP3；NP2+被+NP1+VP+GEI+NP3；NP2，NP1+VP+GEI+NP3”，输出一组句式变换式：“　老干部(NP1)传授(VP)经验(NP2)给(GEI)新干部(NP3)；经验(NP2)被老干部(NP1)传授(VP)给(GEI)新干部(NP3)；经验(NP2)，老干部(NP1)传授(VP)给(GEI)新干部(NP3)”。这样，我们就实现了对“把”

[12] 在随机抽取的 100 句分析结果中，有 41 句的分类结果错误。

字句的自动释义和句式变换程序，输出结果如下图所示：

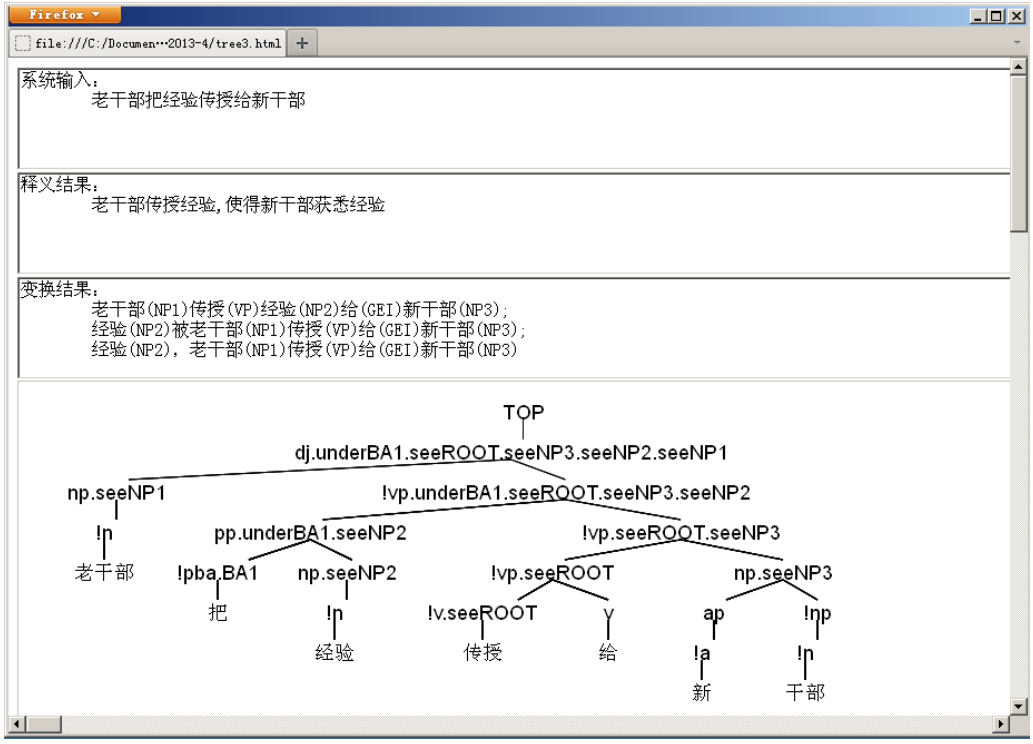

图 6 计算程序实例演示结果

需要说明的是，这部分自动生成的语义解释也许并不流畅，因为我们是采用的人工语言的释义模板。如果要达到自然语言的流畅度，还需要结合大规模的基于 N-Gram 的语言模型的训练。

7. 结论

本文通过对“把”字句的计算分析，实现了一个可以对“把”字句进行框架识别、自动分类和自动释义与句式变换的程序。首先，我们分别采取了基于组块分析的方法和基于完全句法分析的方法来对“把”字句进行框架识别。研究发现，基于完全句法分析的识别方法对“把”字句框架成分的召回率高于基于组块分析的方法。但是，由于预测的数量增多，准确率不如基于组块分析的方法。只有对 NP2 的识别，基于句法分析的方法在召回率和准确率两个方面都优于基于组块分析的方法。前者对 NP2 的召回率达到 96.03%，准确率达到 92.44%；而后者对 NP2 识别的召回率是 90.14%，准确率是 92.03%。接下来，我们采用判别式机器学习的方法来对“把”字句自动分类。实验结果表明，在组块分析的识别基础上的自动分类的准确率是 67.21%，而基于完全句法分析的自动分类的准确率是 61.79%。最后，在自动识别与分类的基础上，我们根据释义模板和变换模板设计了一个“把”字句的自动释义与句式变换程序。当输入端输入一个“把”字句，我们在输出端给出该句的释义结果和相应类别的变换式，从而为机器翻译等应用研究提供帮助。例如，我们可以对 “把”字句的自动释义和句式变换结果进行翻译，再通过自信度计算等策略计算出最为理想的翻译结果。

参考文献

[1] 崔希亮 1995 《“把”字句的若干句法语义问题》,《世界汉语教学》第 3 期。
[2] 陈鑫 2011 《基于主动学习的汉语依存树库构建》，哈尔滨工业大学硕士论文，11-12。

[3] 郭锐 2003 《把字句的语义构造和论元结构》,《语言学论丛》，第 28 辑，北京:商务印书馆。
[4] 邵敬敏 1986 《把字句及其变换句式》，《研究生论文选集 • 语言文字分册》，南京：江苏古籍出版社。
[5] 王力 1943/1985 《中国现代语法》,《王力文集》第二卷，山东教育出版社。
[6] 王璐璐 2013 《基于变换的“把”字句自动释义研究》，北京大学博士论文。
[7] 薛凤生 1987 《试论“把”字句的语义特性》,《语言教学与研究》第 1 期。
[8] 叶向阳 2004 《“把”字句的致使性解释》,《世界汉语教学》第 2 期，25-39。
[9] 袁毓林 1989 《论变换分析方法》,《汉语学习》第 1 期，7-13。
[10] 袁毓林 1998 《语言的认知研究和计算分析》，北京大学出版社。
[11] 袁毓林 2002 《论元角色的层级关系和语义特征》,《世界汉语教学》第 3 期，5。
[12] 袁毓林 2007 《语义角色的精细等级及其在信息处理中的应用》,《中文信息学报》，第 21 卷，第 4 期。
[13] 袁毓林 2008 《基于认知的汉语计算语言学研究》，北京大学出版社。
[14] 詹卫东 2004 《论元结构与句式变换》,《中国语文》第 3 期，209-221。
[15] 张伯江 2000 《论“把”字句的句式语义》,《语言研究》第 1 期。
[16] 张旺熹 2001 《“把”字句的位移图式》,《语言教学与研究》第 3 期。
[17] 周强、张伟、俞士汶 1997 《汉语树库的构建》,《中文信息学报》第 4 期，42-51。
[18] 朱德熙 1982 《语法讲义》，商务印书馆。
[19] 朱德熙 1989 《变换分析中的平行性原则》，《中国语文》第 2 期。
[20] Abney, S. 1991. Parsing by Chunks. In Robert C. Berwick, Steven P. Abney, and Carol Tenny (eds.), *Principle-Based Parsing*. Dordrecht: Kluwer Academic.
[21] Bender, E. 2000. The Syntax of Mandarin ba: Reconsidering the Verbal Analysis. *Journal of East Asian Linguistics* 9, 100–145.
[22] Gao, Qian. 2000. *Argument Structure, HPSG and Chinese Grammar*. Ph. D.thesis, Ohio State University.
[23] Jurafsky D,Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics, 2000. 冯志伟,孙乐译. 自然语言处理综论. 冯志伟序. 北京:电子工业出版社, 2005.7-412.
[24] Lafferty, J., McCallum, A., Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. 18th International Conf. on Machine Learning*. Morgan Kaufmann, 282–289.
[25] Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*, University of Chicago Press, Chicago, IL.
[26] Lipenkova, J. 2010. *A HPSG representation of causativity in the Chinese ba-construction* (Poster). The 17th International Conference on Head-Driven Phrase Structure Grammar, Universit éDenis Diderot Paris.
[27] Lipenkova, J. 2011. *Lexical licensing and obligatory event modifiers in the Chinese ba-construction*. CSSP.
[28] Sag, A. and Wasow, T. 1999. *Syntactic Theory: A Formal Introduction*. CSLI Publications.
[29] Sang**,** E.T.K. and Veenstra J. 1999. Representing Text Chunks. In: *Proceedings of EACL Conference* (EACL 1999).
[30] Zou, Ke. 1995. *The Syntax of the Chinese Ba-constructions and Verb Compounds: A Morpho-syntactic Analysis*. Doctoral Dissertation. University of Southern California.