构建大规模的汉语语块库

周　强
智能技术与系统国家重点实验室，清华大学计算机系，北京 100084

詹卫东
北京大学中文系，北京 100871

任海波
上海师范大学国际文化交流学院，上海 200234

摘要：本文介绍了构建 200 万字的汉语语块库的主要工作，包括设计语块标注体系、总结语块标注规范和协调语块加工流程等，分析了我们的标注体系与英语的 CONLL-2000 语块任务的主要差异，并提出了对现有标注体系的进一步理论思考和在现有语块库上的一些应用设想。

1 引言

构建大规模标注语料库是语料库语言学发展的重要基础。在英语方面，百万词次规模的词性标注语料库——Brown 语料库的建成，直接促使了基于统计的词性标注模型：HMM 模型和自动标注算法：Viterbi 算法的提出和完善。大规模的句法树标注语料库——Penn 树库的建立，则为许多基于统计的自动句法分析模型提供了基础的训练素材。同时，作为一个统一的训练和测试平台，也为不同分析算法处理性能的评估提供了客观的依据。近几年来，随着部分分析技术的不断发展和应用范围的不断扩大，对处于中间层次的语块（chunk）标注语料库的开发也越来越受到重视，出现了一些较大规模的语块标注语料库，如 CONLL-2000 的语块库[TB00]等。

在汉语方面，经过近几年的研究，已经建立了几个较大规模的切分和词性标注语料库，包括清华大学的 200 万字的平衡语料库和北京大学与富士通合作开发的人民日报语料库。在树库构建方面，也已取得一些成果，包括清华大学的汉语测试树库[ZS99]、美国宾州大学的 UPenn 树库[XP00]和台湾中研院的树库项目[HCC00]。但对语块标注和部分句法分析的研究还比较少。

本文介绍了我们在汉语语块标注体系设计和大规模语块库构建方面进行的一些初步探索。下面的第 2 节比较详细地介绍了我们的语块描述体系，并与 CONLL-2000 的标注体系进行了比较，分析了两者的不同之处。第 3 节介绍了我们的语块库构建工作，包括基础语料库资源、语块标注规范和语块加工流程等，并给出了一些基本的语块库统计数据。第 4 节进一步分析了语块与论元结构的关系以及语块与韵律结构的关系等。最后的第 5 节展望了在现有的语块库上可以进一步进行的一些句法分析和知识获取研究设想。

2 语块描述体系

Abney(1991)最早提出了一个完整的语块描述体系。他把语块定义为句子中一组相邻的属于同一个 s-投射（s-projection）的词语的集合，建立了语块与管辖约束（GB）理论的 X-bar 系统的内在联系，从而奠定了这个语块描述体系的比较坚实的理论基础。在此前后，一些应用系统的研究重点则主要集中在名词短语的识别上，其中包括基本名词短语（BaseNP）([Chu88], [RM95])和最长名词短语（MNP）([LZ95], [ZSH00])。在其他语块或基本短语方面的研究则比较少。最近比较完整的工作是 Buchholz & al.(1999)。他们探索了 NP, VP, PP 和 ADJP 等基本短语的自动识别方法。另外，Veenstra(1999)也识别了 NP, VP 和 PP 块。他们的研究为 CONLL-2000 提出的语块共享研究计划打下了基础。

去年举行的自然语言学习国际会议（CONLL-2000）提出的语块共享任务（Chunking Shared Task）旨在开发出一个大规模的英语语块库，为基于统计的不同部分分析方法的探索提供统一的训练和测试库。他们采用了 Abney 的语块描述框架，并对一些语块进行了分解

和细化，其中的一些差异可以从下面的例子中看出来（其中例句 1 采用了 Abney 的标注体系）：

(1) [He] [reckones] [the current account deficit] [will narrow] [to only $1.8 billion] [in September].

(2) [NP He] [VP reckones] [NP the current account deficit] [VP will narrow] [PP to] [NP only $1.8 billion] [PP in] [NP September].

语料则取自 Penn 树库的华尔街日报（WSJ）部分。利用自动程序将分析树标注文本直接映射成不相交、无嵌套的语块标注文本，并保留了原来的大部分句法成分标记。目前抽取的语料规模约为 30 万词，平均每个语块包含 2 个词。表 2 列出了其中最常见的几个语块的信息描述，有关的详细资料可参阅[TB00]。

表 2 CONLL-2000 的常见语块描述

语块标记	语块描述
NP	名词短语
VP	动词短语
PP	介词短语（大部分情况下只包含一个介词）
ADVP	副词短语
SBAR	小句（subordinated clause）(大部分情况下只包含一个从属连词)
ADJP	形容词短语

表 1 我们的语块标记集

语块标记	语块描述
S	主语短语
P	述语短语
O	宾语语块
J	兼语语块
D	状语语块
C	补语语块
T	独立语块
Y	语气块

我们从 2000 年 3 月起，开始进行大规模汉语语料库的语块标注研究。最初的设想是通过语块划分和标注，描述一个句子的基本结构骨架，从而为进一步构建汉语树库，进行深层的句法分析和知识获取打下基础。遵循以下~~几~~两条原则：

- 穷尽性——在完成语块标注的句子中，任何一个词都必须无遗漏地进入某个语块。
- 线性——在完成语块标注的句子中，全部语块将形成一个线性序列，即没有嵌套。

我们设计了包含 8 个标记的语块标记集（详见表 1）。下面是一个具体标注实例：

[D 自/p 古/t 以来/f ，/， [S 人类/n [D 就/d [P 重视/v [O 档案/n 的/u 保存/vN 和/c 利用/vN ，/， [P 设置/v [O 馆库/n 、/、 [P 选派/v [O 专人/n [P 进行/v [O 管理/v 。/。

从语块描述内容上看，两个语块库的差异还是很明显的。CONLL 的语块强调对局部的句法相关词语的描述，侧重于从底向上地把句子分隔成不同的基本短语；我们的语块则强调对句子整体功能块的描述，侧重于自顶向下地描述句子的基本骨架。这种差别使得 CONLL 的语块一般比较简单，平均每个块只包含 1-2 个词语，而我们的语块则比较复杂，有的语块甚至包含 10-20 个词语。但两者具有很好的信息互补性。在适当的条件下，将两者的描述信息进行合并，形成分层次的语块描述体系，并构建相应的语块库，将是一个很有意义的研究课题。

3 语块库构建

3.1 基础语料库

我们的语块加工对象是清华大学的 200 万汉字的平衡语料库（ThCorp）。它的主要语料来源是 90 年代的现代汉语书面语以及准口语（包括剧本、谈话录、演讲录等）的真实文本，按文体分为文学、新闻、学术、应用四类。经过自动切词、词性标注和人工校对，已经形成了准确度很高的切分和词性标注精加工文本，为进一步进行语块信息标注打下了很好的基础。表 3 列出了目前的 ThCorp 的一些基本统计数据，其中‘词项数’包括汉语词和标

点符号，‘汉字数’包括汉字和汉字标点。

表 3 ThCorp 切分和词性标注语料库的基本统计数据

文体	文件数	句子数	词项数	汉字数
学术	29	9846	273017	447288
新闻	376	16921	427649	674566
应用	258	4302	88452	144027
文学	295	38258	740445	1018839
合计	958	69327	1529563	2284720

3.2 语块标注规范

大规模语料库的标注是一个庞大的语言工程项目，需要投入大量的人力和物力。因此，预先制定一部比较完善的语料标注规范，对保证标注结果的规范性和一致性将起到重要作用。但真实文本中涉及到的语言现象又是非常复杂的，不可能通过一部规范就能完全包括。因此，比较好的处理思路是在标注过程中不断发现新问题，对现有规范进行补充和修订，使之能更好地符合新的语言事实。经过不断摸索，我们已初步形成了一套比较完善的汉语语块标注规范，基本上覆盖了目前语料库中遇到的各种语言现象。下面简单地列出其中状语块的基本规范条款，有关的其他详细资料，可参阅[Th00]。

1．副词性成分（词性标记为 d,dB,dD,dN）连续出现作状语，可以整体标注为一个状语语块‘[D’，其他不同类成分连续出现作状语，都必须分别单独标出状语块。

2．名词直接作状语，需单独标注状语块标记‘[D’。

3．动词直接作状语，需单独标注状语块标记‘[D’。

4．形容词直接作状语，需单独标注状语块标记‘[D’。

5．数量词作状语，需单独标注状语块标记‘[D’。这里的数量词主要有：半年、半日、半晌、半天、多年、一辈子、一会、一会儿等。

6．介词结构、方位结构和“地”字结构、数量结构等成分在句中作状语（我们称之为“复杂状语”结构），需单独标注状语块标记‘[D’。特别应注意它们与上面的简单状语连用的情况，这时每个状语块都应显性标注，比如方位结构、“地”字结构状语的左边界，介词结构作状语的右边界等。

3.3 语块加工流程

目前的所有语块信息都是由人工标注的。利用 WORD 编辑器中的宏命令定义不同的快捷键，可以做到每个语块通过一键输入，大大提高了标注效率。初步统计显示，最初的标注速度约为每小时处理 1200 个词。随着对标注规范和加工过程的不断熟悉，标注速度不断提高，1 至 2 月后基本上可以达到每小时处理 2400 个词。

为了保证标注结果的质量，我们设计了两级检查机制。首先，依据语块标注规范，开发自动检查程序，发现大部分不合规范的标注语块，提供标注者进一步确认或修改。这个过程重复数次后，可以大大减少标注“硬伤”。然后，对标注结果进行随机抽样检查，发现并改正遗留的标注错误，直至最终标注质量达到要求为止。

3.4 语块库基本统计

表 4 列出了现有语块库的基本统计数据，包括不同语块总数及语块中的词语分布。表 5 进一步计算了具有不同数目的词语的语块的分布特征，以 5 为界分为 4 个区间：1) 词数<5, 2) 5<=词数<10, 3) 10<=词数<15, 4) 15<=词数。从中可以看出不同语块的分布特点：

- 语气块定义为句尾的一个或多个语气词。由于汉语里多个语气词连用的情况很少，因此其平均词长最小，为 1.01。

- 汉语句子的述语块大多由谓词性成分充当，在我们的标注规范中对它们进行了严格规定，其词语数都不超过 5 个。这些分布特点在两个表中都有很好的体现（词数<5 的语块占 99%以上，平均词长为 1.31）。
- 状语块和补语块的平均词长约为 2，90%以上的语块中的词数都小于 5，表明汉语真实文本中复杂状语和补语出现的频度不是很高。由于它们一般都有明显的边界标志（介词、方位词、助词‘地’、助词‘得’等），因此自动识别难度不太大。
- 兼语块、主语块和宾语块得平均词长较大，特别是宾语块更达到 4.13。主要原因是其中往往包含了复杂的定语。它们是自动识别的难点所在。
- 在我们的标注体系中，独立语块的内容比较杂，包括句子中的插入语、应答语、呼语、同位性插入成分、句中的补充说明部分（一般在括号内）、句首的序号等，因此分布比较特殊。如何对其中的不同情况进行分化处理，将是以后的一个研究课题。

表 4 不同语块的词语分布统计

语块类别	语块总数	词语总数	平均词长
主语	99121	251041	2.53
述语	179605	236104	1.31
宾语	109362	452211	4.13
兼语	5715	12338	2.16
状语	156000	321254	2.06
补语	3113	6431	2.07
独立	5649	14414	2.55
语气	12111	12225	1.01
合计	570676	1306018	2.29

表 5 具有不同长度词语的语块的分布统计

语块类别	语块总数	词数 [0,5)	比率 (%)	词数 [5,10）	比率 (%)	词数 [10,15）	比率 (%)	词数 [15,∞）	比率 (%)
主语	99121	85208	85.96	11023	11.12	1939	1.96	951	0.96
述语	179605	178545	99.41	862	0.48	144	0.08	54	0.03
宾语	109362	75745	69.26	24569	22.47	5888	5.38	3160	2.89
兼语	5715	5134	89.83	482	8.43	70	1.22	29	0.51
状语	156000	141060	90.42	11863	7.60	2151	1.38	926	0.60
补语	3113	2857	91.78	219	7.04	31	1.00	6	0.19
独立	5649	4984	88.23	388	6.87	136	2.41	141	2.49
语气	12111	12111	100.00	0	0.00	0	0.00	0	0.00
合计	570676	505644	88.60	49406	8.66	10359	1.82	5267	0.92

4 语块标注的进一步思考

语法分析的主要内容是语句的结构问题，需要弄清整体中各组成部分之间的关系。语法分析包括以下两个步骤：1）切分分析，即如何把一个语言结构体，如句子、短语等切分为若干组成成分；2）关系分析，即如何分析、整理各成分之间的关系。无论是切分分析还是关系分析，都有许许多多的可能性供我们选择。我们目前进行的语块标注探索，就是希望从中选出一种客观上能较好地反映语言结构的本质，主观上又比较容易被人理解和掌握的句子结构分析和描述方法。从目前的大规模语块库构建实践看，基本上达到了预期目标。在此基础上，我们希望进一步探索语块与论元结构（argument structure）和韵律结构（prosodic structure）之间的内在联系，从而以语块描述体系作为出发点，建立汉语的句法、语义、语音分析的紧密结合体。

论元结构是指词项及其所属的子语类所构成的介于词汇语义和句法之间的一种结构关

系，通常在文献中讨论最多的是作述语的动词和论元之间的结构关系。论元这一概念，从广义上说就是带有论旨角色（thematic role）的名词组，而论旨角色是述语所固有的子语类，通常表示述语所表达事件所涉及的主体、客体等。动词与论旨角色之间形成论旨关系，论旨角色由动词指派给相关的名词组，就成为论元。论元与述语构成论元结构，最终再反映到句法结构中，形成句法关系，相应的论元就担当句子的主语、宾语或其他成分。Alsina(1996)对上述结构关系进行概括和抽象，并与词汇功能语法（LFG）[KB82]中的功能结构（f-structure）和成分结构（c-structure）相结合，形成了图 1 所示的结构关系图。每个不同的结构层次通过对应原则（correspondence principles）建立联系，限制形成合格的句子。

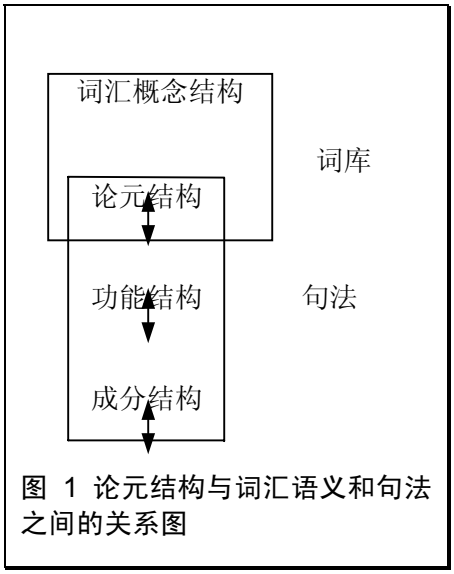

图 1 论元结构与词汇语义和句法之间的关系图

我们目前语块描述的信息基本上相当于图 1 中的功能结构层次。这样，如果能从现有语块库出发，深入分析不同功能语块与句子核心谓词的论元之间的对应关系，就有可能在从句法到语义的分析过程中大大前进一步，从而为进一步进行基于句子（或段落、篇章）的语义分析和知识推理打下很好的基础。

自然语音中的韵律结构是包含有不同韵律信息的层次结构，主要包括：1）韵律词一级的韵律信息，如多音字、变调、词重音等；2）短语语音间隔；3）句子重音等。它们的正确识别对提高文语转换系统的性能有重要作用。Abney(1992)的研究结果表明，他所定义的语块与英语韵律结构的ϕ-短语之间可以建立了很好的一一对应关系。而我们的语块划分由于从句法功能出发，可能与一般的韵律描述有较大差异。但许多研究成果也显示出，韵律结构和句法功能之间有千丝万缕的联系。因此，按照韵律结构分布特点，对现有语块库进行适当改造，合并较小的功能块，分解较大的功能块，可望形成与韵律结构分布基本一致的韵律语块库，从而为基于统计的韵律结构自动识别模型，提供大量有用的训练数据。

5 结语

利用人工标注和机器检查相结合的方法，在现有的切分和词性标注语料库的基础上，我们加工完成了 200 万字规模的汉语语块库。本文简要介绍了在这一过程中的一些初步研究成果，包括汉语功能语块标注体系和经过不断完善的语块标注规范。并通过与英语的 CONLL-2000 共享语块任务的比较，突出了我们工作的特色所在。目前，我们正在此语块库上进行一些新的句法分析和知识获取探索：

- 汉语语块的自动识别：利用机器学习技术，提取有用的识别特征，训练形成有效的语块自动识别模型。
- 动词搭配知识的自动发现：利用语块标注信息，自动学习有用的知识，形成高效的汉语动词语法搭配和词汇搭配自动发现工具。
- 基于语块描述的句法分析器：利用句法分析器，从语块标注串出发，向下分析复杂语块的内部结构，向上形成句子的整体结构，完成对句子的完整层次结构分析。

这些工作继续贯彻了我们最初制定的“标注语料库、自动分析器和语言知识自动发现工具三位一体，同步发展，相互促进，共同提高”的总体研究思路。随着研究工作的不断深入，希望能逐步建立起基于大规模真实文本语料库的汉语句法语义计算平台。这将是我们的长期研究目标。

6 致谢

黄昌宁教授最初提出了语块标注的设想，并亲自标注了大量真实文本例句，积累了宝贵的经

验。靳光谨博士和周明博士对语块标注规范提出了许多建设性意见。下列同学参加了语块库的人工标注和校对工作：清华大学的祝安顺、杨晓明、黄光斌、罗萍、刘淑菊，北京大学的姚静仪、孟贵贤、崔玉珍、黄晖菁、宋作彦、陈园媛、姜南、杜轶、王皓冰、曾汀燕。这里一并表示感谢。本项研究得到国家自然科学基金（项目号：69903007）、国家 973 基金（项目号：G1998030507）和清华大学骨干教师基金资助。

参考文献

[Abn91] Steven Abney(1991). “Parsing by Chunks”, In *Robert Berwick, Steven Abney and Carol Tenny (eds.) Principle-Based Parsing, Kluwer Academic Publishers*.

[Abn92] Steven Abney(1992). “Prosodic Structure, Performance Structure and Phrase Structure”, *Proceedings of Speech and Natural Language Workshop*, pp. 425-428. Morgan Kaufmann Publishers, San Mateo, CA.

[Als96] Alex Alsina (1996). *The Role of Argument Structure in Grammar: Evidence from Romance*. CSLI Lecture Notes No. 62, CSLI Publications: Stanford, California, USA.

[BVD99] Sabine Buchholz, Jorn Veenstra and Walter Daelemans (1999). “Cascaded grammatical relation assignment”, In *Proceedings of EMNLP/VLC-99,* Association for Computational Linguistics.

[Chu88] Kenneth Church(1988). “A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text.” In: *Proceedings of Second Conference on Applied Natural Language Processing*, Austin, Texas, 136-143.

[HCC00] Chu-Ren Huang, Feng-Yi Chen, Keh-Jiann Chen, & al.(2000). “Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface”, *Proceedings of the Second Chinese Language Processing Workshop, HongKong*. 29-37.

[KB82] Ronald Kaplan and Joan Bresnan (1982). “Lexical-Functional Grammar: A Formal System of Representation”, In Joan Bresnan (ed.) *The Mental Representation of Grammatical Relations*. 173-281. MIT Press, Cambridge, Mass.

[LZ95] 李文捷，周明等(1995).“基于语料库的中文最长名词短语的自动提取”，陈力为、袁琦主编，《计算语言学进展与应用》，北京：清华大学出版社，119-124.

[RM95] Lance A. Ramshaw and Mitchell P. Marcus. (1995). “Text chunking using transformation-based learning”, In *Proceedings of Third ACL Workshop on Very Large Corpora*, Association for Computational Linguistics.

[TB00] Erik F. Tjong Kim Sang and Sabine Buchholz. (2000). “Introduction to CoNLL-200 Shared Task: Chunking”. *Proceedings of CoNLL-2000 and LLL-2000*. Lisbon, Portugal. 127-132.

[Th00] “汉语句子的语块标注规范”，清华大学计算机系智能技术与系统国家重点实验室，技术资料，2000 年 6 月。

[Vee99] Jorn Veenstra (1999). “Memory-based text chunking”, In *Nikos Fakotakis (ed.) Machine Learing in human language technology.* workshop at ACAL 99.

[XP00] Xia, Fei, Martha Palmer, & al. (2000) “Developing Guidelines and Ensuring Consistency for Chinese Text Annotation”. In *Proceedings of the second International Conference on Language Resources and Evaluation (LREC-2000), Athens, Greece.*

[ZS99] Qiang Zhou, Maosong Sun. (1999). “Build a Chinese Treebank as the test suite for Chinese parser”, *Proceedings of the workshop MAL’99(Multi-lingual Information Processing and Asian Language Processing),* Beijing, China. p32-36.

[ZSH00] 周强，孙茂松，黄昌宁 (2000). “汉语最长名词短语的自动识别”， 《软件学报》11(2), 195-201.