# 中文 CCG 树库的构建*

宋 彦[1], 黄昌宁[2], 揭春雨[1]

[1]香港城市大学 中文、翻译及语言学系，香港九龙 达之路 83 号

[2]微软亚洲研究院，北京 100080

E-mail: yansong@ctiyu.edu.hk; v-cnh@microsoft.com; ctckit@cityu.edu.hk

**摘 要**：组合范畴语法（CCG）是一种类型驱动的语法，可以高度词例化（lexicalized）并兼顾句法和一定程度上语义的表达，可为深层次的文本分析提供有效支持。将 CCG 应用于真实文本分析需要编制大规模的词库，为了避免为此付出的昂贵人力和资源，一个行之有效的解决方案是利用现有短语句法树库来自动生成 CCG 树库。本文提出在清华中文树库的基础上自动生成 CCG 树库的方案，在预定义的中文句型和基于清华树库的动词子范畴框架的支持下，通过标准转换算法，得到一个包含 32737 句、超过 35 万词次的中文 CCG 树库。该树库通过手工和自动评价验证，又与已有文献报道的多语种 CCG 树库构建工作比较，均证明本文建议算法的有效性。

**关键词**：组合范畴语法；树库；中文句型；动词子范畴框架

# Construction of Chinese CCGbank

Song Yan[1], Huang Chang-ning[2], Chunyu Kit[1]

[1]Department of Chinese, Translation & Linguistics, City University of Hong Kong, Hong Kong

[2]Microsoft Research Asia, Beijing 100080

E-mail: yansong@ctiyu.edu.hk; v-cnh@microsoft.com; ctckit@cityu.edu.hk

**Abstract:** Combinatory Categorial Grammar (CCG) is a type-driven lexicalized grammar formalism with a transparent interface between syntax and semantics, which is essential to in-depth text processing. To apply CCG to real texts, however, a large scale lexicon is required to provide data support. An effective way to achieve this without a huge demand for manpower and resources is to transform an existing treebank into a CCGbank. In this paper, we present an approach of this kind for constructing a Chinese CCGbank from Tsinghua Chinese Treebank, supported by a number of verb sub-categorization and predefined Chinese sentence patterns. The resulted CCGbank includes 32737 sentences with more than 350,000 word tokens. The effectiveness of this approach is confirmed by an evaluation using manually annotated references, in comparison with existing work on several reported CCGbanks.

**Keywords:** Combinatory Categorical Grammar, treebank, Chinese sentence pattern, verb sub-categorization frame

## 1 引言

自动句法分析（parsing）是自然语言处理的一项基本技术，是迈向语义理解的一道门槛。当前信息处理对于深度本文分析的需求有增无减，不但要求描绘句内各成分间的句法关系，还需要在一定程度上刻画词和词之间的语义联系。组合范畴语法（Combinatory Categorial Grammar，CCG）[8]可以为这种需求提供一种显式的表达形式，直接将词汇关系映射到各个句法节点的范畴上，为快速句法分析提供了一种有效的形式化描述。在 2009 年约翰霍普金斯大学举行的夏季研讨班（JHU Summer School 2009）上[1]，研究人员通过采用优化的句法分析算法，使 CCG 句法分析在维基百科（Wikipedia）语料上达到每秒超过 100 句的分析速度，且抽样显示，其分析精度并未有明显损失，说明 CCG 可以用来进行工业规模的句法分析[2]。

实用的句法分析器（parser），特别是以有监督的机器学习方法训练，需要大量的句法树实例

---

* 本文主要工作由第一作者在微软亚洲研究院实习期间完成。

[1] 该研讨班的主题为“Parsing the Web”，旨在通过 CCG 得到快速而有效的跨领域深层句法分析结果。

作训练语料。目前很难找到合适的CCG树库作此用途，对于中文，此等资源[1]尤其短缺。要将CCG应用于中文信息处理，当务之急是构建一个中文CCG树库（CCGbank）。众所周知，从零开始标注一个大规模的句法树库是一项极费人力和资源的工程。考虑到目前已经有一些现成的句法树库，例如宾州英文树库（PTB）、宾州中文树库（CTB）、德文TIGER树库、台湾中研院Sinica中文树库和清华大学中文树库（TCT）等，把这些资源转化成所需的CCG树库，当是一种行之有效的资源建设方案。

本文在清华中文树库的基础上，针对中文句法和句型的特点，通过自动转换方式实现了中文CCG树库的构建。由于清华树库没有空语类和同指索引等标记，我们提出一种识别动词-论元关系的简单方法，识别出句子中每个动词的词汇范畴及其相关的论元，并将识别结果直接标注到相应的CCG句法树上。本文给出所构建的CCG树库的统计数据，并使用人工标注的200句测试集验证了树库构建方法的有效性。

## 2 CCG简介及相关工作

CCG源自范畴语法[1]，是一种类型驱动（type-driven）的词例化语法，它通过词汇范畴显式地提供从句法到语义的接口（interface）[8]。范畴（category）是CCG的基本操作单元，它有两种形式：原子范畴（atomic category）和组合范畴（functor category）。前者是CCG的基本符号集，用来表达基本的词汇类型和句法功能；后者由前者构成，通常的形式为X/Y或X\Y，由“\”（左斜杠）和“/”（右斜杠）分别指示左右两个不同的组合方向，表示该范畴可以向左或右找到变元（argument）Y，并得到组合结果（result）X。在这些词汇范畴的基础上，CCG使用一系列操作规则来完成推导（derivation），如表2.1所示。

表2.1 CCG规则

| 规则类型 | 规则说明 | 形式化描述[2,3] |
|---|---|---|
| 基本规则 | 前向应用 | $X/Y:f+Y:a \Rightarrow_{>} X:f(a)$ |
| | 后向应用 | $Y:a+X\backslash Y:f \Rightarrow_{<} X:f(a)$ |
| 组合规则 | 前向组合 | $X/Y:f+Y/Z:g \Rightarrow_{>B} X/Z:\lambda a.f(g(a))$ |
| | 后向组合 | $Y\backslash Z:g+X\backslash Y:f \Rightarrow_{<B} X\backslash Z:\lambda a.f(g(a))$ |
| | 前向交叉组合 | $X/Y:f+Y\backslash Z:g \Rightarrow_{>Bx} X\backslash Z:\lambda a.f(g(a))$ |
| | 后向交叉组合 | $Y/Z:g+X\backslash Y:f \Rightarrow_{<Bx} X/Z:\lambda a.f(g(a))$ |
| | 前向类型提升 | $X:a \Rightarrow_{>T} T/(T\backslash X):\lambda f.f(a)$ |
| | 后向类型提升 | $X:a \Rightarrow_{<T} T\backslash(T/X):\lambda f.f(a)$ |
| 非组合规则 | 类型转换 | $X \Rightarrow_{T} TOP$ |

CCG树库的自动转换已有一些工作。Hockenmaier首先在PTB上自动转换生成英文CCG树库[3,5]。该工作使用基本的三步转换流程，同时也针对短语树库设计了很多转换规则，为后来的相关工作提供算法和转换规则的标准。Hockenmaier在TIGER树库上的工作[4]，则针对德文的特点，提出了一些非常规的处理方案，包括词序的链接关系、抽取（extraction）类型、并列结构

[1] 2010年Daniel Tse[9]在宾州中文树库上开发了相应的中文CCG树库，然而该树库尚未开放给他人使用。

[2] 本文中，我们使用=>和↔分别表示子节点到父节点和父节点到子节点的方向。另外，在组合逻辑的表达式中，箭头的下标表示CCG组合的不同操作类型。形式化描述中，f(a)表示函数f以a为变元，定义了整个组合过程的逻辑语义表达。

[3] 其中TOP表示所有可能的转换类型。

(coordination)等特殊情况，第一次将英语之外的句法资源转换为 CCG 树库。Tse 使用 Hockenmaier 的算法，从 CTB 转换出中文 CCG 树库[9]，在预处理时把一些难以转换的中文句型作为特殊结构进行处理，例如，主题句（topicalization）、主语省略（pro-drop）、非动词谓语结构（zero-copula）等。此外，Tse 还采用后处理方式来完善输出的 CCG 句法树，包括整理主、宾语抽取结构（subject/object extraction）和更正修饰词的范畴等。

# 3 构建中文 CCG 树库

我们的工作采用清华中文树库（TCT）作为原始资源，是因为考虑到：1）TCT 的规模和覆盖率相对较大；2）TCT 提供了每个短语的中心语的显式标记[1][11]。下面详细陈述我们使用的转换算法及针对中文句型的特殊处理。

## 3.1 树库转换算法

### 3.1.1 确定句法成分类型

这是树库转换的第一步，目标是将一棵子树（treelet）中的中心语（head）、补足语（complement）和附加语（adjunct）区分出来，便于后续的二叉化和范畴指派工作。TCT 句法树中每个短语的中心语已经被显式标示，我们需要小心加以区分的主要是补足语和附加语。其中，补足语一般都是可以充当动词核心论元（core argument）的成分，如 NP、SP、S 等，通常用原子范畴表示，并进入与之相应的述语动词的组合范畴。一般来说，除中心语和补足语以外其他语法成分都属于附加语。

### 3.1.2 句法树二叉化

二叉化（binarization）是将多叉树（平坦结构）重组成二叉结构，从而把整棵句法树统一为标准二叉树，以便按照句法成分类型来描述每棵子树中两个子节点之间的关系。二叉化遵循以下原则：1）中心语左边的节点，一律往右分叉（right-branching）；2）中心语右边的节点，一律往左分叉（left-branching）。这样，二叉化以后，整个结构的节点间关系并不改变，中心语左边的成分仍在其左，右边的仍在其右。图 3.1 给出了一个平坦结构二叉化的示例，其中 h 表示 P 的中心语。

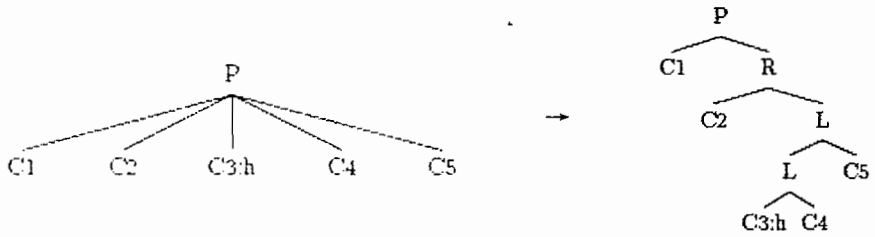


图 3.1 二叉化示例

### 3.1.3 范畴指派

句法树二叉化以后，就可以自上而下对其节点指派范畴。对中心语与补足语、中心语与附加语两种不同的语法关系，指派范畴时应分别遵循以下规则（箭头左边表示一棵子树的根节点，右边是它的两个子结点）：

$R \rightarrow R/C + C$ （其中 R/C 是中心语，C 是补足语），与

$R \rightarrow R + R\backslash R$ （其中 R 是中心语，R\R 是附加语）。

[1] 本文使用的 TCT 语料来自 2009 年中文信息学会举办的 ParsEval 中文句法分析评测活动，它是 TCT 的一个子集，其中每个短语标记用整数后缀（0，1，2，…）标明了中心语的位置，但 TCT 原有的短语结构标记（如 ZW, SB 等）已被删除。

一般来说，补足语用其自身的原子范畴来描述，例如 NP，S 等；而附加语则用它所修饰的中心语的范畴来描述，以确保其父节点可以得到与其中心语一致的范畴。除这两条规则之外，我们还针对并列结构中的连接词、标点符号等使用一种非组合规则：$R \rightarrow R + p$，其中心语直接继承根节点的范畴。

## 3.2 动词及其论元的抽取

对述语-论元关系（predicate-argument relationship）的描写是 CCG 区别于传统上下文无关文法的一个显著特性。与 PTB 和 CTB 不同，TCT 树库并未提供句法移位的完整标注，应用转换算法难以单独完成合理的述语-论元关系指派。为此，我们设计了一个轻量级的动词子范畴框架（verb sub-categorization frame）抽取算法。为了保证其可靠性，该算法仅对动词进行处理，其基本运作基于以下三个假设：1）大量句子的表述方式遵从正常的 SVO（主语-述语-宾语）架构；2）非正常移位（因而形成长距离依存关系）的论元主要是受事（patient）论元，它们在正常结构中应该直接跟随在相应的及物动词之后；3）如果在整个语料库中某个动词没有携带宾语的证据，就把该动词归为不及物动词。我们通过统计 TCT 中动词后面直接携带的句法成分，判定一个动词是否及物，以及它可以携带何种类型的补足语[1]的信息。我们认为短语 np，vp，mp，sp，tp 和句子 S 可以充当动词的补足语成分，其他成分则标注为附加语。最终，我们从 TCT 中得到一个动词子范畴框架的数据库，其中包含所有被标注出来的及物动词及其携带的补足语，包括该补足语出现的频度。后面的评测结果表明，通过动词子范畴框架数据库来判断一个动词是否存在补足语，以及该补足语是否以长距离依存关系出现，可以有效弥补标准转换算法的不足。

## 3.3 中文特殊句型的 CCG 推导

针对特定树库资源的 CCG 转换过程与语言类型和语料标注格式紧密相关，前述的标准转换算法难于覆盖 TCT 的全部句法树。同时，考虑到中文表达的灵活性，很多特殊的句法结构较难得到正确的分析结果[6]。我们列出树库中 10 种典型的句法结构，包括“的”字结构、并列结构、非动词谓语句（如名词谓语句、形容词谓语句和主谓谓语句）、谓词性宾语、兼语句、连动句、被动结构、无主（语）句、存在句以及独立成分等，并针对每种句型设计了专门的转换方案，细节请参见[7]。考虑到篇幅限制，我们在此仅例示“的”字结构和被动结构的处理方案。

### 3.3.1 “的”字结构

“的”字结构一直以来都是中文处理的一个难题，不但由于它具有灵活的表示形式，更由于其内部复杂的依存关系。总的来说，绝大多数“的”字结构都是名词短语，在 TCT 树库中标注为一个平坦的句法结构，符合*[X 的]*或者*[X 的 Y]*这样的形式。针对前者，我们将助词“的”视为整个结构的中心语，$X$ 是它的补足语；对于后者，我们将*[X 的]*当作一个定语语块，而 $Y$ 是定语语块所修饰的中心语（即名词短语*[X 的 Y]*的中心语）；助词 “的”仍是定语语块的中心语，如果定语语块中的 $X$ 是一个动词短语（VP）或小句（S），就需要进一步判断以下两种情况：

1）如果 $Y$ 是定语从句 $X$ 中的述语动词 $v$ 的一个论元（如宾语或主语），就要给动词 $v$ 和 $Y$ 建立相应的述语-论元关系，并指派合适的范畴，如图 3.3 所示。为了实现 CCG 的句法树推导，这里需要对 $X$ 中的主语 *NP* 进行类型提升，以便使它与动词先组合，然后整个 $X$ 部分表示成一个需要寻找宾语的范畴。

2）否则（即 $Y$ 不是 $X$ 中述语动词的一个论元），定语语快整体作为 $Y$ 的附加语，如图 3.2 所示。

[1] 虽然我们使用动词及其论元的关系来描述动词子范畴化，实际上我们的算法仅支持寻找可能的补足语，而并不区分其是否为真正的受事宾语。

这里，我们仍根据前述的动词子范畴框架来判断动词类型及其相应的补足语。

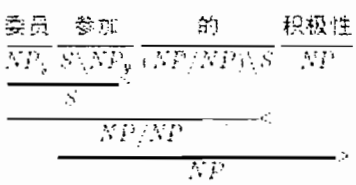


图 3.2 “的”字结构的 CCG 分析结果（一）

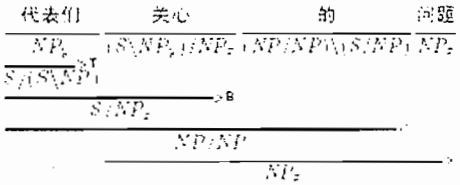


图 3.3 “的”字结构的 CCG 分析结果（二）

### 3.3.2 被动结构

中文里的被动结构通常都由明显的标记（marker）如“把”、“被”等介词引出。在 TCT 中，这些介词通常包含在一个介词短语中。文献[10]用 HPSG 处理这种结构时，描述了三种操作方法，其一是将这些标记作为动词处理，其二是作为介词，第三种方案则是将它们定义成一种格标记（case marker）。考虑到 TCT 的标注规范，本文采用介词形式的处理方案更合理。在 TCT 中，几乎所有的被动结构都放在固定的 pp+vp 组合中，因此使用简单的模式匹配方式即可激活针对被动结构的处理规则。典型情况下，述语动词的施事（agent）经常被置于被动标记之后，例如：*[dj-1 [np-2 班里/n 的/uJDE 战友/n ] [vp-1 [pp-0 被/p [np-1 搏斗/vN 声/n ] ] 惊醒/v ] ]*，我们有如图 3.4 的 CCG 表示。

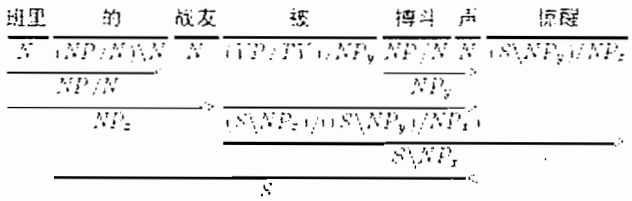


图 3.4 被动结构的 CCG 分析结果

## 4 实验结果及评价

### 4.1 CCG 树库的统计结果及验证

我们用以上方法完成了所用 TCT 树库中 32737 个句子的 CCG 转换，句子覆盖率达到 99.9%，余下的 33 句因为特殊标记或者非正常的句型结构而超出了我们算法的处理能力范围。与之相比，英文 CCG 树库[5]的句子覆盖率为 99.44%，未转换的句子数为 274，德文 CCG 树库[4]在 50474 句的基础上得到的句子覆盖率仅为 92.4%，从 CTB 转换的 CCG 树库[9]在 28295 句的基础上达到了 99.76%的句子覆盖率。我们相对较高句子覆盖率的原因可以归结为：其一，TCT 已经标示好的中心语对 CCG 的转换有非常大的帮助；其二，大部分 TCT 树结构都已经是二叉结构，因此在进行结构转换的时候，降低了出现错误结构的风险；其三，针对特殊中文句型的操作涵盖了绝大多数难以被标准算法正常处理的句子，而这部分对于提升整体句子覆盖率有很大帮助。

具体地，在所得到的 CCG 树库中，一共有 10 个原子范畴，包括 M（量词）、MP（数量短语）、NP（名词及名词短语）、SP（方位词及方位短语）、TP（时间短语）、PP（介词短语）、S（句子）、conj（连接词或连接标记）、p（标点符号）以及 dlc（独立语标记），其中 dlc 是直接来自 TCT 的标记。在此基础上，我们一共得到了 763 个不同的范畴类型，其中 208 个出现的频度超过 10 次，279 个仅仅出现一次。图 4.1 显示了范畴类型的个数随语料库规模增长的变化。我们也统计了它们在树库中的组合规则，有近 1600 条，其中超过 400 条出现了 10 次以上。这些统计符合一般语法的观点，即有限的重要范畴及其组合规则在集中大量地使用。

在词汇方面，一共在 23641 个词（word type）上得到了 41733 个词汇范畴，其中一部分词有为数众多的不同范畴，承担不同的句法功能，而一些词仅有一个范畴，在整个语料中扮演固定的句法成分。作为案例，表 4.1 列出词“学”的所有词汇范畴。

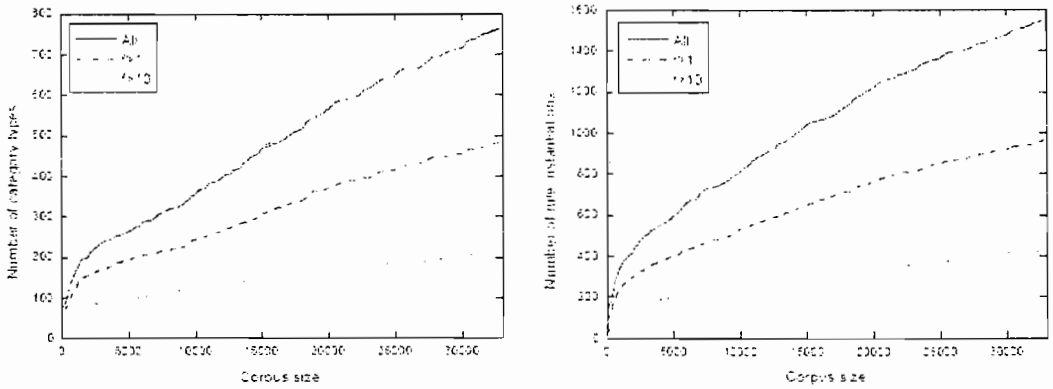

图 4.1　范畴类型（左）和规则的数目（右）与语料库规模的增长关系

表 4.1　词汇范畴示例

| 词语 | TCT 词性标注 | 范畴 | 范畴频次 | log*P(category\|word=学)* |
|---|---|---|---|---|
| 学 | n | NP | 4 | -2.3978952727983707 |
| 学 | v | S\NP | 8 | -1.7047480922384253 |
| 学 | v | [S\NP]/NP | 23 | -0.6486954179891115 |
| 学 | v | [S\NP]/[S\NP] | 3 | -2.6855773452501515 |
| 学 | v | [S\[S\NP]]/NP | 4 | -2.3978952727983707 |
| 学 | v | [[S\NP]/NP]/[S\NP] | 1 | -3.784189633918261 |
| 学 | v | [[S\NP]\PP]/NP | 1 | -3.784189633918261 |

## 4.2　基于人工标注的评价

前述的统计分析给出对整个语料库的宏观描述，但我们无法从中得知句库质量。与文献[4,5,9]中的工作不同，我们还在整个句库中抽取出一部分句子进行了人工 CCG 标注，然后作为标准与自动转换的结果进行比较，对自动转换结果评估。该工作相当于直接比较两个句子的句法分析结果，我们使用已被广泛采用的自动句法分析评测方法，即通过计算两个句子中正确的句法成分及其覆盖的范围是否一致，从而得到定量的评价结果，以惯常的准确率（Precision）、召回率（Recall）以及基于这两个指标的 F 值来体现。为了保证选取的评测数据具有代表性，同时又不至于动用过多的资源进行标注，我们需要平衡数据选取和人工标注的负担。因此我们对整个语料库按照长度进行了统计，发现 20 词以下的句子占总数的 90%以上，我们从中选取总共 200 句，按长度分两类，各 100 句，作为测试语料。测试结果如表 4.2 所示。

表 4.2　基于人工标注测试语料的评价结果

| 类别 | 方法 | 准确率 | 召回率 | F 值 | 词汇范畴正确率 |
|---|---|---|---|---|---|
| 第一类(1~10 词) | 标准转换 | 0.9427 | 0.9458 | 0.9443 | 0.9201 |
| | +动词子范畴 | 0.9607 | 0.9639 | 0.9623 | 0.9418 |
| 第二类(11~20 词) | 标准转换 | 0.9566 | 0.9479 | 0.9523 | 0.9573 |
| | +动词子范畴 | 0.9882 | 0.9818 | 0.9850 | 0.9855 |

其中，“+动词子范畴”表示在标准转换的基础上使用动词子范畴框架来帮助识别动词及其论元。从中可以看出，使用了动词子范畴框架以后，转换的质量得到了明显提高，在两类测试集中各个指标上分别提升了约 2%和 3%，特别在词汇范畴的准确度上也有更明显的提高。这 200 句的分析结果一定程度上反映出，整体的转换质量较为可靠。后经人工检查，发现拥有完全正确的 CCG 句法树的句子超过了 70%，也证实该转换算法的有效性以及得到的 CCG 树库的可靠性。

## 5 结论

本文提出了一种将中文 TCT 短语树库转成 CCG 树库的方法，通过宏观统计并和人工标注对比，验证了该方法的有效性。相比已有工作，我们有如下创新：第一，使用一个轻量级的动词子范畴框架抽取结果，有效提升了在句子中识别动词及其论元关系的正确率；第二，提出了针对不同中文句型的特殊处理方案，为其中很多典型结构设计了有效的转换规则；第三，直接使用了人工标注结果来评价树库质量，尽管我们使用的测试集规模较小（200 句），相信评测结果能适当反映转换结果的整体质量。除此之外，由于 CCG 树库的转换方法与特定语种和树库资源紧密相关，因此在不同的树库上转换的难度也各不相同。我们的工作也为那些在没有空语类标注或者动词论元关系描述的树库上实现 CCG 转换提供了一个有价值的参考。

## 参 考 文 献


[1] Bar-Hillel, Y.: A quasi-arithmetical notation for syntactic description. Language, 29(1), 47-58 (1953).

[2] Clark, S., Copestake, A., Curran, J.R., Zhang, Y., Herbelot, A., Haggerty, J., Ahn, B.G, Wyk, C.V., Roesner, J., Kummerfeld, J., Dawborn, T.: Large-scale syntactic processing: Parsing the web. Final Report of the 2009 JHU CLSP Workshop (Oct 2009).

[3] Hockenmaier, J.: Data and Models for Statistical Parsing with Combinatory Categorial Grammar. Ph.D. thesis, University of Edinburgh (2003).

[4] Hockenmaier, J.: Creating a CCGbank and a wide-coverage CCG lexicon for German. In Proceedings of ACL-2006. pp.505–512. Sydney, Australia (July 2006).

[5] Hockenmaier, J., Steedman, M.: CCGbank: A corpus of CCG derivations and dependency structures extracted from the penn treebank. Computational Linguistics 33(3), 355-396 (2007)..

[6] Liu, Y., Pan, W., Gu, W.: Modern Chinese Grammar (2002).

[7] Huang, C.N., Song, Y.: Chinese CCGbank Construction from Tsinghua Chinese Treebank. In Proceedings of the Roundtable Conference on Linguistic Corpus and Corpus Linguistics in the Chinese Context, Hong Kong, (2011) (forthcoming).

[8] Steedman, M.: The Syntactic Process (2000).

[9] Tse, D., Curran, J.R.: Chinese ccgbank: extracting ccg derivations from the PENN Chinese Treebank. In Proceedings of COLING-2010. pp. 1083–1091. Beijing, China (August 2010).

[10] Yu, K., Yusuke, M., Wang, X., Matsuzaki, T., Tsujii, J.: Semi-automatically developing Chinese HPSG grammar from the PENN Chinese Treebank for deep parsing. In Proceedings of COLING 2010: Posters. pp. 1417–1425. Beijing, China (August 2010).

[11] Zhou, Q.: Annotation scheme for Chinese Treebank. Journal of Chinese Information Processing 18(4), (2004).