Encyclopedia of Chinese Language and Linguistics

Volume 3

Men–Ser

ENCYCLOPEDIA OF CHINESE LANGUAGE AND LINGUISTICS

Volume 3
Men–Ser

General Editor
Rint Sybesma

Associate Editors
Wolfgang Behr
Yueguo Gu
Zev Handel
C.-T. James Huang
James Myers

BRILL

LEIDEN • BOSTON
2017

Typeface for the Latin, Greek, and Cyrillic scripts: “Brill”. See and download: brill.com/brill-typeface.

ISBN 978-90-04-18643-9 (hardback, set)
ISBN 978-90-04-26227-0 (hardback, vol. 1)
ISBN 978-90-04-26223-2 (hardback, vol. 2)
ISBN 978-90-04-26224-9 (hardback, vol. 3)
ISBN 978-90-04-26225-6 (hardback, vol. 4)
ISBN 978-90-04-26226-3 (hardback, vol. 5)

This book is printed on acid-free paper and produced in a sustainable manner.

- Parts of speech (syntactic categories): noun (*míngcí* 名詞); pronoun (*dàicí* 代詞); verb (*dòngcí* 動詞); adjective (*xíngróngcí* 形容詞); numeral (*shùcí* 數詞); measure (*liàngcí* 量詞); preposition (*jiècí* 介詞); adverb (*fùcí* 副詞); conjunction (*liáncí* 連詞); particle (*zhùcí* 助詞); and interjection (*tàncí* 嘆詞). These categories are taken more or less verbatim from L1 Chinese grammar without much discussion. They are so strictly defined syntactically that arbitrary category assignments result, e.g., defining the disposal marker *bǎ* as a preposition, since it appears to occupy the same syntactic position as rightful prepositions;
- Grammatical functions: subject (*zhǔyǔ* 主語); object (*bīnyǔ* 賓語); attribute (*dìngyǔ* 定語); adverbial (*zhuàngyǔ* 狀語); complement (*bǔyǔ* 補語); and predicate (*wèiyǔ* 謂語). Though these are functions, they are in most cases defined syntactically, giving rise to confusing labels such as "agentive object" *shīshì bīnyǔ* 施事賓語 'agentive object', e.g., Liú *et al.* (1996);
- Sentence types: declarative sentence (*chénshùjù* 陳述句); interrogative sentence (*yíwènjù* 疑問句); imperative sentence (*qíshǐjù* 祈使句); and interjection sentence (*gǎntànjù* 感嘆句);
- Sentence structures: Simplex sentence (*dānjù* 單句); complex/compound sentence (*fùjù* 複句); and discourse (*piānzhāng* 篇章); and
- Specific structures: disposal sentence (*bǎzìjù* 把字句); passive sentence (*bèidòngjù* 被動句); command sentence (*jiānyǔjù* 兼語句); serial sentence (*liándòngjù* 連動句); cleft sentence (*fēnlièjù* 分裂句); equation sentence (*pànduànjù* 判斷句); topic sentence (*huàtíjù* 話題句); focus sentence (*jiāodiǎnjù* 焦點句); etc. The listing in this category varies a great deal from author to author.

4. Concluding Remarks

The field of L2 Chinese has had a long history, since the 1950s, and its pedagogical grammar component has followed its L1 counterpart closely, in fact too closely. The distinguishing features between L1 and L2 grammars have begun to emerge in the field and more discussion of principles underlying the construction of a pedagogical grammar, such as presented in Teng (2010), will prove to be helpful, before a healthier L2 Chinese pedagogical grammar becomes available to better serve the field.

Bibliography

Liú Yuèhuá 劉月華, Pān Wényú 潘文娛 and Gù Wěi 故韡, *Shíyòng xiàndài Hànyǔ yǔfǎ* 實用現代漢語語法 [Practical Modern Chinese grammar], traditional characters edition, Taipei 台北: Shīdà Shūyuàn 師大書苑, 1996.

Lù Qìnghé 陆庆和, *Shíyòng duìwài Hànyǔ jiāoxué yǔfǎ* 实用对外汉语教学语法 [Practical L2 Chinese pedagogical grammar], Běijīng 北京: Běijīng Dàxué 北京大学出版社, 2006.

Lǚ Wénhuá 吕文华, *Duìwài Hànyǔ jiāoxué yǔfǎ tànsuǒ* 对外汉语教学语法探索 [Inquiries into L2 Chinese pedagogical grammar], Běijīng 北京: Yǔwén 语文出版社, 1994.

Ross, Claudia and Jing-heng Sheng Ma, *Modern Mandarin Chinese Grammar: A Practical Guide*, London: Routledge, 2006.

Teng Shou-hsin 邓守信, *Duìwài Hànyǔ jiāoxué yǔfǎ* 对外汉语教学语法 [L2 Chinese pedagogical grammar], Běijīng 北京: Běijīng yǔyán dàxué 北京语言大学出版社, 2010.

Xing, Janet, *Teaching and Learning Chinese as a Second Language: A Pedagogical Grammar*, Hong Kong: Hong Kong University Press, 2006.

Shou-hsin Teng

Peking University Treebank

Ever since the 1990s, as statistical methods became the main stream in the field of natural language processing, increasing attention has been paid to deep tagging. The institutions that started to build the Chinese Treebank around 2000 include the University of Pennsylvania (USA), Academia Sinica (Táiwān), and Tsinghua University and Peking University (mainland China) (Xue and Xia 2000, Xue *et al.* 2005; Huang *et al.* 2000; Zhān 2000; Zhōu 2004; for more information see the appendix). This article focuses on the characteristics of the Peking University (PKU) Chinese Treebank.

1. Text Sampling

One of the major designing targets of the PKU Chinese Treebank is to support research into the basic syntactic structures of modern Chinese. As a consequence, primary attention is paid to the selection of materials from standard modern Chinese texts. The example sentences from elementary and middle school textbooks and research works on Chinese syntax are taken to constitute ideal material for this purpose. Besides this, the Treebank also includes language material from news reports and government publications. The average number of words from the latter two types is significantly higher than from the former two. The table below shows detailed information regarding the language samples of the PKU Chinese Treebank when its first edition was released in 2006.

2. Tagging System

The PKU Treebank adopts the so-called light tagging model. In terms of syntactic structure, the principle of least assumption is taken as a fundamental tagging policy to describe the basic sentence structures. This means that this is done in traditional structuralist terms (Zhū 1982) and not in terms of the multi-layered ideas of generative grammar (deep vs. surface structure), or the dependency grammar model based on the description of the word relations. Following the principle of an extensible hierarchy, all the syntactic tags are divided into three levels: (1) sentence, (2) phrases, and (3) words. The functional categorization of phrases and words is highlighted; 14 categories have been established for phrases (in total 21 if sub-categories are included), and 26 for words (in total 98 with sub-categories included). In this way, the users of the Treebank can conveniently choose different packages of information. As for the head element of syntactic structure, the Treebank marks it with an exclamation marker: !. The Treebank also includes information about the texts such as the titles, the inserted elements, etc.

3. Mode of Tagging

The PKU Treebank integrates automated and manual tagging. The specific process is indicated in the following illustration.

In the illustration in Figure 1, sentence-splitting (step 1) is automatically processed by computer with little manual correction. Word tokenization and part of speech tagging (step 2) is automatically processed by programs of Chinese word segmentation and POS tagging with little manual correction. The draft version of the Treebank (step 3) is achieved automatically by a Chinese syntactic parser. A visualized TreeEditors is then used to assist manual proofreading of the draft version sentence by sentence (step 4). This process enables the syntactic analysis of each sentence to fit the syntactic tagging requirements of the Treebank. The "tree syntactical knowledge

Genres	Number of sentences	Number of words	Number of Chinese characters
Chinese language textbooks for primary school to high school	31,928 (62.40%)	536,927 (66.75%)	744,563 (64.21%)
Example sentences from grammar books and papers	14,084 (27.53%)	128,499 (15.97%)	174,123 (15.02%)
Newspaper	3,550 (6.94%)	93,796 (11.66%)	165,880 (14.31%)
White paper published by the Chinese government	1,601 (3.13%)	45,165 (5.61%)	74,949 (6.46%)
Total	51,163 (100%)	804,387 (100%)	1,159,515 (100%)

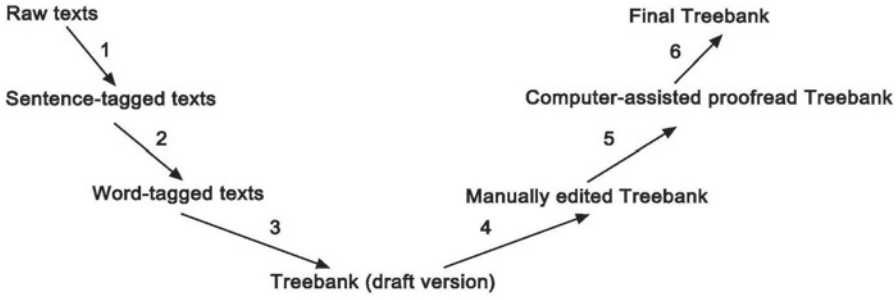


Figure 1. Flowchart of compiling the PKU Treebank.

extraction tool" and the "tree grammaticality judgement checking tool" in the TreeEditor are utilized to check consistency of tagging (step 5). This helps to detect inconsistencies of tagging that may occur within phrases, which therefore can be further distinguished and corrected manually. Finally, the "tree checking tool" is applied to perform an overall consistency check, making sure that no mistakes were made against tagging rules after which the final Treebank texts are generated (step 6).

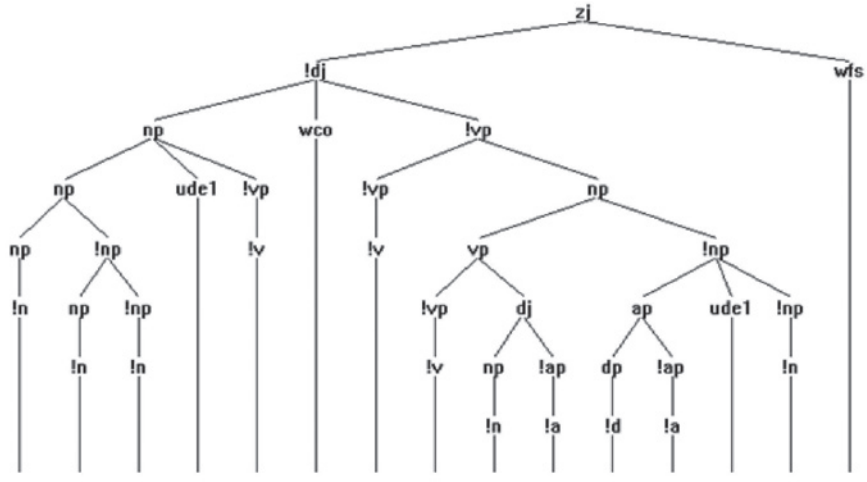
Among the procedures, the fourth and fifth steps are vital to ensure the quality of the Treebank by combining the automatic programs and manual annotation. The fourth step adopts the conventional post-correction model that follows the natural order of the texts while the fifth uses the transverse correction model across the texts. The former correction model pays attention to the partials of each sentence, analyzing the syntactic structure of sentences whereas the latter focuses on the whole syntactic system, inspecting whether the Treebank is consistent in dealing with the same grammatical phenomena.

Figure 2 is a sample of the tagging format of one sentence from the final Treebank texts and the image of its tree as they appear in the TreeEditor.

4. Knowledge Extraction and Improvement of the Parser

After the Treebank reaches a certain scale, information about phrase structure and its frequency can be extracted to enrich the knowledge base of the syntactic parser which is rule-based, and to increase the accuracy of analysis (i.e., to improve the results of step 3). The process of building treebanks thus becomes an ever-improving cycle. In addition, examining the rules that have low frequency among the phrase structure rules open up two options. First, they may turn out to be tagging mistakes that need to be corrected. Second, they may exemplify special cases of within the syntactic system of Chinese. They may, for instance, not be regular phrases, but set constructions, which could be of use for those who are involved in basic theoretical research on Chinese syntax. For example, the phrase *bèi tā chǎo de* 被他吵的 /by him make.noise SUB/ 'be annoyed by him' is a "rare" tree (the type of verb does not generally occur in a passive sentence like this one), and thus a set construction. Systematically examining the features of phrase structural rules can also help understand the ambiguous structures of Chinese syntax. For instance, the fact that the same phrase order or the same word order forms different trees inevitably indicates that this kind of order contains potentially ambiguous syntactic structures. Through statistical analysis of these orders and possible tree structures that these orders can be associated with, one can understand the level of ambiguity of potentially ambiguous structures in Chinese, which in turn can lead to information for generating pointed strategies for the elimination of ambiguity in Chinese natural language processing.

(zj (!dj (np (np (np (!n (妇女)) !np (np (!n (经济)) !np (!n (地位)))) ude1 (的) !vp (!v (提高))) wco (，) !vp (!vp (!v (是)) np (vp (!vp (!v (实现)) dj (np (!n (男女)) !ap (!a (平等))))) !np (ap (dp (!d (最)) !ap (!a (重要))) ude1 (的) !np (!n (基础)))))) wfs (。)))



妇女 经济 地位 的 提高 ， 是 实现 男女 平等 最 重要 的 基础 。

Fùnǚ jīngjì dìwèi de tígāo , shì shíxiàn nán-nǚ píngděng zuì zhòngyào de jīchū .

women economy status SUB increase, be achieve man.woman equality most important SUB basis

'The increase of the economic status of women is the most important basis to achieve the equality between men and women.'

(Source: Chinese government report 1994 *The current situation of Chinese women*)

Figure 2. Sample bracketed sentence—full syntactic structure parsed and corrected manually.

Appendix

Table 1. The basic information of four representative Chinese treebanks

Chinese treebanks	**Scale**	**Types of language materials**	**Frame of grammatical theories**	**phrase structure tags**
Sinica Treebank of Táiwān (3.0 version) Oct. 2005	360,000 words	Literature, life, social, science, philosophy, art	Case Grammar theory	6 tags for phrase function; annotated syntactic head and semantic head; 54 tags for semantic characters.

Table 1. (*cont.*)

Chinese treebanks	Scale	Types of language materials	Frame of grammatical theories	phrase structure tags
PKU Treebank (1.0 version) Jan.2006	800,000 words	Chinese textbooks, exemplary sentences of Chinese patterns, news reports and government publications	Conventional structural grammar theory (PSG)	21 tags for phrase function; annotated syntactic head.
Tsinghua University Treebank (1.0 version) Oct. 2004	1,000,000 words	Literature, news, academic works, applied	Conventional structural grammar theory (PSG)	16 tags for phrase function; 27 tags for structural relation.
Penn Chinese Treebank	1,200,000 words	News, radio news, radio talks and blogs	Generative Grammar (GB theory)	23 tags for phrase function; 33 tags for phrase function character.

Bibliography

Abeillé, Anne, ed., *Treebanks: Building and Using Parsed Corpora*, Dordrecht: Kluwer Academic Publishers, 2003.

Huang, Chu-Ren, Feng-Yi Chen, Keh-Jiann Chen, Zhao-ming Gao and Kuang-Yu Chen, "Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface", in: Martha Palmer, Mitch Marcus, Aravind Joshi and Fei Xia, eds., *Proceedings of the Second Chinese Language Processing Workshop*, Hong Kong, 2000, 29–37.

Marcus, Mitchell P., Beatrice Santorini and Mary Ann Marcinkiewicz, "Building a Large Annotated Corpus of English: the Penn Treebank", *Computational Linguistics* 9/2, 1993, 313–330.

Xue, Nianwen and Fei Xia, "The Bracketing Guidelines for the Penn Chinese Treebank (3.0)", http://www.cis.upenn.edu/~chinese/parseguide.3rd.ch.pdf, 2000.

Xue, Nianwen, Fei Xia, Fu-Dong Chiou and Martha Palmer, "The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus", *Natural Language Engineering* 11/2, 2005, 207–238.

Zhān Wèidōng 詹卫东, *Miànxiàng zhōngwén xìnxī chǔlǐ de xiàndài Hànyǔ duǎnyǔ jiégòu guīzé yánjiū* 面向中文信息处理的现代汉语短语结构规则研究 [Studying the modern Chinese phrase-structure rules in dealing with Chinese language processing], Běijīng 北京: Qīnghuá Dàxué 清华大学出版社, 2000.

Zhān Wèidōng 詹卫东, "Shùkù zài Hànyǔ yǔfǎ fǔzhù jiàoxué zhōng de yìngyòng chūtàn 树库在汉语语法辅助教学中的应用初探" [The application of the treebank to assist Chinese grammar instruction: a preliminary investigation], *Journal of Technology and Chinese Language Teaching* 3/2, 2012a, 16–29, http://www.tclt.us/journal/2012v3n2/zhan.pdf.

Zhān Wèidōng 詹卫东, "Cóng yǔyán gōngchéng de jiǎodù kàn 'zhōngxīn kuòzhǎn tiáojiàn' yǔ 'bìngliè tiáojiàn' 从语言工程的角度看 '中心扩展条件' 与 '并列条件'" [An investigation on the violation of the Head Expansion Principle and syntactic mismatches in coordination constructions in Chinese], *Yǔyán Kēxué* 语言科学 5, 2012b, 449–463.

Zhōu Qiáng 周强, Zhān Wèidōng 詹卫东 and Rén Hǎibō 任海波, "Gòujiàn dàguīmó de Hànyǔ yǔkuàikù 构建大规模的汉语语块库" [Building a large-sacle chunk annotated corpus], in: Huáng Chāngníng 黄昌宁 and Zhāng Pǔ 张普, ed., *Zìrán yǔyán lǐjiě yǔ jīqì fānyì* 自然语言理解与机器翻译 [Natural language understanding and the computer translation], Běijīng 北京: Qīnhuá Dàxué 清华大学出版社, 2001, 102–107.

Zhōu Qiáng 周强, "Hànyǔ jùfǎ shùkù biāozhù tǐxì 汉语句法树库标注体系" [Annotation scheme for Chinese treebank], *Zhōngwén Xìnxī Xuébào* 中文信息学报 4, 2004, 1–8.

Zhū Déxī 朱德熙, *Yǔfǎ jiǎngyì* 语法讲义 [Lectures on syntax], Běijīng 北京: Shāngwù 商务印书馆, 1982.

Weidong Zhan