

Chinese Sentence Pattern Structure (SPS) Treebank V1.0

Authors: Weiming Peng, Min Zhao, Yuchen Song, Jing He, Tianbao Song, Dongdong Guo, Shuqin Zhu, Yinbin Zhang, Jingbo Sun, Zuntian Wei, Jiajia Hu, Jihua Song, Zhifang Sui, Ning Wang

Introduction

The Chinese Sentence Pattern Structure (SPS) Treebank is constructed based on the annotation scheme of diagrammatic Chinese syntactic analysis. It implements Sentence Constituent Analysis (SCA) as its core concept and emphasizes the significance of Sentence Pattern Structure. The method of diagrammatic Chinese syntactic analysis can be derived from famous Linguist Jinxi Li's *The New Chinese Grammar*, in which he named his grammar system as Sentence-based Grammar. Therefore, the historical SPS Treebanks have been called Li's Grammar Treebank, Sentence-based syntactic Treebank, and diagrammatic syntactic Treebank.

The SPS Treebank adopts the innovative and comprehensive annotation scheme containing 3-layer information: lexical sense and structural mode for dynamic words, syntactical structure for clauses, and inter-clause relation within complex sentence and sentence clusters. All the structures can be visualized by a Jbw-viewer tool, and the data is stored in XML format.

The SPS Treebank project began at Beijing Normal University and Peking University in 2012. The project aims to provide a visual and intuitional structure representation of the Chinese discourses. The SPS Treebank V1.0 release contains 27 chapters / 5,016 sentences extracted from some classic selected works including modern Mandarin and ancient Chinese. The SPS Treebank project is on-going and more data will be released in future versions.

Annotation Scheme

(1) Tag Set of the XML

Table 1 shows the Element names for sentence constituents and words (using part of speech, POS as Tag name). In general, the sentence constituent (for content word) is encoded as 3 letters, the position of the function word (i.e. constituent for function word) is encoded as 2 letters, and the word is encoded as 1 letter. The POS system complies with the *Modern Chinese Dictionary*, which also provides the lexical item for the sense encoding (see the @sen in Table 2) of the word in Mandarin.

Table 1 XML Element for constituents and words (part of speech, POS)

Tag	Constituent	Tag	POS
para	Paragraph	n	Noun
ju	Sentence	t	Time word
xj	Clause	f	Localizer
sbj	Subject	m	Numeral
prd	Predicate	q	Measure word
obj	Object	r	Pronoun
att	Attribute	v	Verb
adv	Adverbial	a	Adjective
cmp	Complement	d	Adverbial
ind	Independent constituent	p	Preposition
pp	Position of preposition	c	Conjunction
cc	Position of conjunction	u	Auxiliary word
uu	Position of auxiliary word “的、之/地/得” as a connective symbol in att/adv/cmp	e	Exclamation
un	Position of auxiliary word combined with NP or single word	o	Onomatopoeia
uv	Position of auxiliary word combined with VP or Clause	w	Punctuation
ff	Position of localizer	x	Default

Table 2 XML Attributes and corresponding values

Attribute	Value and its meaning
ju/@dct (dictionary type or lexicon source)	0: Modern Chinese Dictionary (6th ed) for Mandarin 1: Dictionary of Ancient Chinese Common Word (4th ed) for ancient Chinese
<POS>/@sen (Sense of the word)	3-digit number. The first indicates different entries of the word form in the dictionary, and the last two indicate the ordinal index of the sense under the specific entry.
<POS>/@mod	The structural mode of the dynamic word, more details later.
prd/@scp (governing scope of the element prd)	V / VO / VC / VOO / VCO / VOC , where V, O, and C indicate the predicate, object, and complement respectively.
cc/@fun (function of element cc, which is used as a placeholder to represent the relation between the constituents)	COO (between NPs in coordinate structure); APP (between NPs in appositive structure); UNI (between VPs in union predicate structure); SER (between VPs in serial verb predicate structure); PVT (between VPs in pivotal predicate structure); SYN (between VPs in synthetic predicate structure);

<p>$xj/@upt$</p> <p>(the alignment position of the dotted line or curve between related clauses)</p>	<p>Two formats: “$m:n$” (if the related clauses are within the same sentence) or “$k:m:n$”(if the related clauses belong to different sentences), where m indicates the id of the related xj, n indicates the aligned index (after which word) in the xj, and k indicates the offset of the related ju. More details later.</p>
<p>$xj/@rel$</p> <p>(logic semantic relation between related clauses)</p>	<p>时间(Temporal)、因果(Causal)、假设(Hypothetical)、条件(Conditional)、让步(Concessive);</p> <p>承接(Continuous)、并列(Coordinative)、选择(Alternative)、递进(Progressive)、转折(Adversative);</p>
<p>$xj/@sub$</p> <p>(subordinate side)</p>	<p>↑ and ↓, indicate whether the subordinate is upper or lower between the related clauses. (this attribute only exists if the value of the @rel is the first five in the above list)</p>
<p>$xj/@tpc$</p> <p>(topic guidance)</p>	<p>+ (topic refs to the sbj or topic of the connected clause);</p> <p>— (topic refs to the obj of the connected clause)</p> <p>± (topic refs to the whole connected clause)</p> <p>↳ (the connected clause is a leading clause, namely the current and below clauses are the quoted clause)</p> <p>↪ (topic refs to the sbj of the nearest leading clause above)</p>

(2) SPS Diagram Formula

In the visualized representation, the sentence constituents and syntactic relations are depicted by various diagrammatic lines or curves, which follow the SPS Diagram Formula as shown in Figure 1.

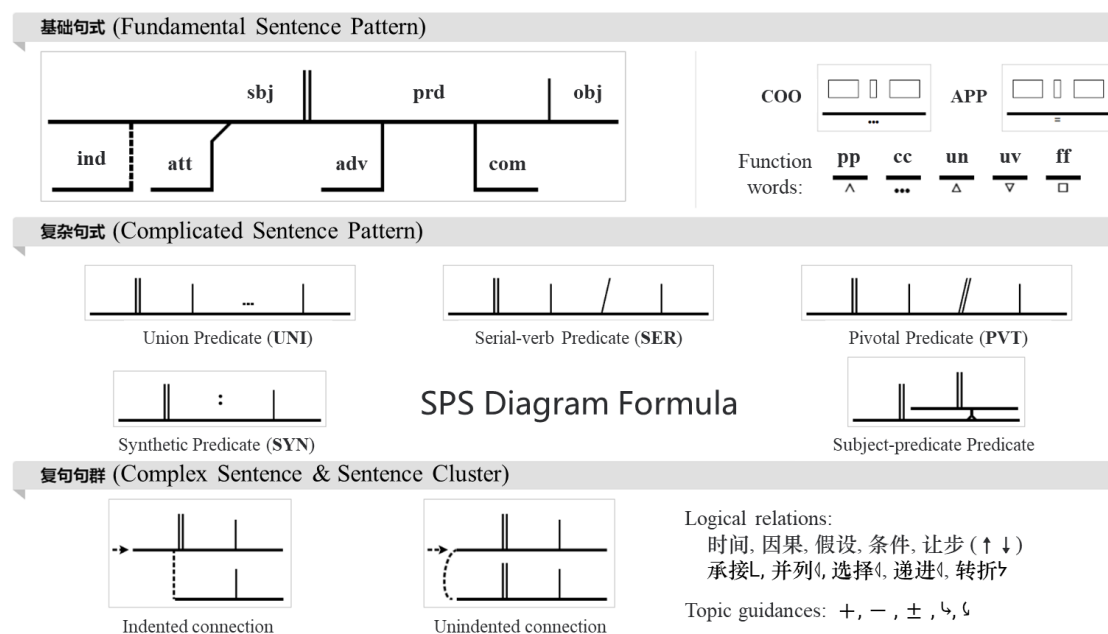


Figure 1 SPS Diagram Formula

The words are put on the horizontal lines in the diagram. The in-vocabulary word is directly annotated the POS and sense code provided by the lexicon. As for the out-of-vocabulary, namely the dynamic word, it is split into a sequence of morphemes that are

in the lexicon. A symbol representing the lexical relation is put between the “<POS><sen>” code of the morphemes, and the POS of the whole dynamic word and a colon is in front of them. The “<POS><syllable num>?(<relation symbol><POS><syllable num>?)+” is stored in @mod the POS node of the whole dynamic word, as shown in Figure 2. The <syllable num> indicates the character number of the morpheme, it will be omitted when equals one.

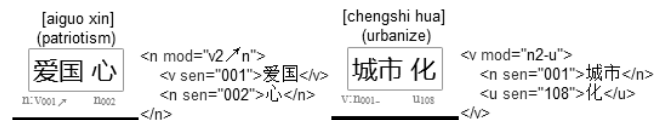


Figure 2 Diagrammatic Style and XML Structure of Dynamic Words

All the lexical relation symbols are illustrated in Table 3.

Table 3 Symbols of Lexical Relation

Symbol	Relationship	Examples
...	Coordination	花 ... 草 [hua cao] (flowers and plants); 中 [zhong](middle)... 小学 [xiaoxue] (primary school)
↗	Attribute-head	鸡 [ji](chicken) ↗ 蛋 [dan](egg); 文字 [wenzi](character) ↗ 改革 [gaige](reform); 北京 [Beijing] ↗ 师范 [shifan](normal) ↗ 大学 [daxue](university);
→	Adverbial-head	极 [ji](extremely) → 具 [ju](possess); 深 [shen](deeply) → 感 [gan](feel); 代 [dai](replace) → 写 [xie](write); 改 [gai](instead) → 用 [yong](use)
←	Verb-complement	赶 [gan](drive) ← 跑 [pao](run); 看 [kan](see) ← 清 [qing](clear); 拿 [na](grasp) ← 下 [xia](down); 举 [ju](lift) ← 起 [qi](up)
	Verb-obj	调 [tiao](mix) 酒 [jiu](wine) ↗ 师 [shi](worker)
	Sbj-prd	你 [ni](you) 争 [zheng](fight)... 我 [wo](I) 夺 [duo](capture)
·	Reduplication	看·看 [kan kan](have a look); 研究·研究 [yanjiu yanjiu](try a research); 看·一·看 [kan yi kan](have a look); 看·了·看 [kan le kan](have a look); 看·不·看 [kan bu kan](to look or not); 看·没·看 [kan mei kan](see or not)
-	Other Lexical Relation	桌 [zhuo](desk)-上 [shang](upside); 一 [yi](one)-只 [zhi]; 一 [yi](one)-大 [da](big)-碗 [wan](bowl); 拿 [na](pick)-得 [de]-起 [qi](up); 看 [kan](look/see)-了 [le]; 看 [kan](look)-着 [zhe]; 看 [kan](look)-过 [guo]; 同学 [tongxue](classmate)-们 [men](-s); 学习 [xuexi](learn)-者 [zhe](-er); 华 山 [Huashan](Hua Mountain)-之 [zhi](of)- 巅 [dian](summit); 付 [fu](put)- 诸 [zhu](into)-实践 [shijian](practice); 翩然 [pianran](lightly)-而 [er]-至 [zhi](come)
◇	Over-Lexical Relation	十 [shi]-元 [yuan] ◇ 八 [ba]-角 [jiao] ◇ 六 [liu]-分 [fen] 呢 [ne] ◇ 吗 [ma]

The above has been published in the following papers which the users could refer to for further information.

- Weiming Peng, Jihua Song, ZhiFang Sui, Dongdong Guo. Formal Schema of Diagrammatic Chinese Syntactic Analysis. The 16th Chinese Lexical Semantics Workshop (CLSW2015). Springer, 2015: 701-710.

- Dongdong Guo, Shuqin Zhu, Weiming Peng, Jihua Song, Yinbing Zhang. Construction of the Dynamic Word Structural Mode Knowledge Base for the International Chinese Teaching. The 17th Chinese Lexical Semantics Workshop (CLSW 2016), Springer, 2016, 10085: 251.

In paragraph-level diagrammatic analysis, the text is split firstly into sentences by punctuation (“。|?|!|:”, and “;” as appropriate), and then the clauses are diagrammatically analyzed using the fundamental and complicated sentence pattern in the SPS Diagram Formula. At last, the diagrams of the clauses within the complex sentence & sentence groups are connected by dotted vertical lines and dotted curves, both of which can be labeled with logical relation and topic guidance.

There are two types of connections:

1) Dotted vertical line with indent: the indented clause shares the leading constituents in the connected clause before the dotted line. An example is given in Figure 3.

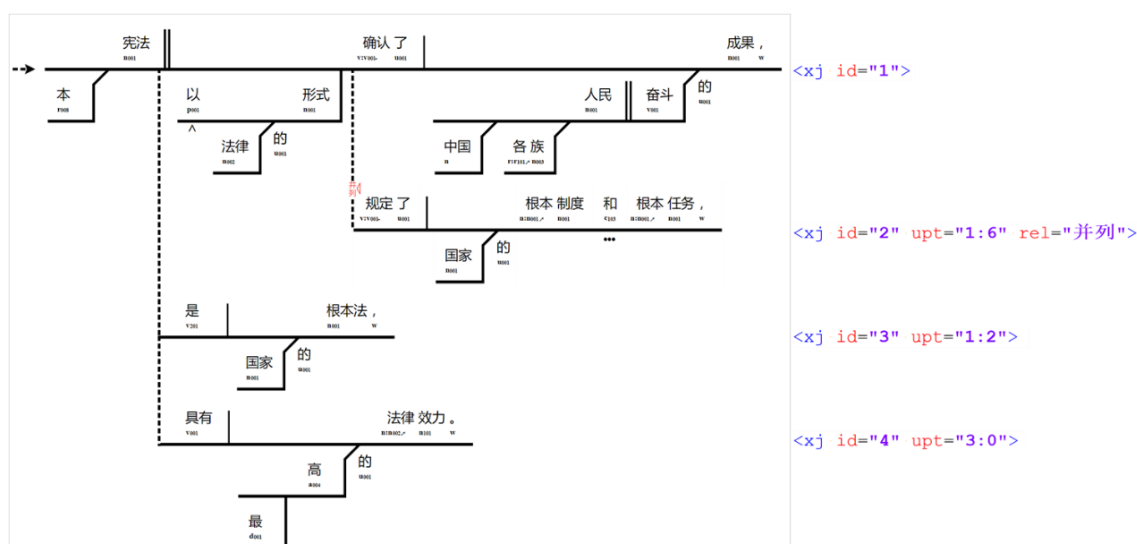


Figure 3 Indented connection between clauses

2) Dotted curve without indent: there is no shared leading constituents between the connected clauses. An example is given in Figure 3, where “1:s”(@upt) means the aligned position is the “start” point of the first clause.

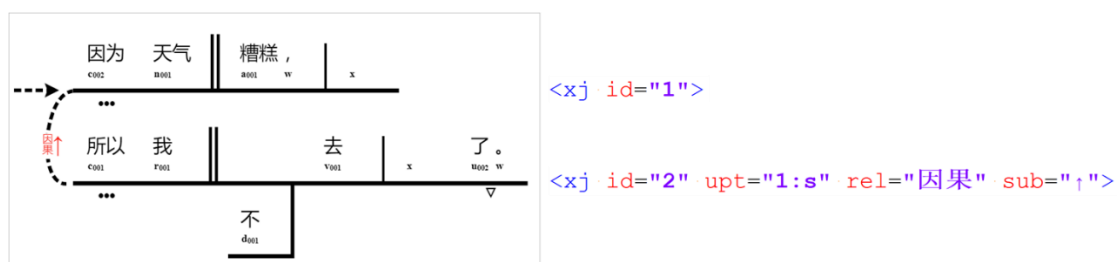


Figure 4 Unindented connection between clauses

Figure 5 provides an example that contains 4 types of topic guidance among clauses,

where the first two and the last four clauses (xj) make up the first and second sentences (ju) respectively. The first two clauses are the leading and quoted clauses. The topic of the 3rd clause could refer to the subject of the leading clause (1st) directly, but since it is separated by the quotation (2nd) which might expand to a long text, it's a better choice using the “ㄣ” as shown. “-1:2:s”(@upt in the third clause) means aligning to the “start” point of the 2nd clause of the preceding one sentence. That is, the first number in the 3-part @upt uses the relative offset other than the @id of the ju.

Since the topics of the last three clauses refer to the object of the 3rd clause semantically, there are 2 choices for the connection of the 5th clause: connecting to the 4th with a “+” or to the 3rd with a “-”. The former is preferred because of the principle of proximity.

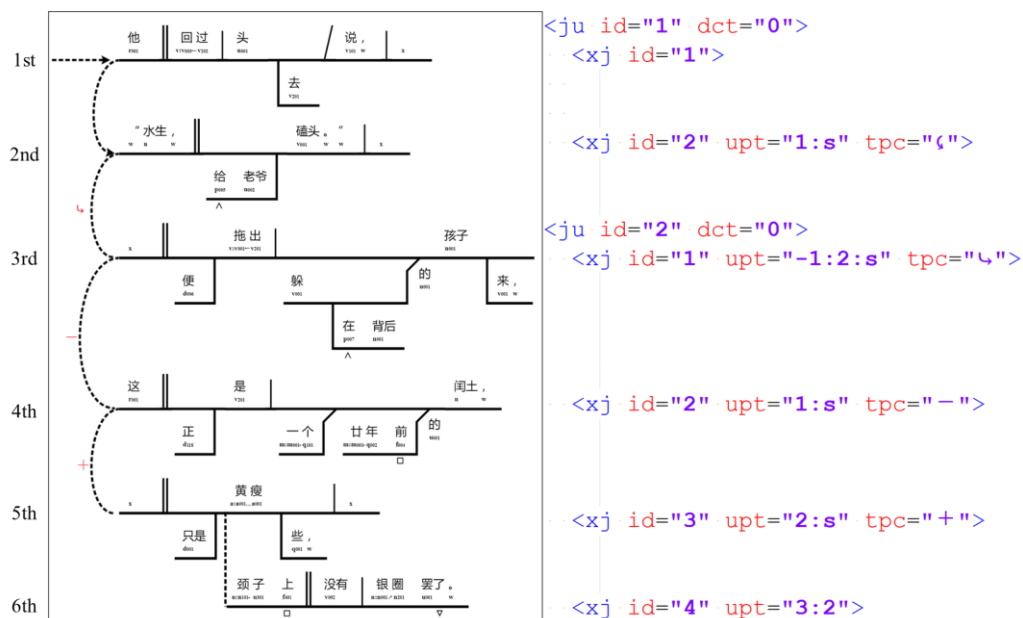


Figure 5 Topic guidance example (1)

Figure 6 provides an example shows the remainder topic guidance value “±” which means the first clause wholly serves as the topic of the second clause.

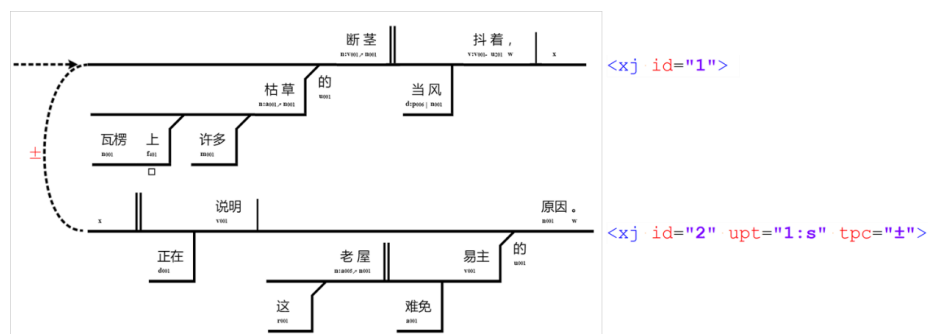


Figure 6 Topic guidance example (2)

Data

The 27 chapters in this release are some classic selected works including modern Mandarin and ancient Chinese. The details are shown in Table 4.

Table 4 Text sources and statistics

Filename	Bookname	Chaptername	Char	Sent
Luxun-01.xml	Selected Work of Luxun (《鲁迅全集》)	Kong Yiji (《孔乙己》)	2602	119
Luxun-02.xml		A Small Matter (《一件小事》)	1027	38
Luxun-03.xml		My Old Home (《故乡》)	4964	223
Luxun-04.xml		Village Opera (《社戏》)	5845	178
Luxun-05.xml		What Happened After Nora Left (《娜拉走后怎样》)	3603	124
Luxun-06.xml		On the Toppling of the Leifeng Pagoda (《论雷峰塔的倒掉》)	1164	36
Luxun-07.xml		Casual Writings Under the Lamplight (《灯下漫笔》)	3993	138
Luxun-08.xml		In Memory of Miss Liu Hezhen (《纪念刘和珍君》)	2347	92
Maozedong-01.xml	Selected Work of Mao Zedong (《毛泽东选集》)	On Practice (《实践论》)	9320	194
Maozedong-02.xml		On Contradiction (《矛盾论》)	23134	577
Feixiaotong-01.xml	From the Soil: The Foundations of Chinese Society (《乡土中国》)	Special Characteristics of Rural Society (《乡土本色》)	3816	130
Feixiaotong-02.xml		Bringing Literacy to the Countryside (《文字下乡》)	3925	122
Feixiaotong-03.xml		More Thoughts on Bringing Literacy to the Countryside (《再论文字下乡》)	3546	117
Feixiaotong-04.xml		Chaxugeju: The Differential Mode of Association (《差序格局》)	4731	163
Caoxueqin-01.xml	A Dream in Red Mansions (《红楼梦》)	Zhen Shiyin in a Dream Sees the Jade of Spiritual Understanding; Jia Yucun in His Obscurity Is Charmed by a Maid (甄士隐梦幻识通灵 贾雨村风尘怀闺秀)	6756	343
Caoxueqin-02.xml		Lady Jia Dies in the City of Yangzhou; Leng Zixing Describes the Rong Mansion (贾夫人仙逝扬州城 冷子兴演说荣国府)	5675	299
Caoxueqin-03.xml		Lin Ruhai Recommends a Tutor to His Brother-in-Law; The Lady Dowager Sends for Her Motherless Grand-Daughter (托内兄如海荐西宾 接外孙贾母惜孤女)	7916	452
Caoxueqin-04.xml		An Ill-Fated Girl Meets an ill-Fated Man; A Confounded Monk Ends a Confounded Case (薄命女偏逢薄命郎 葫芦僧判断葫芦案)	5398	250
Caoxueqin-05.xml		Jia Baoyu visits the Land of Illusion; The fairy Disenchantment performs the 'Dream of Golden Days' (贾宝玉神游太虚境 警幻仙曲演红楼梦)	7342	437
Confucius-01.xml	The Analects of Confucius (《论语》)	Xue Er (《学而》)	654	56
Confucius-02.xml		Wei Zheng (《为政》)	801	72
Confucius-03.xml		Ba Yi (《八佾》)	962	92
Confucius-04.xml		Li Ren (《里仁》)	680	70

Confucius-05.xml		Gong Ye Chang (《公冶长》)	1188	115
Confucius-06.xml		Yong Ye (《雍也》)	1107	112
Mencius-01.xml	Mencius (《孟子》)	Liang Hui Wang I (《梁惠王上》)	3048	210
Mencius-02.xml		Liang Hui Wang II (《梁惠王下》)	3723	257
Total:			119267	5016

The data is provided in the UTF-8 encoding. Each file contains all the 3-layer information that comply with the annotation scheme. All files were automatically verified and manually checked.

Tool




The Jbw-viewer.exe is an SPS diagram visualization toolkit built by Electron (<https://www.electronjs.org/>). The core code is written in Vue/JavaScript language, and published at <https://github.com/bnucip/jbwviewer>. The corresponding editor version of the viewer is at <http://www.jubenwei.com/>.



Figure 7 Main interface of the Jbw-viewer

As shown in Figure 7, the main interface contains 3 blocks: the top is the button bar; the left is the text list; and the main block shows the diagram of the selected sentence/paragraph. The 4 buttons on the top bar are:

- (1) : Show the SPS Diagram Formula.

- (2) : Open a .xml file in this treebank to load the data. All the paragraphs will be listed in the left block.
- (3) : Show the XML and the SPS expressions of the current diagram.
- (4) : About the author and contact.

Acknowledgement

This work is supported in part by the National Science Foundation of China (62007004) and Yingmin Cultural Education Fund of Beijing Normal University.

Updates

We will continue to release more annotated data of the Chinese Sentence Pattern Structure (SPS) Treebank. Please visit our website (<http://www.jubenwei.com/>) for the latest news.

Copyright

©2024 Weiming Peng, ©2010-2024 Beijing Han-Sky Education Technology Co., Ltd.