# 文構造文法に基づいた中国語文法資源 CSSG およびその特許分野での応用

王向莉[1]，宮尾祐介[2]，李元[3]

[1]日本特許情報機構　[2]国立情報学研究所　[3]東京大学

{xiangli_wang@japio.or.jp, yusuke@nii.ac.jp, liyuan@is.s.u-tokyo.ac.jp}

## 1. はじめに

従来の自然言語処理用の中国語文法資源は Penn Chinese Treebank, Peking University Treebank, Tsinghua University Treebank などがあげられる（Yu et al. 2010）。これらの中国語文法資源はいずれも句構造文法(Phrase Structure Grammar; PSG)という文法枠組みに基づいて構築されたツリーバンクであり、統計的構文解析手法のための学習データとして用いられている。これらの文法資源を用いた中国語構文解析手法の精度は英語に比べるとはるかに低く、いまだに実用レベルに達していない(Levy and Manning 2003）。

王、宮崎（2007）は句構造文法に基づく規則は規則間の衝突による不整合が避けられないため、構文曖昧性が爆発的に生じることと、句構造文法は特に中国語のような動詞の形態的変化などのない孤立語を精密に解析しにくいことを指摘した。そして、句構造文法より精密に中国語文を解析できる文構造文法 SSG（Sentence Structure Grammar）を提案した。

既存の中国語文法資源から学習された PSG 規則で中国語を解析する手法の精度が低い原因は PSG という文法枠組み自体は中国語の構文を精密に取り扱えないことにあると考えられる。よって、精度の高い中国語構文解析を実現するために、中国語を精密に解析できる文法枠組みに基づいた新しい文法資源の開発が必要となる。

本稿では、PSG と SSG の２つの文法枠組みを、中国語文を解析する精密さという視点から比較する。孤立語である中国語にある構文的な制約を PSG 規則に取り込みにくいが、SSG 規則にはそれをうまく取り込むことによって、SSG のほうが精密に中国語文を解析できることを示す。また SSG という文法枠組みに基づいて開発された新しい中国語文法資源 CSSG（Chinese Sentence Structure Grammar）を紹介する。最後に CSSG の特許分野の応用について簡単に紹介する。

## 2. 文構造文法 SSG

文構造文法 SSG は依存文法、句構造文法と格文法の特性を併せ持つ文法枠組みである。文構造文法の考え方は以下の 3 点である。

1) 文を述語と述語を中心とした構文要素からなる文構造で解釈する。
2) 文の意味構造を述語と述語を中心とした構文要素との意味上の依存関係として構文木上で直接表示する。
3) 格文法(Fillmore 1968)の観点によって述語動詞を分類する。

図 1 は SSG の構文木を示す。

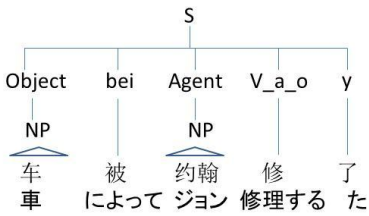


図 1:「車はジョンによって修理された」の SSG 構文木

## 3. SSG と PSG の精密性の比較

中国語では英語のような動詞の形態的な変化などの字面上の構文的な制約がほぼないが、構文要素の順序、構文要素の共起関係、述語と文型の共起関係は中国語を精密に解析する

ための重要な制約になる。

たとえば、「也/も」と「没/ない」はいずれも副詞で、述語の直前に位置できる(文 1a、1b)。文 1c のように、二つの副詞は同時に現れることが可能である。しかし、「也/も」は「没/ない」の後ろに位置することができないという制約があり、文 1d は非文である。このような制約は構文要素の順序制約である。

| PSG 規則 | SSG 規則 |
|---|---|
| S→ NP VP | S→ Agent d1 v |
| VP→ d1 VP | S→ Agent d2 v |
| VP→ d2 VP | S→ Agent d1 d2 v |
| VP→ d3 VP | S→ Agent v |
| VP→ VP PP | S→ Agent v PP |
| NP → 约翰 | Agent → NP |
| VP → 去 | NP → 约翰 |
| PP → 于 彼得 | PP → 于 彼得 |
| d1→ 也 | v → 去 |
| d2→ 没 | d1→ 也 |
| d3→ 很 | d2→ 没 |
| | d3→ 很 |

表 1：シンプルな句構造文法と文構造文法

PSG は構文要素の順序制約を規則にうまく取りこむのが難しい。正しい文 1a、1b、1c を句構造文法 PSG で解析するために、表 1 に示す PSG 規則が必要である。しかし、これらの規則によって、非文 1d も図 2 のように解析されてしまい、非文 1d を除外できない。その一方、文構造文法 SSG は、このような制約を精密に規則に取り込める。文 1a、1b、1c を SSG で解析するために、表 1 に示す SSG 規則が必要である。これらの SSG 規則は非文 1d を解析せずに除外できる。

1a. 约翰　也　去
　　ジョン　も　行く
　　(ジョンも行く)

1b. 约翰　没　去
　　ジョン　ない　行く
　　(ジョンは行かない)

1c. 约翰　也　没　去
　　ジョン　も　ない　行く
　　(ジョンも行かない)

1d. *约翰 没 也 去
　　ジョン　ない　も　行く
　　(ジョンはも行かない)

構文要素の共起関係上の構文的な制約について例を挙げる。たとえば、程度副詞と比較表現は同時に現れないという制約がある。文 2a、2b は正しい文だが、文 2c は非文である。

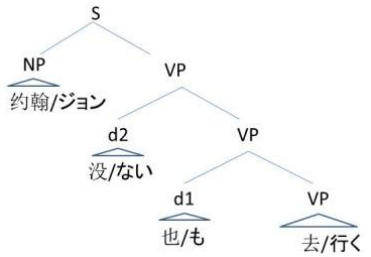


図 2：非文 1d の構文木

PSG はこのような共起制限を規則に取り込みにくい。PSG で文 2a、2b を解析するために、表 1 に示す PSG 規則が必要である。しかし、これらの規則は非文 2c を図 3 のように解析してしまうので、それを非文として除外できない。SSG はこのような共起関係上の制約を精密に規則に取り込める。表 1 に示す SSG 規則で文 2a、2b を解析できるが、非文 2c を解析せずに除外できる。

2a. 约翰　很　高
　　ジョン　とても　高い
　　(ジョンは背がとても高い)

2b. 约翰　高　于　彼得
　　ジョン　高い　より　ペテロ
　　(ジョンはペテロより背が高い)

2c. *约翰　很　高　于　彼得

ジョン とても 高い より ペテロ

(ジョンはとてもペテロより背が高い)

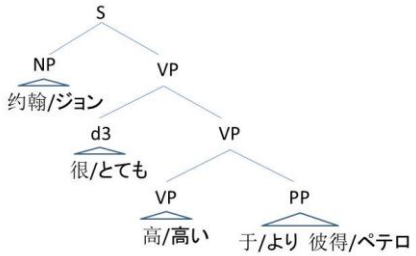


図 3：非文 2c の構文木

述語と文型の共起制約とは、ある種の動詞が、ある文型に現れないという制約である。以下に示す文型 1 は中国語の 1 つの常用な文型である。対象格 Object は文頭に位置し、主格 Agent は受動を表す機能語「被/によって」の後ろに位置する。そして、文末に場所格 Location が現れる。文型 1 は「放/置く」のような動詞と共起可能だが、「读/読む」と「买/買う」のような動詞とは共起できない。そのため、3b、3c は非文である。PSG はこのような制約を取り扱えないが、SSG はそれを利用できる。

文型１：Object by Agent V at Location

3a. 书 被 约翰 放 在 桌上

本 によって ジョン 置く に テーブルの上

(本はジョンにテーブルの上に置かれる)

3b. *书 被 约翰 读 在 桌上

本 によって ジョン 読む に テーブルの上

(本はジョンにテーブルの上に読まれる)

3c. *书 被 约翰 买 在 桌上

本 によって ジョン 買う に テーブルの上

(本はジョンにテーブルの上に買われる)

PSG 規則に取り込みにくい中国語にある重要な構文的制約は、文構造文法 SSG の規則にうまく取り込むことができるため、SSG という文法枠組みは PSG より中国語を精密に解析できる。

## 4. 中国語文法資源 CSSG

CSSG は文構造文法 SSG に基づいて長年にわたってチャート法パーサ上に実装された新しい中国語文法資源である。

| 文型 | 例文 |
|---|---|
| 一般構文 | 他放几本书在桌上<br>(彼は何冊の本をテーブルの上に置く) |
| 「把」文型 | 他把书放在桌上<br>(彼は本をテーブルの上に置く) |
| 「被」文型 | 书被他放在桌上<br>(本は彼によってテーブルの上に置かれる)<br>书被放在桌上<br>(本はテーブルの上に置かれる) |
| 「由」文型 | 书由他放在桌上<br>(本は彼にテーブルの上に置いてもらう) |
| 「得」文型 | 书放得很整齐<br>(本はとてもきれいに置かれる)<br>他把书放得很整齐<br>(彼は本をとてもきれいに置く)<br>书被他放得很整齐<br>(本は彼によってとてもきれいに置かれる) |
| 述語の重複 | 他放书放得很整齐<br>(彼は本をとてもきれいに置く) |
| 目的語前置 | 书放在桌上<br>(本はテーブルの上に置かれる)<br>书他都放在桌上<br>(本は彼がテーブルの上に置かれる) |
| 主題文 | 约翰他把书放在桌上<br>(ジョン、彼は本をテーブルの上に置く) |
| 各構文の関係節 | 把书放在桌上的他<br>(本をテーブルの上に置いた彼)<br>被他放在桌上的书<br>(彼によってテーブルの上に置かれた本) |

表 2：「放/置く」類動詞に対応する文型の一部

CSSG 資源は 7652 の文法規則と 3.6 万単語の構文解析用単語辞書からなる。単語分割さ

れた中国語文(例えば、「车 被 约翰 修 了(車はジョンによって修理された)」) を CSSG に入力すると、図 1 のような SSG 構文木が出力される。

広い構文的網羅性を持たせることを考慮して、CSSG 開発手法を工夫した。CSSG では中国語の述語動詞は格文法の観点によって 32 種類に分類される。そのほか、「是/である」、「进行/行う」などの特殊な述語も収集される。種類ごとに述語動詞を中心とした文型を網羅的に収集し、それぞれの文型に対応する SSG 規則を記述する。たとえば、「放/置く」のような動詞は、主格 Agent、対象格 Object、場所格 Location を必要とする動詞である。表 2 に示すように、このような動詞を中心とした文型の例文を予め網羅的に収集する。次に、収集された例文を網羅的に解析する SSG 規則を記述する。このようにして開発された CSSG は広い構文カバー率を実現した（王、宮崎 2007)。

## 5. CSSG の特許分野での応用について

近年、中国語特許文献の量は著しく増えている。精密性の高い文法枠組み SSG に基づいた高精度の中国語特許文解析を実現することが期待される。CSSG を利用して大量の中国語特許文を解析し、中国語特許文 SSG ツリーバンクを開発することは、その第一歩である。

Japio は、現在 CSSG を用いた中国語特許文タイトルツリーバンクの開発について研究を行っている。この作業は以下の 6 つのステップに分けられる。

1. データを選定
2. データを単語分割
3. 人手で単語分割結果を修正
4. 構文解析辞書に未知語を収録
5. データを CSSG で構文解析
6. 構文木の選択と修正

ステップ 1、2、3、4 はすでに完了した。表 2 に示すように、文字数によって 2.9 万件の特許文タイトルを選定した。この 2.9 万タイトルを京大形態素解析ツール Kytea で単語分割してから人手で修正した。ステップ 5 と 6 はこれから実施する予定である。

| タイトル文字数 | タイトル数 |
|---|---|
| 1-10 文字 | 20000 |
| 11-15 文字 | 5000 |
| 16-20 文字 | 2000 |
| 20 文字以上 | 2000 |
| 合計 | 29000 |

表 3：データの選定

## 6. おわりに

中国語を精密に解析できる文法枠組み SSG に基づいて、新しい中国語文法資源 CSSG を開発した。Japio は精度の高い中国語特許文解析を実現するために、CSSG を用いた中国語特許文タイトルツリーバンクの開発を研究している。将来は豊かな構文情報を出力する CSSG 資源を用いた新しい機械翻訳手法を検討したい。

## 参考文献


Yu, Kun, Yusuke Miyao, Takuya Matsuzaki, Xiangli Wang, Yaozhong Zhang, Kiyotaka Uchimoto, Junichi Tsujii. *Comparison of Chinese Treebanks for Corpus-oriented HPSG Grammar Development.* Journal of Natural Language Processing (Special Issue on Empirical Methods for Asian Language Processing). April 2010.

王向莉, 宮崎正弘. *文構造文法に基づく中国語構文解析*. 言語処理論文誌. vol.14, No.2. April 2007.

Fillmore, Charles J. (1968). *The Case for Case*. In Bach and Harms (Ed.): *Universals in Linguistic Theory*. New York: Holt, Rinehart, and Winston, 1-88.

Roger Levy and Christopher Manning. *Is it Harder to Parse Chinese, or the Chinese Treebank?* Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL2003).