ツリーバンキングのための文法枠組みに関する考察

王向莉[1]，松崎拓也[1]，宮尾祐介[2]，Kun Yu[1]，李元[1]，辻井潤一[134]

[1]東京大学
[2]国立情報学研究所
[3]School of Computer Science, University of Manchester
[4]National Center for Text Mining
Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-0033, Japan
{xiangli, matuzaki, kunyu, liyuan, tsujii}@is.s.u-tokyo.ac.jp

1　はじめに

ツリーバンクは自然言語処理及び言語学研究のための重要な資源である。ツリーバンクはある文法枠組みに基づいてテキストに統語構造を付与することで作成される。そのため、選択された文法枠組みはツリーバンクから得られる文法情報の種類を決めるだけでなく、ツリーバンクの構築の効率と作成されたツリーバンクの品質、およびツリーバンクを構築する際の方法論にも深くかかわっている。

本稿では、いくつかの代表的な文法枠組みについて、表示される情報の種類、アノテーターにとっての表示の直観性、および文法規則の一般性などの特性に着目しながら整理し、それらの特性が各枠組みによるツリーバンキングに与える影響について論じる。最後に、(i)アノテーターにとって表示が分かりやすい文法、(ii)文法規則の一般性のよい文法、との相反する特性をもつ2つの文法枠組みを組み合わせたアノテーション手法について、その構想を示す。

2　文法枠組みごとの整理

2.1 依存文法 DG

DGでは、図1に示すように、統語構造がある単語とその従属部の関係として定義され、句ノードのような情報がない。PDT(J. Hajic et al., 2000)は依存文法に基づいて作成された典型的なツリーバンクである。PDTは形態素情報、構文構造情報、意味構造情報の三つの段階でアノテートされた。

2.2 句構造流文法

ここでの句構造流文法はチョムスキーの提案およびそれに基づいて発展してきたすべての枠組みを指す。本稿では、特に、文脈自由句構造文法および語彙化文法の一種である HPSG の2つについて考える。

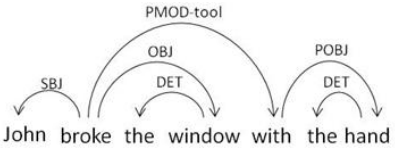

図1：依存文法に基づく構文木

2.2.1　文脈自由句構造文法 CF-PSG

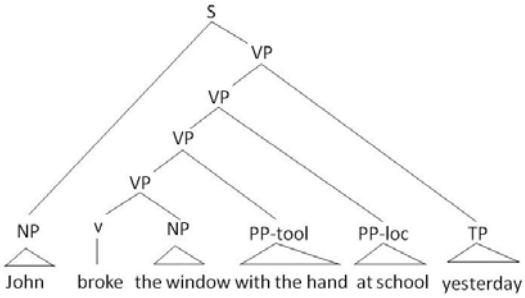

図2：句構造文法に基づく構文木

CF-PSGはツリーバンキングのためによく選ばれる文法枠組みである。Penn Treebank (Marth et al., 2005) は典型的なCF-PSGに基づくツリーバンクである。CF-PSGは図2に示すように文1aを句構造で解釈する。CF-PSGに基づくツリーバンクでは、意味構造を直接表示しない場合が多い。例えば、1aと1bは同じ意味構造を持つと考えられるが、それぞれ文の句構造による表示では、この意味の同一性は直接表示されない。また、句構造による表示では、文1cと文1dのように構文木上の述語ー項の位置関係と意味役割が1対1に

対応しない場合に意味構造の同一性を直接読み取るのは困難である。

1a. John broke the window with the hand at school yesterday

1b. the window was broken by John with the hand at school yesterday

1c. John broke the window

1d. the window broke

2.2.2 主辞駆動句構造文法 HPSG

語彙化文法の一種である主辞駆動句構造文法 HPSG(Pollard and Sag, 1994) は PSG の拡張であり、CF-PSG と同じように文を句構造で解釈するが、構文木の各ノードに置かれるデータ構造（Sign）の中で、意味構造が直接表示される。この反面、Sign による表示は一般に複雑なものとなり、表示されている統語・意味構造をアノテーターが直観的に把握することは難しいと考えられる。

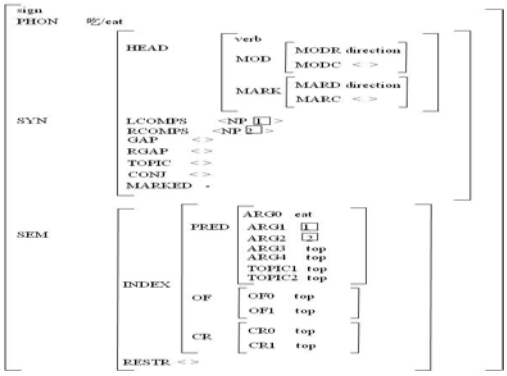

図 3：HPSG の語彙項目の例

2.3 文構造文法

文構造文法 SSG（王,宮崎 2007）は依存文法と句構造流文法の特性を併せ持つ文法枠組みである。CF-PSG と比べると、基本的な区別が二つある。

1) 文を細かな句構造で解釈するのでなく、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。具体的には、各述語に対する必須の構文要素および付加的な構文要素を、述語とともに構文木上の 1 つのレベルにまとめて表示する。

2) 文の意味構造を、述語と述語の周囲の構文要素との意味的依存関係として構文木上で直接表示する。

文 1a を例にして、SSG はどのように文を解釈するかについて説明する。文の述語は”broke”であり、その前の名詞句”John”が主語であるので、”Sn”で表示する。名詞句”the window”は目的語であるため、”On”で表示する。前置詞句”with the hand”は道具で、 “at school”は場所で、 時間詞句“yesterday”は時間の要素であるため、それぞれ PP-tool、PP-loc と TP で表す。すべての要素が図 3 に示すように文構造規則 1)に記述する。

図 4 と図 5 に示すように、文 3a と文 3b はそれぞれ、規則１）と規則２)で解析する。2 つの文のどれにあっても、”John”は意味上の主語であり、”the window”は意味上の目的語である。

規則 1) s → Sn V On PP-tool PP-loc TP

規則 2) s → On BE V BY Sn PP-tool PP-loc TP

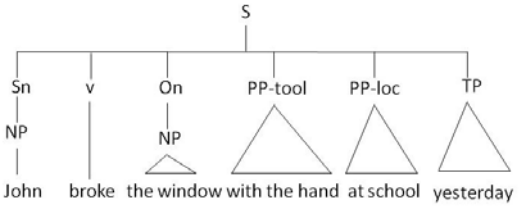

図 4：文構造文法に基づく構文木(1)

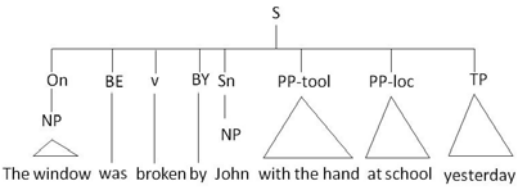

図 5：文構造文法に基づく構文木(2)

SSG では、文 2a のように複数の述部を持つ並列構造を扱うのが難しい場合がある。これは、主語位置の NP がそれぞれの節に対して異なる意味役割（意味上の主語と目的語）を持つため、主語位置の NP に対応するノードのラベルとしてそれらを表示することができないためである。

このような場合、図 6 のように、主語が必要である節に CL_Sn_gap、目的語が必要となる節に CL_On_gap というラベルを与えることで意味・

統語構造を両方表示するが、図 4 のような単純な場合と異なる取り扱いとなり、また、意味的な依存関係を表示から直観的に読み取ることは難しいという問題がある。ただし、現実的にはこのような構造が現れる頻度はそれほど高くないと考えられる。

2a. John slept on road and was robbed.

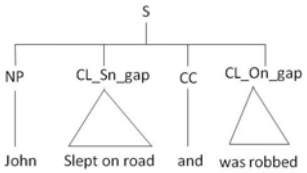

図 6：SSG の節の処理

3　文法枠組みの比較

3.1 文法情報の豊かさと直観性の比較

図 1 に示すように、DG は述語とそれに直接依存する従属部の意味上の関係を直観的に分かりやすく表示できるが、句構造の情報が欠けている。そのため、例えば単純な名詞句などであっても、構文要素のかたまりが直観的に把握しにくい。

これに対し、PSG では句構造を利用することで、スコープなどの情報を表示できる場合がある。また、PSG に基づくコーパスである Penn Treebank では、句構造に加えて -SBJ、-OBJ といった文法機能タグを句ラベルに付加する形でアノテートしている。しかし、述語・項関係のように、構文木上の位置関係と必ずしも一対一に対応しないような情報をアノテートする場合、句ラベルをさらに拡張することで表示する方法は（原理上は可能であるにせよ）アノテーターにとって見やすい表示であるとは言えないだろう。HPSG のような語彙化文法は PSG で表示されるような構文構造に加え、述語項構造のような意味情報を表示するためのシステムを含んでいる。しかし、既に図 3 に示したように統語情報と意味情報、さらに両者の関係を同時に含む表示は非常に複雑になる場合があり、これを直接アノテーターに提示するための表現形とするのは難しい。

SSG は構文構造情報と意味構造情報を分けずに 1 つの文構造規則で表示するため、図 4 と図 5 に示すように、意味上の主語 Sn と意味上の目的語 On のような述語と項、および修飾句との意味関係を構文木上で直接表示できる。また、図 5 に示したような場合を除けば、構文木上の一つのレベルに述語とそれに依存する句が並ぶため、CF-PSG や HPSG の意味表示のような複雑な記法を必要としない。また、ある程度まで句構造の情報を表示するため、特に頻度の多い名詞句などは構文要素としてのかたまりを直観的に把握できる。さらに、空範疇と co-indexing の仕組みを導入することで、HPSG では解析が難しい、文 3a のような例を図 7 のように述語項関係が見やすい形で表示することも可能である。

3a. John likes apples and Mary oranges.

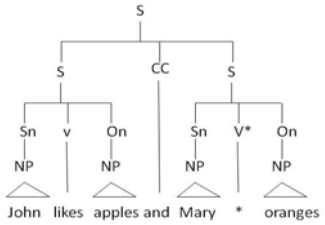

図 7：SSG でアノテートしやすい文例

3.2 方法論の比較

文法枠組みの文法情報の豊かさと直観性がツリーバンキングの方法論にかかわっている。従来のツリーバンクは静的な方法と動的な方法のどちらかに従って作られたものである。

静的方法論とは、複数のアノテーターが自分の言語的な直感に頼って、1 つずつの構文木をアノテートするという方法である。従来の DG および PSG に基づくツリーバンクは静的な方法に従って構築されるものが多い。

動的な方法とは、あらかじめ文法規則を用意しておき、その文法規則にしたがって解析した結果を文に付与する方法である。動的な方法に従って、語彙化文法の一種である主辞駆動句構造文法HPSGに基づいて構築されたツリーバンクである。

DG や PSG のような文法が比較的乏しい文法枠組みに基づいてツリーバンキングをする場合、広範

囲の文を被覆するためには過剰生成する文法を使わざるを得ない場合がしばしばあるため、動的な方法に向いていない。

HPSG のような詳細な文法的制約を表現できる文法は、DG や PSG よりも、動的な方法に向いている。その一方、HPSG のような語彙化文法は、その統語・意味構造の表示が非常に複雑なため、文法をあらかじめ用意することなしで、アノテーターが一文ずつアノテートするのは非常に難しい。そのため、語彙化文法は静的な方法に向いていないと言える。

SSGはDGと同様に表示の直感性に優れ、PSGと同程度の単純な形式をもつため、静的方法によるアノテーションが可能だと考えられる。また、各文法規則を述語によって語彙化することで語彙化文法と同様の詳細な制約が記述できるため、動的方法論でツリーバンキングすることも考えられる。

4 新しい方法論

統語構造・意味構造がともに表示でき、かつ直観性のよい SSG をインターフェスにし、SSG に基づいてツリーバンキングをするのと並行し、文法変換規則を使って、アノテートされた SSG 構文木を、ほかの文法枠組み（例えば一般性に優れた文法を記述可能であるが、表示が複雑である HPSG のような語彙化文法）における構文木に変換し、同時に複数の枠組みでツリーバンクを構築する方法が考えられる。

ここで、HPSG を例として、この方法論を実現する可能性を検討する。Miyao (2006) は、CF-PSG による解析を HPSG による解析へと変換するルールを人手によって作成することで、PennTreebank を HPSG ツリーバンクへと半自動的に変換する方法を提案している。このような手法に基づいて、アノテートされた平坦な SSG 木を HPSG 木へ変換することが考えられる。図 8 に文 1a から、変換された部分導出木を示す。SSG の各構文規則における述語は、HPSG における主辞にほぼ対応し、HPSG における解析へと変換する際に必要となる統語・意味情報はおおむね SSG のひとつの文法規則に含まれているため、SSG の各文法規則を単位として、HPSG での解析へと変換する規則を作成することはそれほど難しくないと考えられる。

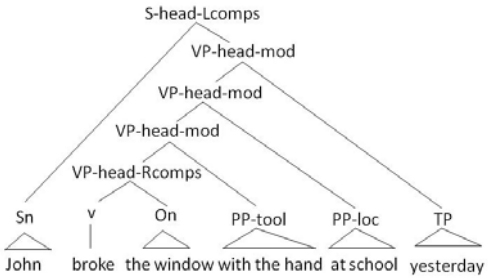

図 8：部分導入木

5 結論と展望

本稿では、ツリーバンキングという側面から、依存文法 DG、PSG 流文法および文構造文法 SSG を比較した。同時に複数の文法枠組みのツリーバンキングをする方法を検討した。

参考文献

Martha Palmer and Daniel Gildea and Paul Kngsbury (2005). *The Proposition Bank: An Annotated Corpus of Semantic Role*. In Computational Linguistics. Vol. 31 Issue 1, March 2005.

Stephan Oepen, Dan Flickinger, Kristina Toutanova, Christoper D. Manning.(2002). LinGo Redwords: A Rich and Dynamic Treebank for HPSG. In Proc. TLT 2002.

Mitchell P. Marcus, B. Santorini and Mary Ann Marcinkiewicz (1994).*Building A Large Annotated Corpus of English: The Penn Treebank.* Computational Linguistics, Vol. 19, No. 2. (1994), pp. 313-330.

Bond F., S. Fujita, C. Hashimoto, D. Kasahara, S. Nariyama, E. Nichols, A. Ohtani, T. Tanaka, S. Amano (2004).*The Hinoki Treebank: Working Toard Text Understanding.* In LINC-04.

Carl Pollard, Ivan A. Sag (1994). *Head-Driven Phrase Structure Grammar.* Chicago: University of Chicago Press.

王向莉, 宮崎正弘(2007). *文構造文法に基づく中国語構文解析*. 自然言語処理, vol.14 no.2, pp.69-93.

Jan Hajic, Alena Bohmova, Eva Hajicova, Barbora Vidova Hladka (2000). The Prague Dependency Treebank: A *Three-Level Annotation Scenario*. In A. Abeillé (ed.): *Treebanks: Building and Using Parsed Corpora*, Amsterdam:Kluwer, 2000, pp. 103-127.

Miyao, Yusuke (2006). *From Linguistic Theory to Syntactic Analysis: Corpus-Oriented Grammar Development and Feature Forest Model*. PHD Thesis.