HPSGに基づくフリーな日本語ツリーバンクの構築

Francis Bond　栗林 孝行　橋本 力♠
情報通信研究機構
♠ 山形大学大学院理工学研究科
bond@ieee.org, kuribayashi@khn.nict.go.jp, ch@yz.yamagata-u.ac.jp

1　はじめに

本論文では、現在我々が構築しているツリーバンクを紹介する。このツリーバンクは、Jacyというオープンソースの日本語HPSG文法を用いて、パブリックドメインのコーパスに詳細な統語情報と意味情報を付与する形で構築している。このツリーバンクの特長は、**i)**自由に配布可能なこと、**ii)**統語と意味に関して詳細なアノテーションがなされていることの2点である。[1]

2　ツリーバンクの素材

本ツリーバンクでは、Jacy (Siegel and Bender, 2002)と呼ばれる計算機用日本語HPSG文法と、田中コーパス(Tanaka, 2001)と呼ばれるパブリックドメインの日英対訳コーパスの日本語部分の2つが素材として用いられている。

2.1　日本語HPSG文法：Jacy

ツリーバンクの統語・意味情報は、Jacyが依拠しているHPSG (Head-driven Phrase Structure Grammar) (Pollard and Sag, 1994)の枠組に基づいている。HPSGは「制約に基づく語彙主義の文法」と呼ばれ、次のような特長を持つ。

1. 統語と意味を同時並行的に解析する
2. 少数の原理により規則性を捉え、豊かな語彙情報により個別的な現象に対応する
3. 素性構造に基づく厳密な形式化により、言語処理技術との親和性が高い

制約に基づく語彙主義の文法として、HPSG以外に、LFG (Lexical-Functional Grammar) (Bresnan and Kaplan, 1982)やCCG (Combinatory Categorial Grammar) (Steedman, 2000)等がある。これらの文法理論もHPSGと同様に上記の特長を備えているが、日本語文法の研究の蓄積はHPSGに比べて少ない。

Jacy以前にも計算機用の日本語HPSG文法はいくつか開発されてきた(Gunji, 1987; Nagata, 1992; 大谷ら, 2000; 金山ら, 2000) 。Gunji (1987)と大谷ら (2000)の文法は、小規模ながら精緻な理論的記述を重視し、一方、Nagata (1992)と金山ら (2000)の文法は、理論的な緻密さよりも実用性を重視したと言える。

Jacyは、従来のものと違い、理論的な緻密さと実用性（語彙数は5万語以上）の両方を兼ね備えている。さらにJacyには、従来のものには無い次のような特長がある。

1. フリーウェアとして公開されている
2. 茶筌[2]を用いた未知語処理機能を備えている
3. 言語現象の記述能力に優れ、言語処理技術との親和性も高い最小再帰意味論(Minimal Recursion Semantics: MRS) (Copestake et al., 2005)を意味記述に採用している
4. Jacyと同一の理論と開発ツールに基づいて、互換性が非常に高い、数多くの言語の計算機用文法が構築されている

上記に関して、Butt et al. (2002)でも、LFGに基づいて、互換性の高い多言語の計算機用文法を構築している。しかし、それらの文法はフリーウェアとして公開されてはいない。

2.2　田中コーパス

田中コーパスとは、もともと、兵庫大学の学生によって収集・作成された212,000文対の日英対訳用例集を故・田中康仁氏が編纂したものである。その後、JMDictの和英辞書プロジェクト(Breen, 2003)で例文として利用され[3]、そこで非文や重複を修正・削除して、最終的に約16万文対となった。

[1] 本研究のツリーバンクと構築時に利用した日本語文法はhttp://wiki.delph-in.net/moin/JacyTopから入手可能である。

[2] http://chasen.aist-nara.ac.jp/

[3] さらに近年、TATOEBA (http://wwwcyg.utc.fr/tatoeba/)において、フランス語訳など多くの情報が徐々に追加されている。

表 1: ツリーバンクの現状

	Type	Number	%
OK	良い木（1本）	7,809	52.06
	良い木（数本）	679	4.53
NG	良い木無し	1,604	10.69
	解析無し	2,826	18.84
	解析エラー	2,082	14.01
	総数	15,000	100

田中コーパスの最大の魅力は自由に利用・配布することが可能な点であり、それが本研究でこのコーパスを採用した理由である。なお本研究では、この中の15,000文を対象にツリーバンクを構築する。

対訳文対の例を(1)に挙げる。和文と英文は元の田中コーパスから、フランス語訳はTATOEBAからである。

(1) あの木の枝に数羽の鳥がとまっている。

“Some birds are sitting on the branch of that tree.” (en)

“Des oiseaux se reposent sur la branche de cet arbre.” (fr)

収録されている文は平均11.6形態素／文のシンプルなものが多いが、文法的なバリエーションは豊富なため言語資料として申し分ない。

3 田中ツリーバンク

3.1 構築手法

ツリーバンクは、一文ごとに、Jacyによる（複数の）解析結果（構文木と意味表象）の中から正しいと思われるものを選択する形で構築していく。この選択作業が統語・意味情報のアノテーションに相当する。図1に解析結果（上部が構文木で、下部が意味表象）の例を挙げる。

この「解析→アノテーション」サイクルは通常複数回行われ、徐々にツリーバンクが完成に近づく。1サイクルごとに文法を改良してツリーバンクを更新するが、その更新は差分に対してのみ効率的に適用される(Oepen et al., 2004)。

3.2 進捗状況

現在、ツリーバンク構築の第1サイクルを終えたところであり、現状は表 1の通りである。つまり、全15,000文中、56.59% (52.06+4.53)にあたる8,488文(7,809+679)が正しく解析され、ツリーバンクに組み込まれている。

3.2.1 解析に成功したもの

正しい解析結果（表1の「OK」）は「良い木（1本）」と「良い木（数本）」の2つに分けられる。前者は、一文に対してただひとつの正しい解析が出力されたもの、後者は、複数の正しい解析が出力されたものである。後者に関して、本研究では、構文構造がわずかに異なるが意味表象は同一であるような複数の解析結果は、その意味表象が正しければ全て正しい解析とみなしている。

3.2.2 解析に失敗したもの

解析に失敗したもの（表1の「NG」）は43.41%にあたる6,512 文であり、これらは今現在ツリーバンクには含まれていない。解析の失敗は「良い木無し」「解析無し」「解析エラー」の3つに分けられる。

「良い木無し」とは、Jacyの出力した解析の中に正しいものが含まれていない場合である。これはさらに、文法規則の不足に起因するものと、田中コーパスの入力ミスに起因するものの2つに分けられる。前者に関して、例えば(2)では「恐怖心を隠すため」が副詞節を構成するが、Jacyの出力の中に当該箇所が副詞節を構成している解析が存在しなかった。また(3)では「山の眺め」が全体で1つの名詞句になるべきだが、出力の中にそのような解析がなかった。

(2) 彼女は恐怖心を隠すため笑った。

(3) この部屋からの山の眺めは最高だ。

後者の田中コーパスの入力ミスに起因する「良い木無し」の例として(4)がある。

(4) 彼女の保険体育の抗議はさっぱり理解できない。

これは「保健」が「保険」、「講義」が「抗議」として誤って入力されているものだが、たとえ正しく解析されてもその意味表象はどうしても誤ったものになってしまう。現在、このような入力ミスの修正作業が進んでいる。

「解析無し」は解析結果がひとつも出力されなかった場合であり、「良い木無し」と同様、文法規則の不足に起因するものと、田中コーパスの入力ミスに起因するものの2つに分けられる。(5)は、「お+連用形+する」という敬語表現を解析するための規則が無かったため

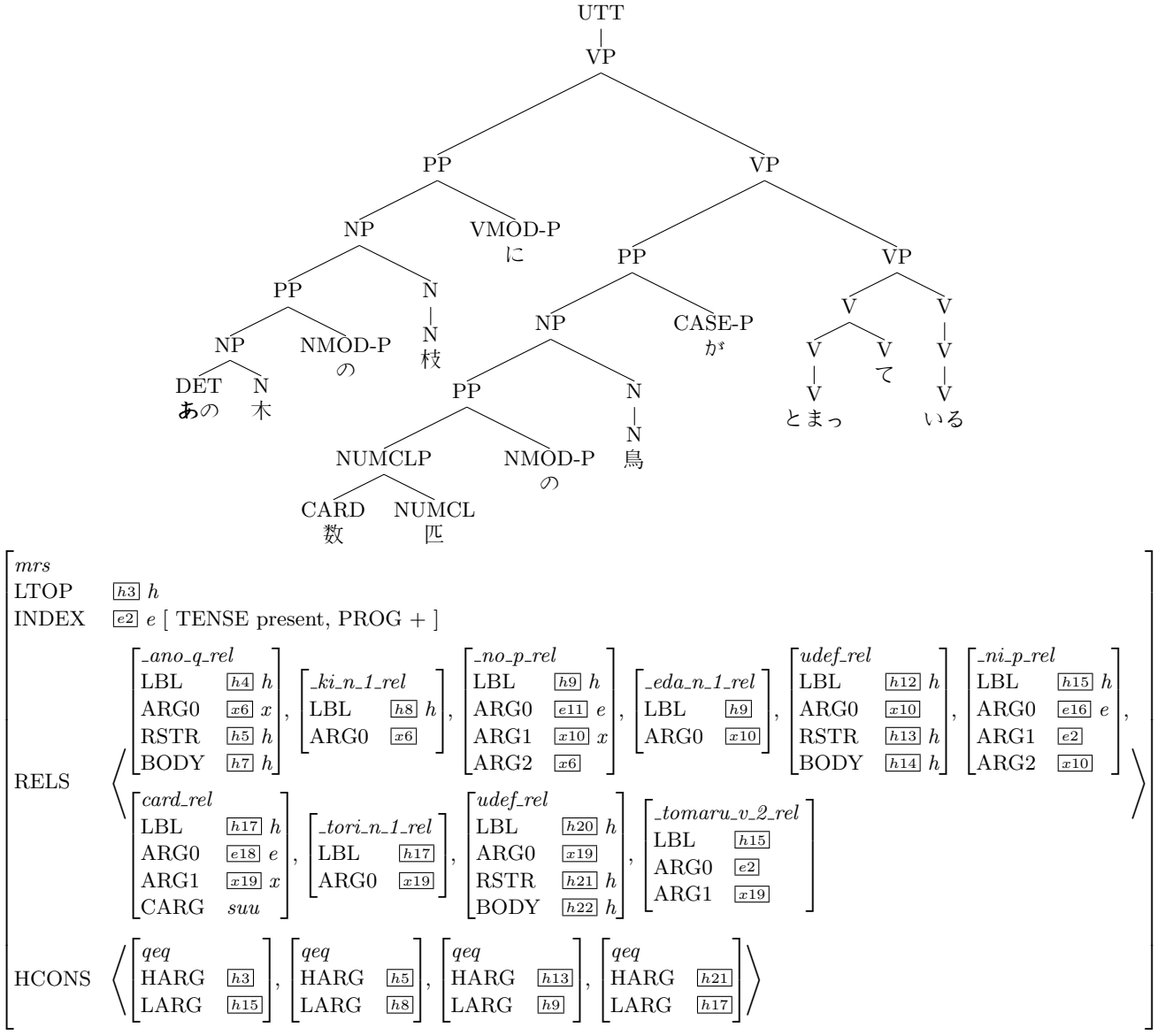

図 1: (1) 「あの木の枝に数羽の鳥がとまっている」の解析木とMRS

に、(6)は日常的に使う口語表現を解析する規則が無かったために「解析無し」となった。

(5) 喜んでご招待をお受けします。

(6) 先方のお名前をどうぞ。

(7)はコーパスの入力ミスであり、本来は「彼はよい医者になるだろう。」である。

(7) 彼はよい医者ななるだろう。

「解析エラー」は、語彙不足あるいは処理のタイムアウトに起因するエラーであり、「解析無し」と同様、解析結果が全く得られない。現在、語彙の追加と解析器のメモリ割り当て量の増量で対応を試みている。

3.3 ツリーバンクの応用

完成した田中ツリーバンクは、NLPシステムの学習データや評価データ、あるいは、日本語文法記述のための言語資料として使用できる。

実際、田中ツリーバンクと互換の「檜」ツリーバンク(Bond et al., 2006)は、以下の研究において学習データあるいは言語資料として使用された。

Tanaka et al. (2007)とFujita et al. (2007)、Blunsom and Baldwin (2006)は、それぞれ、「檜」ツリーバンクを学習データとして、語義曖昧性解消システム、Parse Rankingシステム、Super Taggerを学習した。これらの研究では、「檜」ツリーバンクの詳細

な統語・意味情報を活用することで、非常に高い精度を得ることに成功した。田中ツリーバンクには、前述の通り、これと同様の詳細な統語・意味情報が付与されている。

Zhang et al. (2007)では、Jacy等のHPSG文法のための語彙獲得システムを開発し、その評価に「檜」ツリーバンクを使用している。

Hashimoto et al. (2007)では、「檜」ツリーバンクの言語的情報と、ツリーバンク構築に使用した文法（Jacy）に実装されている文法規則をマッシュアップして、例文つき文法辞典とも言える文書を自動生成した。

4 おわりに

本稿では、現在我々が構築している日本語ツリーバンクについて述べた。その特長は、**i)**自由に配布可能なことと、**ii)**統語と意味に関して詳細なアノテーションがなされていることの2点である。

今後もアノテーションのサイクルを継続し、新しいバージョンの公開を目指している。

その後、京大コーパス(Kurohashi and Nagao, 2003)と比較するため、新聞文からHPSGのツリーバンクを構築する予定である。

References

Phil Blunsom and Timothy Baldwin. 2006. Multilingual deep lexical acquisition for HPSGs via supertagging. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 164–171. Association for Computational Linguistics, Sydney, Australia.

Francis Bond, Sanae Fujita, and Takaaki Tanaka. 2006. The Hinoki syntactic and semantic treebank of Japanese. *Language Resources and Evaluation*, 40(3–4):253–261. (Special issue on Asian language technology).

James W. Breen. 2003. Word usage examples in an electronic dictionary. In *Papillon (Multi-lingual Dictionary) Project Workshop*. Sapporo.

Joan Bresnan and Ronald M. Kaplan. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. In J. Bresnan, editor, *The Mental Representation of Grammatical Relations*, pages 173–281. MIT Press, Cambridge, MA.

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The parallel grammar project. In *Proceedings of COLING 2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal Recursion Semantics. An introduction. *Research on Language and Computation*, 3(4):281–332.

Sanae Fujita, Francis Bond, Stephan Oepen, and Takaaki Tanaka. 2007. Exploiting semantic information for HPSG parse selection. In *ACL 2007 Workshop on Deep Linguistic Processing*, pages 25–32. Prague, Czech Republic.

Takao Gunji. 1987. *Japanese Phrase Structure Grammar: A Unification-Based Approach*. D. Reidel (Kluwer), Dordrecht.

Chikara Hashimoto, Francis Bond, and Melanie Siegel. 2007. Semi-automatic documentation of an implemented linguistic grammar augmented with a treebank. *Language Resources and Evaluation*. (Special issue on Asian language technology).

Sadao Kurohashi and Makoto Nagao. 2003. Building a Japanese parsed corpus — while improving the parsing system. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, chapter 14, pages 249–260. Kluwer Academic Publishers.

Masaaki Nagata. 1992. An Empirical Study on Rule Granularity and Unification Interleaving Toward an Efficient Unification-Based Parsing System. In *Proceedings of the 14th Conference on Computational Linguistics (COLING'92)*, pages 177–183. Morristown, NJ, USA.

Stephan Oepen, Dan Flickinger, and Francis Bond. 2004. Towards holistic grammar engineering and testing — grafting treebank maintenance into the grammar revision cycle. In *Beyond Shallow Analyses — Formalisms and Statistical Modelling for Deep Analysis (Workshop at IJCNLP-2004)*. Hainan Island.

Carl Pollard and Ivan A. Sag. 1994. *Head Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

Melanie Siegel and Emily M. Bender. 2002. Efficient deep processing of Japanese. In *Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization at the 19th International Conference on Computational Linguistics*, pages 1–8. Taipei.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press.

Takaaki Tanaka, Francis Bond, Timothy Baldwin, Sanae Fujita, and Chikara Hashimoto. 2007. Word sense disambiguation incorporating lexical and structural semantic information. In *The 2007 Joint Meeting of the Conference on Empirical Methods on Natural Language Processing (EMNLP) and the Conference on Natural Language Learning (CoNLL)*, pages 477–485. Prague.

Yasuhito Tanaka. 2001. Compilation of a multilingual parallel corpus. In *Proceedings of PACLING 2001*, pages 265–268. Kyushu.

Yi Zhang, Timothy Baldwin, and Valia Kordoni. 2007. The corpus and the lexicon: Standardising deep lexical acquisition evaluation. In *ACL 2007 Workshop on Deep Linguistic Processing*, pages 152–159. Prague, Czech Republic.

金山博, 鳥澤健太郎, 光石豊, and 辻井潤一. 2000. 3つ以下の候補から係り先を選択する係り受け解析モデル. 自然言語処理, 7(5):487–490.

大谷朗, 宮田高志, and 松本裕治. 2000. HPSGにもとづく日本語文法について— 実装に向けての精緻化・拡張—. 自然言語処理, 7(5):19–49.