4B1-1

統語・意味コーパスの統合と再解釈による大規模な日本語CCG文法の開発

Integrating Multiple Dependency Corpora for Inducing Wide-coverage Japanese CCG Resources

植松 すみれ[*1] Sumire Uematsu　松崎 拓也[*2] Takuya Matsuzaki　花岡 洋輝[*3] Hiroki Hanaoka　宮尾 祐介[*4] Yusuke Miyao　美馬 秀樹[*1] Hideki Mima

[*1]東京大学知の構造化センター
Center for Knowledge Structuring, The University of Tokyo

[*2]国立情報学研究所社会共有知研究センター
Research Center for Community Knowledge, National Institute of Infomatics

[*3]東京大学大学院情報理工学系研究科コンピュータ科学専攻
Department of Computer Science, Graduate School of Information Science and Technology, The University of Tokyo

[*4]国立情報学研究所コンテンツ科学研究系
Digital Content and Media Sciences Research Division, National Institute of Infomatics

This paper describes a method of inducing wide-coverage CCG resources for Japanese. While deep parsers with corpus-induced grammars have been emerging for some languages, those for Japanese have not been widely studied mainly because most Japanese syntactic resources are dependency-based and are not trivialy converted to phrase structures. Our method first integrates multiple dependency-based corpora into phrase structure trees and then converts the trees into CCG derivations. The method is empirically evaluated in terms of the coverage of the obtained CCG lexicon and the accuracy of parsing with the grammar.

1. はじめに

日本語の構文・意味解析の研究においては，文節間の係り受け解析を行った後に述語項構造の解析を行うパイプライン処理のアプローチが一般的である [Kudo 02][Hayashibe 11]．それに対し，英語，ドイツ語等では HPSG，CCG といった語彙化文法理論に基づき，句構造解析と述語項構造解析を統合的に行う解析も実現されている [Miyao 08][Hockenmaier 06]．とくにコーパスから自動獲得した文法を用いた場合，実世界テキストの解析にも成果を上げており，出力された述語項構造を利用した応用タスクや，出力に DRS(談話表示構造) を割り当てることでさらに深い解析を行う研究も行われている [Bos 04]．このような語彙化文法に基づいた解析を日本語においても実現することで，日本語意味解析の発展が期待される．

本稿では，京都大学テキストコーパス (京大コーパス)[Kurohashi 98] 並びに関連するコーパスから，CCG 文法理論に基づく日本語文法を自動獲得する手法について述べ，実際に獲得した文法を解析に用いた際の精度評価を示す．また自動獲得の際に問題となる構文情報アノテーションの不足など，語彙化文法獲得の点から見た日本語テキストリソースの課題についても述べる．

語彙化文法理論に基づく包括的な日本語文法構築の試みとしては，古くは HPSG 理論に基づく JPSG[Gunji 87] があり，またコーパスからの自動獲得を行うものとしては [Yoshida 05],[小嶋 06] などが挙げられる．ただし，実際に実世界テキストを用いて解析と評価を行った例は著者らが知る限り [Yoshida 05] のみであり，CCG に関しては今回が初の報告となると考えられる．

連絡先: 植松 すみれ，東京大学知の構造化センター，東京都文京区本郷 3-7-1，03-5841-0897，uematsu@cks.u-tokyo.ac.jp

2. 語彙化文法に基づく日本語文法の自動獲得

2.1 CCG 理論に基づく日本語文法

CCG は語彙化文法理論の一種であり，少数の組み合わせ規則と詳細な辞書情報によって言語のふるまいを説明する．詳細な辞書とは各単語についてその統語的・意味的ふるまいを細かく指定する，ということである．CCG の場合，統語的ふるまいは統語カテゴリと呼ばれる記法 (図 1 の $S\backslash NP$ など) で指定され，意味論上の役目はラムダ式を用いて表される (図 1 では省略)．文の解析では単語に統語カテゴリを割り当て，組み合わせ規則に従って句を構成しながらカテゴリの合成・ラムダ式操作を行なう事で，文の統語構造と意味表現が得られる．

本研究では，CCG 理論に基づいた日本語文法論 [戸次 10](以下戸次文法と呼ぶ) を目標とすべき文法の基礎とした．ただし，2.3 節で述べる問題等から戸次文法の説明する全ての現象を扱うわけではなく，用言のかき混ぜ，受け身，使役など重要かつ頻出する現象を正確に扱い，他の部分については簡略化した文法を目標として設定した．意味表現についてもラムダ式ではなく，用言について簡単な述語項構造を定義し深層格 (動詞を基本形に直した際のガヲニト格) に対する認識精度を評価した．

2.2 コーパス指向文法開発

本研究ではコーパス指向文法開発と呼ばれるアプローチを適用して語彙化文法による日本語解析器を構築した．そのような解析器には，文法 (組み合わせ規則と辞書)，解析アルゴリズム，曖昧性解消モデルが必要となるが，組み合わせ規則は少数であり，解析アルゴリズムは基本的に CKY と変わらないことから，いかに辞書と曖昧性解消モデルを得るかが課題となる．コーパス指向文法開発は図 2 に示したように，構文情報がついたコーパスから目標文法での正しい導出木コーパスを作ることで，辞書と曖昧性解消モデルの学習データをまとめて得るアプローチである．元のコーパス内に導出木の構築に十分な情

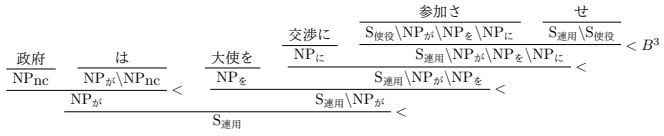

図 1: 「政府は大使を交渉に参加させ (た)」の CCG 導出木. 「大使を」と「交渉に」の部分構造は省略してある

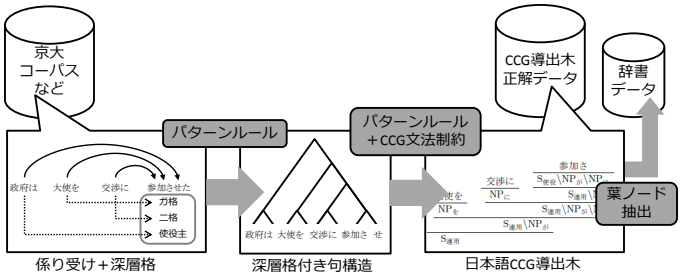

図 2: コーパス指向文法開発の流れ

報があり，正しい導出木を構成できれば，その葉ノードの ⟨ 単語, カテゴリ ⟩ の組から辞書をつくる事ができる．また曖昧性解消モデルは，導出木コーパスでの導出木を正例，同じ文に対して獲得した辞書（と組み合わせ規則）から導出可能な導出木を負例として学習することできる．

2.3 日本語文法獲得における課題

語彙化文法を獲得する元となるデータとして，同一文に対する統語構造アノテーションと意味構造アノテーションがともに必要となる．日本語に対する大規模な統語構造つきコーパスとしては EDR 日本語コーパスおよび京大コーパスがある．本研究では，現在も活発な付加アノテーションが行われている京大コーパスを用いることにした．京大コーパスは統語情報として形態素解析，文節境界および文節間の係り受け関係のアノテーションを含み，同一テキストに対するアノテーションとして，NAIST テキストコーパス [飯田 10] による述語項構造および照応関係の情報，また,「と」コーパス [Hanaoka 10] による助詞「と」でマークされた項と述語の関係の情報がある．本研究では京大コーパス，NAIST コーパスおよび「と」コーパスから得られる情報を統合的に利用する．

目標となる CCG 導出木を得るに当たって，上記 3 コーパスからは読み取れない統語・意味構造は，発見的ルールによる情報付加，あるいは人手による付加的アノテーションなど，何らかの方法で補う必要がある．そのような情報付加が必要な例として，まず，京大コーパスからは得られない句構造の情報が挙げられる．句構造がないと読み取れない情報としては，助詞，助動詞のスコープや並列された句の範囲など基本的なものに加え，固有名詞，普通名詞，量化された普通名詞といった名詞句の下位区分 (非終端記号) の認定など，意味的な分類に近いものも含まれる．

さらに，意味構造と統語構造を併せた形での付加情報が必要になる例として，NAIST コーパスによる述語-項関係のうち，目標文法によって直接決まる関係と，照応やある種の推論を経由して決まる関係との識別が挙げられる．格要素と述語，連体節と被修飾名詞句など，項となる句と述語となる句の関係が統語的に単純な場合，文法的に決まる項の認定は比較的容易である．しかし，表層的な依存関係がない「太郎が + (走って + 転んだ)」における「太郎」と「走って」のような句の間の関係は，NAIST コーパスではゼロ照応としてアノテーションされている一方で，我々の文法では文法によって認識すべき関係となっている．このように，獲得の元となるコーパスにおける分析枠組みと獲得すべき目標文法における分析のギャップが大きいケースでは，やや複雑な処理による識別が必要となる．

3. コーパス変換

コーパス指向文法開発で日本語文法を獲得する場合 2.3 節で述べたように京大コーパスの係り受け構造，NAIST コーパスの述語項構造などを統合して CCG 導出木へ変換する必要がある．目標とする CCG 文法に依存しない中間表現として句構造データを想定し，第 1 ステップとして京大コーパス並びに関連コーパスから，それらのアノテーションを統合した句構造データへの変換，第 2 ステップとして句構造データから目標 CCG 文法の導出木への変換，の 2 段階に分けて変換を行うこととした (図 3).

3.1 第 1 段階：係り受けから句構造へ

係り受け構造から句構造への変換は，大きく分けて (1) 文節内の構造の認定 (2) 文節間の係り受け関係から句の間の関係への変換 の 2 段階からなる．(1) では，文節内の形態素列を 2 分木としてまとめ上げる．この際，体言を中心とした文節の場合は，中心となる複合名詞を右下がりの木になるようにまとめ，最後に助詞を右上がりの順に付加した．用言を中心とし，中心となる用言にいくつかの助動詞や形式名詞が続く形の文節の場合は，全体が右上がりの木構造とすることを原則とした．いずれの場合も，非終端記号はそれぞれの句の最後の形態素 (記号や一部の接尾辞は除く) の品詞ラベルをそのまま用いることとした．

文節内の句構造と非終端記号を決定する処理は，例外的なケースを小さな CFG として表したうえで，その CFG と上記の原則を併せ，決定的な構文解析を各文節について行う処理として実現した．CFG の規則を詳細化することで，より精緻な分析が可能になるが，複合名詞の内部構造や，複合名詞が全体として固有名詞として振舞うかどうか，といった決定は，本来的に人手によるアノテーションを必要とするため，現在の結果は近似的なものである．

(2) の文節間の係り受け関係を句構造に変換する処理は，(1) で決定した各文節に対する部分木を，係り受け構造の通りに組み合わせることで行う．この際，係り側の句の部分木を，受け側の部分木のどの位置に接合するかを決定する必要がある．接合する位置の決定は，係り側の句の種類，最右の形態素の活用形，京都コーパスの係り受け関係ラベル (並列句か否か) などを基にしたルールに従って行った．

3.2 第 2 段階：句構造から CCG 導出木へ

句構造から導出木の変換においては，原則として句構造と木の形がほぼ同じ導出木を想定し，句構造に統合された述語項構造アノテーションなどから導出木の素性に制約をかける方法をとった．具体的には (1) 句構造木のノードに対応する導出木ノードにカテゴリ制約をかける，(2) 二分木で表された句構造の各枝分かれについて，導出木での組み合わせ規則を割り当てる，(3) 導出木の根ノードに S(文を表すカテゴリ) を割り当てた上で (2) で指定された組み合わせ規則を逆に適用して導出木を獲得する，(4) 導出木の葉ノードに制約をかける，の 4 段階で変換を行った．

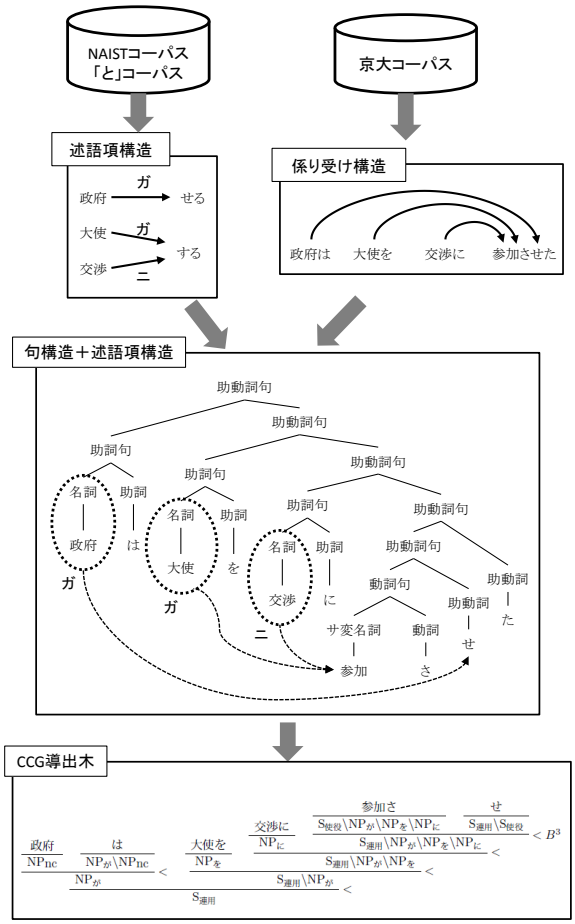

図 3: コーパス変換例．文は図 1 と同じ

(1) のノードに対してカテゴリ制約をかける段階では，文法上カテゴリの形が定まっている形態素のカテゴリに予め制約をかける（例えば助動詞にはカテゴリ $S\backslash S$ など），用言に対し項となる格助詞句 (句構造に統合された述語項構造アノテーションにより判定) のノードに格の制約を与える，など導出木の特定部分に文法制約を与えることとした．

(2) では句構造の各枝分かれについて，副詞句と用言のような修飾句と主辞の句の組，格助詞句と動詞句のように項となる句と主辞の句の組など，左右のノードの性質ごとに組み合わせ規則を割り当てた．

(3) では導出木に (2) で割り当てた組み合わせ規則による制約をかけ，(1) で与えた制約と整合性のある導出木をつくることとした．制約が相互矛盾する場合は変換が失敗することとなり，想定文法で扱えない文法現象を含む文や，制約条件の誤り等を検出することができるようになっている．

最終段階では導出木の葉ノードに制約をかけ，全てのカテゴリ素性を決定するようにした．特に動詞に対応する葉ノードの場合，活用形，態などによりカテゴリに受け身接続形か否か，終止形か否かなどの素性を与えた．またかき混ぜを含んだり，ガ格の省略がある動詞の場合，標準形カテゴリ (ガ格の項があり，かつヲ格，ニ格，ト格の順に項がならぶ) を想定し，実際に得られたカテゴリは標準型からの派生カテゴリとして辞書に登録した．

	総文数	変換文数	変換成功率 (%)
第 1 段階目	24,283	24,116	99.312
第 2 段階目	24,116	22,820	94.626

表 1: 訓練セットにおける CCG 導出木への変換成功率

形態素に対するカテゴリ正解率	
開発セット	90.86
テストセット	90.69

カテゴリ間依存関係	LP	LR	LF	UP	UR	UF
開発	82.55	82.73	82.64	90.02	90.22	90.12
テスト	82.40	82.59	82.50	89.95	90.15	90.05
述語項構造						
	LP	LR	LF	UP	UR	UF
開発	78.66	76.45	77.54	86.67	84.23	85.43
テスト	78.48	76.75	77.61	86.65	84.74	85.68

表 2: 開発・テスト各セットでの解析精度 (%)．LP, LR, LF はそれぞれラベル付きの適合率，再現率，F1 値を表す．UP, UR，UF はそれぞれラベルなしの適合率，再現率，F1 値

4. 実験と評価

京大コーパス Version 4.0，NAIST コーパス Version 1.5，「と」コーパス Version 1.0 を用いて実際に文法を抽出し，獲得した文法による解析実験を行った．実験では京大コーパスを用いた係り受け解析の実験 [Kudo 02] と同様にデータを分割し，訓練セット，開発セット，テストセットを用意した．

4.1 コーパス変換と辞書抽出

まず文法開発でのコーパス変換については，京大コーパスおよび関連コーパスから句構造への変換，句構造から CCG 導出木への変換を行った結果，訓練セット (全 24,283 文) から 22,820 の導出木が得られ，93.98%の文を導出木へ変換することができた．表 1 は変換の各段階における変換成功率であり，第 2 段階の句構造から CCG 導出木への変換において，変換時にかけた制約の相互矛盾によって文法制約を満たさない導出木が検出されていることがわかる．

次に導出木の葉ノードから ⟨ 形態素基本形, 品詞, 活用形, カテゴリ ⟩ の組を抽出し，辞書を構成した．導出木コーパス (総語数は 615,121 語，カテゴリ種総数 662) から得られた辞書エントリ，つまり上記の 4 つ組の種類数は 84,620 個となり，カテゴリ曖昧性，つまり ⟨ 形態素基本形, 品詞, 活用形 ⟩ に対する平均カテゴリ数は 12.440 となった．これらに未知語用エントリを加え，得られた辞書の総エントリ数は 86,232 となった．

4.2 辞書の被覆率とパーザの解析精度

辞書被覆率とは，未知文中の形態素に対して正しいカテゴリが辞書に登録されている割合を示すものである．今回は訓練セットから抽出した辞書の被覆率を開発セット，テストセットに対して測定した．測定では，開発・テストセットに文法開発時と同様の係り受け—導出木変換を施して得られる導出木を正しい木とみなし，その導出木の葉ノードから 4 つ組 ⟨ 形態素基本形, 品詞, 活用形, カテゴリ ⟩ を取り出し，これが辞書に含まれている場合被覆されているとした．開発セットの 100 語以下の文に対して被覆率を測った結果，この文法の辞書による被覆率は 99.4503 であり高被覆な辞書が得られたといえる．

訓練セットから得られた文法を使ってテストセットの文を解

析した結果の精度を表 2 に示す．解析には [Miyao 08] で用いられた語彙化文法用パーザと曖昧性解消モデルを用い，入力は正解形態素付きの文とした．またテスト文に対する正解としては，テストセットに訓練セットと同様の変換を行って得られた導出木を正解の導出木として扱った．評価については 2.1 節で述べた述語項構造の他にカテゴリ依存関係とよぶ基準でも行った．カテゴリ依存関係とは例えば図 1 の「参加さ」ように述語となる形態素に割り当てられた組み合わせカテゴリ（図 1 で $S\backslash NP\backslash NP$) とその項となるカテゴリ (同じ例の NP) の head word を組とした関係である．正確には 5 つ組 ⟨ 組み合わせカテゴリが割り当てられた形態素，組み合わせカテゴリ，組み合わせカテゴリの引数カテゴリ，引数カテゴリの主辞となる形態素 ⟩ を想定し，すべての要素が正解と一致した場合にはラベル付きで正解とし，1 番目と 5 番目の組が正解に含まれればラベルなし正解とした．述語項構造による評価では，4 つ組 ⟨ 用言，その用言のとる格 (ガヲニト格) の組み合わせ，格，格を埋める句の主辞となる形態素 ⟩ を想定し同様に測定した．

今回の解析精度と英語における CCG パーザの精度 [Clark 07] とは，言語や実装の違いもあり直接比べられるものではない．ただし我々の日本語文法の被覆率は英語 CCG 文法の被覆率 (99.63%) とほぼ同じレベルであるのに対し，単語へのカテゴリ割り当て正解率と解析精度は英語の場合と比べて 4〜5 ポイント低くなっている（英語のカテゴリ割り当て正解率は 94.32%，解析精度はカテゴリ間関係 LF で 87.64)．単語単位，構造単位で曖昧性解消での課題が示唆された．

5. おわりに

文の統語構造と意味表現を同時に解析する深い日本語解析器の実現のため，コーパス指向文法開発とよばれるアプローチで高被覆な日本語 CCG 文法を開発し，未知文の解析実験によって評価を行った．

今回の文法開発では，深い日本語解析器構築の第一段階として，CCG 文法理論に基づいた戸次文法を基礎とし，動詞のかき混ぜ，受け身，使役等の重要な現象は正確に扱うが，他の部分については簡略化した文法を目標とし開発を行った．コーパス指向文法開発では，予め句構造によってあらわされた統語情報コーパスの存在が前提とされることが多いが，日本語において一般的な係り受け表現で表された統語情報を句構造へ変換する方法は自明でない．本稿では複数の日本語コーパスに含まれる統語情報と意味情報を統合して句構造へ変換し，さらに CCG 導出木へと変換する手法を示した．またその過程で明らかになった現状の日本語資源に「不足」するアノテーションについても指摘を行った．

また得られた文法を用いて未知文の解析を行ったところ，辞書の被覆率は高く，実世界テキストの解析に充分であると考えられる．ただし解析精度は英語 CCG の場合と比べてやや低くなっており，曖昧性解消に課題があることが示唆された．

今後の課題としては解析器の精度向上はもちろんのこと，さらに精緻な文法を目指すために導出木変換を正確にすること，変換の質を評価することの 2 点があげられる．1 点目については，精密な文法を想定して導出木変換を行うには，現状で発見的ルールで行っている文節内句構造の決定などをより正確に行う必要がある．そのためには上述したような「不足」情報をアノテーションすることなどが考えられる．また 2 点目では，今回の実験では係り受け—導出木変換が完全に正しいと仮定した上で精度評価等をおこなっているが，全ての導出木が想定した文法内で妥当である保証はない．CCG 専門家による変換後導出木 (の一部) の評価を行うことで，変換の質を推測し，また改良の指針を得ることができるだろう．

参考文献

[Bos 04] Bos, J., Clark, S., Steedman, M., Curran, J. R., and Hockenmaier, J.: Wide-Coverage Semantic Representations from a CCG Parser, in *Proceedings of COLING '04*, pp. 1240–1246, Geneva, Switzerland (2004)

[Clark 07] Clark, S. and Curran, J. R.: Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models, *Computational Linguistics*, Vol. 33, No. 4 (2007)

[Gunji 87] Gunji, T.: *Japanese Phrase Structure Grammar: A Unification-based Approach*, D. Reidel (1987)

[Hanaoka 10] Hanaoka, H., Mima, H., and Tsujii, J.: A Japanese Particle Corpus Built by Example-Based Annotation, in *Proceedings of LREC'10* (2010)

[Hayashibe 11] Hayashibe, Y., Komachi, M., and Matsumoto, Y.: Japanese Predicate Argument Structure Analysis Exploiting Argument Position and Type, in *Proceedings of the 5th IJCNLP*, pp. 201–209 (2011)

[Hockenmaier 06] Hockenmaier, J.: Creating a CCGbank and a wide-coverage CCG lexicon for German, in *Proceedings of the Joint Conference of COLING/ACL* (2006)

[Kudo 02] Kudo, T. and Matsumoto, Y.: Japanese Dependency Analyisis using Cascaded Chunking, in *Proceedings of CoNLL 2002* (2002)

[Kurohashi 98] Kurohashi, S. and Nagao, M.: Building a Japanese Parsed Corpus while Improving the Parsing System, in *Proceedings of LREC 1998* (1998)

[Miyao 08] Miyao, Y. and Tsujii, J.: Feature Forest Models for Probabilistic HPSG Parsing, *Computational Linguistics*, Vol. 34, No. 1, pp. 35–80 (2008)

[Yoshida 05] Yoshida, K.: Corpus-Oriented Development of Japanese HPSG Parsers, in *the 43rd ACL Student Research Workshop* (2005)

[小嶋 06] 小嶋 大起, 戸次 大介, 宮尾 祐介, 潤一 辻井 : 日本語 CCG の語彙項目獲得 (語彙・概念の獲得と同義語), 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2006, No. 124 (2006)

[戸次 10] 戸次 大介 : 日本語文法の形式理論—活用体系・統語構造・意味合成—, 日本語研究叢書, くろしお出版 (2010)

[飯田 10] 飯田 龍, 小町 守, 井之上 直也, 乾 健太郎, 松本 裕治 : 述語項構造と照応関係のアノテーション : NAIST テキストコーパス構築の経験から, 自然言語処理, Vol. 17, No. 2, pp. 25–50 (2010)