統語・意味解析情報付き日本語コーパスの構築

プラシャント・パルデシ　　吉本　啓
国立国語研究所　　東北大学

要旨

国立国語研究所共同研究プロジェクトで開発中の統語･意味解析情報付きコーパス NINJAL Parsed Corpus of Modern Japanese (NPCMJ) は，日本語に対し句構造をタグ付けした初めてのコーパスである。これにより，日本語研究者や日本語教員が構文にもとづいて言語データを検索，入手することが容易になる。本稿では，まず NPCMJ 開発の動機について述べ，さらにコーパス構築の基本方針および手続きについて説明する。続いて NPCMJ の基本特徴について述べ，述語-項構造のアノテーションに関して，他の先行コーパスとの比較を行う。またウェブでの公開とウェブ・インタフェースについて解説し，隣接分野との連携について述べる。最後に本コーパスの意義について考察する。

1　はじめに

コーパスは日本語研究の手段としても広範囲に利用されるようになっている。しかし，従来日本語に関して利用可能なコーパスは，文節（アクセント句）を単位として形態素や係り受けに関する情報を付加したものに限られていた (Kurohashi & Nagao 2003, Maekawa et al. 2014)。文法研究者の多くが関心を持つ様々な構文に関して，これらのコーパスはきわめて限定された知識しか提供しない。

そこで私たちは国立国語研究所共同研究プロジェクトとして，現代日本語のテクストに対し，統語・意味解析情報を付加したコーパス NINJAL Parsed Corpus of Modern Japanese (NPCMJ; バトラー他 2016) の開発を始め，順次公開中である。NPCMJ は，文の意味を直接反映する句構造 (Phrase Structure) を日本語に対してタグ付けした，初の本格的で一般に入手可能なコーパスである。これによって，日本語の研究や教育に携わる人々が関心のある文法事象や文型にマッチする言語データを容易に入手できるようになる。また同時に，その時々の必要や対象データに応じてダイナミックに文法情報を抽出して利用することを可能にする。

---

* 本研究は日本学術振興会科研費基盤 (B) 15H03210，基盤 (C) 16K02654，および国立国語研究所共同研究プロジェクト「統語・意味解析コーパスの開発と言語研究」の助成を受けた。

NPCMJ 開発の最大の目的は日本語文法研究に資することにあるので，この報告ではこの観点から説明を行う。しかし，NPCMJ が提供する精緻な統語・意味情報は，外国語としての日本語の教育や自然言語処理にも有力な手段を提供すると考えられる。

以下では，まず第 2 節で NPCMJ 開発の動機について述べる。続いて第 3 節でアノテーションの基本方針を，第 4 節でコーパス構築の手順について説明する。第 5 節では本コーパスの特徴について述べる。第 6 節では，述語-項構造を取り上げて，NPCMJ と既存の日本語コーパスとの比較を行う。第 7 節で，コーパスの公開とウェブ上でのインタフェース利用について解説する。第 8 節で，今後の改訂および隣接分野との連携について説明する。最後に，第 9 節で，将来にわたる日本語研究において占める本コーパスの意義について考察する。

2　開発の動機

2.1　なぜ統語解析情報が必要か

日本語についても近年コーパスが整備され，一般にも利用可能になっている。しかし，それらは文節へのセグメンテーションと文節間の係り受け情報の付加および形態素情報のアノテーションから成っている。そのため，文法に関心を持つ研究者にとっては限定的な意義しか持たない。

まず，形態素情報だけではなぜ不十分かについて説明する。文の理解にあたって，文法役割 (格) が最重要な情報であることはよく知られている。日本語において文法役割は格助詞 (「が，を，に」など) により表示されることが多いが，それだけに限られない。文法役割を持つ名詞句に「は，も」等の係助詞が付加され，格助詞は使用されないこともある。また，格助詞が脱落したり，あるいは名詞句が省略されて格助詞も使われないことがある。さらに，連体修飾の一種である「内の関係」の場合，修飾される主名詞は連体節の中の述語に対し文法役割を果たすが，ここでも格助詞は使われない。他方，一つの格助詞が表示する文法役割は一つとは限らない。例えば，格助詞「が」は主語を表す他に，「たい」を伴うなどのいくつかの構文で直接目的語を表示する。このように，文法役割を表示する形態論的手段 (省略されている場合を含む) と表示される文法役割とは，多対多の関係にある。そのため，特定の文法役割を持つ名詞句を過不足なく検索するためには形態論的情報だけでは不十分で，文法役割をあらかじめアノテートしておく必要がある。

さらに，日本語の文を文節へとセグメントし，文節間の関係を係り受けとして捉えることは，文の意味理解という点で問題を生じる。文節とは，日本語の音声・音韻的側面から見た単位であるアクセント句とほぼ同一であり，橋本 (1934) によって文の基本的構成要素とされた。文節は，1 個の自立語 = 内容語 (名詞，形容詞，動詞，副詞等) およびそれに後続する 0 個以上の付属語 = 機能語 (助詞または助動詞) から成る。文節にもとづく文解析/生成モデルは日本語の音声・音韻処理に利用され，また構造が単純

であることから，初期の言語処理で使用された。
例えば次の文で，

(1) a. [冷蔵庫 に] [牛乳 が] [入ってい ない]
b. [[冷蔵庫に牛乳が入ってい] ない]

(1a) の文節セグメンテーションに従うと，否定辞「ない」は「入っている」のみをスコープとして否定していると解釈されることになる。しかし，この場合，牛乳そのものが存在しないと考えられるので，(1b) のように「ない」が文の残りの部分をスコープとする統語構造を与える必要がある。

一般に，付属語が「独立」したスコープを持つ統語構造がしばしば必要とされることは，すでに 1960 年代から指摘されていた (三上 1970: 33-34)。このような，意味にもとづいて文を解析して得られる統語構造は句構造 (Phrase Structure) と呼ばれる。意味を反映する正確な検索結果を得るには，句構造をタグ付けしたコーパスが必要である。

2.2 なぜ意味解析情報が必要か

文の伝える意味が，表層的な統語構造が伝える以上のものを含むことは広く知られている。

(2) a. 昨日**買っ**たりんご
b. 昨日**買っ**て冷蔵庫に入れたりんご
c. 昨日**買っ**て冷蔵庫に入れたと思ったりんご

例えば動詞「買う」が直接目的語として取る名詞を言語データの中から網羅的にリストアップする課題を与えられた場合，(2a) のように直接目的語が主名詞として，その述語を含む関係節により修飾されている例を除外することはできない。主名詞には関係節内での文法役割を示す情報は表示されていない。(2a) の場合は当該の述語と直接目的語とが隣接していることが手掛かりになると思われるかもしれない。しかし，両者の間は (2b,c) のように隔てられていることもある。このように，関係節による修飾は非有界依存構文 (unbounded dependency) の一種なので，両者の距離は理論上無限大でありうる。

NPCMJ では，自動解析結果に対し人手で修正を加えて得られる統語構造 (句構造) 解析結果を自動意味解析し，論理意味表示 (述語論理式) を得る (Butler 2015)。これが，非有界依存構文を含む複雑な文の解析に利用される。

(2a-c) を例とすると，これらの文の論理意味表示は共通して $\exists xy(\ldots$ 買う$(x,y) \wedge$ りんご$(y) \ldots)$ を含む。述語「買う」の第 2 項が「りんご」の第 1 項と共通であることから，「りんご」が「買う」の直接目的語であることが導かれる。

非有界依存構文における述語の項や複文中の従属節の主語の多くは，表層的統語構造の自動意味解析を通じて同定される。このような場合について，NPCMJ では照応情報を示すためのインデクス付けが不要であり，このことがアノテーションにおける負担軽減をもたらしている。

3　アノテーションの方法

NPCMJ の統語アノテーションの方式は，ペン通時コーパス (Penn Historical Corpora; Santorini 2010) に従っている。文の統語構造はラベル付きの括弧によって表示される。文の中のすべての単語に対して品詞情報を表す品詞タグ (N, ADJI, VB, P 等) がタグ付けされる。また，句に対しては，統語タグ (NP, PP, IP 等) が付加される。CP, IP, NP, および PP については, 通常拡張タグを付加してより細分化されたカテゴリー情報を表示する。ラベル付き括弧表示による統語構造については，下記の図 1A を参照のこと。図 1A は，図 1B の木表示と全く同じ統語情報を表す。

本アノテーションでは X-bar 理論に従い，すべての種類の句に対し同様の比較的フラットな構造を与えている。句のヘッド (N や P 等) が原則的にそれと同一カテゴリーの句 (NP, PP 等) を投射するに際し，両者の間に中間的なカテゴリーを想定せず，修飾語句や補語句と同一レベルの姉妹となる。ヘッドはつねに句の右端にあらわれる (図 2 を参照のこと)。このような統語構造を採用するのは，一つには，木構造の検索や変換を簡単に行うためである。また，これによって，統語構造の埋め込みによるスコープへの干渉を防ぐことができ，柔軟なスコープ包含関係の指定を可能にする。

また，品詞タグや統語タグに対して，必要に応じて拡張タグを付加してその文法機能を表示し，曖昧な統語構造からの正確な意味情報の抽出に利用する。図 1A, 1B に統語構造アノテーションの例を示す。統語タグ PP (助詞句) の後に拡張タグ SBJ や OB1 を付けて主語や直接目的語であることを示し, 日本語における格表示の曖昧性の問題を解決している。品詞タグ P (助詞) には拡張タグ ROLE や OPTR を付加し，それぞれ格助詞および係助詞であることを示す。さらに，文 (節) を表す統語タグである IP に対しては，主節を表す MAT が与えられている。

図 1A：統語アノテーションのラベル付き括弧表示

```
( (IP-MAT (PP-SBJ (NP (NPR (WORD 太郎)))
                  (P-OPTR (WORD は)))
          (PP-OB1 (NP (N (WORD 貴重品)))
                  (P-ROLE (WORD を)))
          (ADVP (ADV (WORD うっかり)))
          (VB (WORD 捨て))
          (P-CONN (WORD て))
          (VB2 (WORD しまっ))
          (AXD (WORD た))
          (PU (WORD 。)))
  (ID 16_textbook_kisonihongo;page_14;JP))
```

図 1B: 統語アノテーションの木構造表示

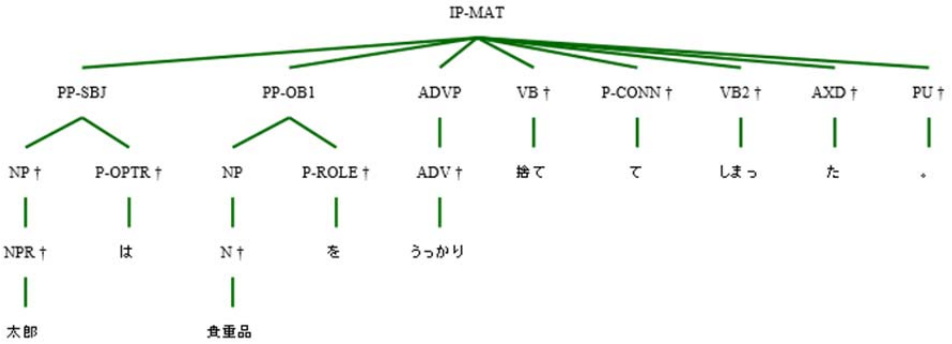
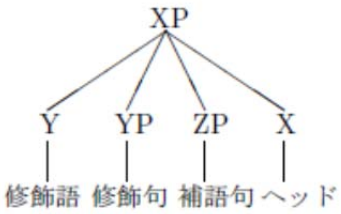


図 2: 統語構造のスキーマ



4 パイプライン

日本語のテクスト (文字の連鎖) を文-句-単語へと分割してそれぞれのレベルのラベリングを行い，それらの間の関係を統語構造として表示するための過程を説明する (Horn et al. 2017, ホーン 2017)。

最初に，テクストは Unidic 辞書 (Den et al. 2008) を使用する形態素解析器 Mecab (Kudo et al. 2014) により解析される。これにより，テクストは単一の形態素から成る短単位に分割され，それぞれに品詞，活用形，語彙素 (レンマ) 等の情報が付与される。これをさらにもう 1 つの形態素解析器 Comainu (Kozawa et al. 2014) に掛け，接辞と語幹や複合語をまとめ，より長い単位である長単位を得るためのチャンキングを行う。この段階で，もう一度品詞等の情報が付与されている。

NPCMJ のセグメンテーションと品詞ラベル付けの方針は，単一の意味のまとまりを成す終端ノードは出来るだけ長く取り，機能的な役割を果たす要素は独立した単語として扱う，というものである。この方針に従って，長単位/短単位のセグメンテーションが取捨選択される。場合により，短単位がさらに分割されることもあるし，複数の長単位をチャンキングすることもある。

こうして得られた，より洗練されたセグメントと品詞ラベルの対を統計的統語解析器に与えることによって，統語解析木を得る。統語解析器としては，手修正済みの NPCMJ

データで学習を行った Berkeley パーザー (Petrov & Klein 2003) を用いている。

以上の機械的処理によって得られた結果は，日本語ネイティブのアノテーターにより修正を受ける。アノテーターによる修正を経た統語解析木は，ダイナミック意味論の一種である Scope Control Theory をインプリメントした意味解析システムによって自動処理され (Butler 2015)，その結果として論理意味表示 (述語論理式) が得られる。

5 アノテーションの特徴

第 3 節のアノテーションの基本的方針を踏まえて，さらに以下の原則を立ててアノテーションを行っている。

第一に，主要文法役割を果たす句に対して，各々の役割を明示する拡張タグを付加する。助詞を伴う句については PP の後に，助詞を伴わない名詞句の場合は NP の後に，SBJ (主語)，OB1 (第 1 目的語—ほぼ直接目的語に相当)，OB2 (第 2 目的語—ほぼ間接目的語に相当) 等の拡張タグを付ける (図 1A, 1B を参照のこと)。

また，いくつかの単語が緊密に連結して 1 つの機能語として働くものは，1 つの助詞 (P) やモーダル助動詞 (MD) として扱っている。単一の複合助詞 (P) としてラベル付けされるものには，

うえに，という，として，に当たって，に関して，に対して，によって，を通じて

等がある。また，単一のモーダル助動詞 (MD) として扱われる連語には

かもしれない，相違ない，違いない

等がある。

関係節に修飾される主名詞が関係節の中で項や付加詞として機能する，いわゆる「内の関係」の関係節の場合は，関係節 (IP-REL とタグ付けされる) 内に空所 (トレース) に相当するノードを与え，さらに NP に対する拡張タグの形で文法役割を明示する。(3a) の統語アノテーションを (3b) に示す。

(3) a. 伊藤さんが好きなピザ

b.
```
( (FRAG (NP (IP-REL (NP-OB1 *T*)
                    (PP-SBJ (NP (NPR (WORD 伊藤さん)))
                            (P-ROLE (WORD が)))
                    (ADJN (WORD 好き))
                    (AX (WORD な)))
            (N (WORD ピザ))))
    (ID 19_textbook_purple_basic;page_30;JP))
```

この例では，関係節 IP-REL の中にトレースに関する情報 (NP-OB1 *T*) が与えられている。トレースと主名詞とが同一であるとの情報は自動意味解析によって得られるので，通常行われているような，両者を関係付けるためのインデクスは不要である。トレースを持たない名詞修飾節 (いわゆる「外の関係」の場合) に対しては，IP-EMB のタグが

与えられる。

さらに，主語または目的語が動詞の主要文法役割として求められるにもかかわらず文中で表現されていない多くの場合について，それらをゼロ代名詞として明示する。次の例文では，第一目的語が省略されていることが (NP-OB1 *pro*) のアノテーションとして示されている。

```
(4)  (IP-MAT (NP-OB1 *pro*)
             (PP-SBJ (NP (WPRO 誰)
                         (P か))
                     (P-CORE が))
             (VB 助ける)
             (MD だろう)
             (PU 。))
```

ただし，複文中の従属節において主語が明示されていなくても，それと同一指示の名詞句が主節中に存在してコントロール関係にある場合はディフォールトとして自動意味解析によりコントロール関係が補完されるので，ゼロ代名詞のタギングは行われない。下の例では，条件節である従属節が主語として主節の主語 = 聞き手を継承している。

```
(5)  ((CP-IMP (IP-SUB (NP-SBJ *hearer*)
                      (PP-CND (IP-ADV (PP-CONJ (IP-ADV (VB (WORD 行く)))
                                               (P-OPTR (WORD にしろ)))
                                      (PU (WORD ， ))
                                      (VB (WORD 行か))
                                      (NEG (WORD ない)))
                              (P-OPTR (WORD にしろ)))
                      (PU (WORD ， ))
                      (NP-OB2 *speaker*)
                      (ADVP (ADV (WORD 後で)))
                      (PP-OB1 (NP (N (WORD 電話)))
                              (P-ROLE (WORD を)))
                      (VB (WORD 下さい)))
            (PU (WORD 。)))
       (ID 366_textbook_djg_advanced;page_418;JP))
```

また，等位関係にある複数の節の間で，前件の左方外部にある構成素をその文法役割も含めて共有することがある (Across the Board 抽出)。このような構文においてもディフォールトとして意味計算により構成素の共有が行われるので，ゼロ代名詞としてはアノテーションを行わない。

上でも述べたように，自動意味解析を利用することで非有界依存のような複雑な構文

でもインデクスのタグ付けが不要であることが NPCMJ の大きな特徴である。これによって，コーパス構築の作業量が著しく軽減される。例外となるのは，外置 (extraposition)，数量詞遊離 (floating quantifier)，主要部内在型関係節 (head-internal relative clause) のいずれかの構文で，意味処理上の要請から，語句を実際に出現する位置以外の場所に関係づける必要が生じる場合である。

6 他のコーパスとの比較

これまでに公表された多くのコーパスでは，文の基本的意味を構成する述語-項構造のアノテーションに重点が置かれてきた。NPCMJ においては主要な文法情報が網羅されており，提供される文法情報は述語-項構造だけに限られない。しかし，比較のために，ここでは述語-項構造だけを取り上げて論じることにする。

日本語の書き言葉に関して述語-項構造をタグ付けしたコーパスとしてはこれまでに，京都大学テキストコーパス (Kurohashi & Nagao 2003) の一部に格情報を付けたもの (Kawahara et al. 2002, 河原他 2002)，NAIST テキストコーパス (Iida et al. 2007, 飯田他 2010)，GDA コーパス (橋田 2005)，および BCCWJ-PAS (小野・飯田 2011) が存在する。これらの概要を NPCMJ のそれとともに表 3 に示す。

京都大学テキストコーパス，NAIST テキストコーパス，BCCWJ-PAS については，各項が伴う格助詞にもとづく述語-項関係のラベル付けを行っている。京都大学テキストコーパスでは出現形をそのまま使用している。他の 2 者では，受動構文や使役構文において格交替が行われる場合，埋め込まれた基本述語の取る格関係を採用するという正規化を行っている。GDA は主題役割という意味情報を提供している点でユニークである。ただし，GDA がタグ付けを行っているのはゼロ代名詞が出現して文外の先行詞を指示する場合に限定されており，すべての述語について述語-項関係を明らかにしたも

表 3：日本語述語–項構造コーパスの比較

	事例数 (文)	述語-項関係	文内の項	ゼロ代名詞項	網羅性
京都大学テキストコーパス	約 5 千	表層格 (出現形)	○	○	×
NAIST テキストコーパス	約 4 万	表層格 (正規化)	○	○	×
BCCWJ-PAS	約 1 万 9 千	表層格 (正規化)	○	○	×
GDA	約 3 万 7 千	主題役割	×	○	×
NPCMJ	約 3 万/6 万	文法役割+ 意味役割	○	○	○

のではない。これに対して，他の 3 つは文内の項も扱っている。事例数の多さやアノテーションの内容を考えると，NPCMJ と比較することが出来るのは NAIST テキストコーパスのみと言ってよい。

NPCMJ は，2019 年 3 月時点で約 3 万文，完成時の 2022 年 3 月で約 6 万文のタグ付けを行う予定であり，全コーパスの中で最大である。述語-項関係については，まず統語情報である文法役割についてタグ付けが行われている。主要文法役割としては，SBJ (主語)，SBJ2 (二重主語構文の第二主語)，LGS (受動文の論理的主語)，OB1 (第一目的語—ほぼ直接目的語に相当)，OB2 (第二目的語—ほぼ間接目的語に相当) がある。日本語においては，上記の受動構文や使役構文における格交替の他にも，他動性の弱い述語の主語のガ/ニの揺れ，可能文や願望文の直接目的語のヲ/ガの交替，連体修飾節内の述語のガ/ノの交替が見られる。また，「は/も」等の係助詞が名詞句に付加されると，その主要文法役割は曖昧になる。このような表層格の違いを超えて，尊敬語化における主語や受動文化における直接目的語等，文法役割の概念が日本語の文法において重要な役割を果たすことが観察されている。このようなことから，述語-項関係は表層格ではなく文法役割として把握することが日本語文法研究者にとっては自然なことである。

NAIST テキストコーパスにおいては，述語-項関係のアノテーションは「が」「を」「に」の 3 つの格助詞に限定されている。これに対して，NPCMJ では，主要文法役割であれば他の格助詞によって表示されるものでもタグ付けが行われている。また，任意文法役割についても，その意味的な性質に応じて，LOC (場所)，TMP (時間)，MSR (時間軸上の範囲または頻度)，ADV (その他の福祉的意味) 等のタグを付加する作業を進めている。

NPCMJ はさらに，論理意味表示における述語の項の順序という形で意味役割に関する情報を提供する。これは，PropBank (Palmer et al. 2010) における ARG0, ARG1 等の数値にもとづく述語項ラベルに相当する。NPCMJ の構築において，自動統語解析結果に人手による修正を加えてタグ付けされるのは表層における文法役割のみである。受動構文や使役構文や関係節において，埋め込まれた基本述語が取る項についての情報は，この段階では欠如している。この統語解析情報が自動意味解析システムに入力されることにより，論理意味表示 (述語論理式) が得られる (Butler 2015)。そこでは，すべての基本述語がそのすべての項についての情報と関係付けられている。文法役割にほぼ同等の意味役割情報が，表層統語情報から自動意味解析によって得られるのである。これにより，述語-項構造付与の人手コストをかなり軽減することができる。

以上に述べたように，NPCMJ はデータ量が多いことと言語学研究上で重要な文法役割を直接タグ付けしていること，また主要文法役割にとどまらず任意文法役割についても分類してタグ付けしている点で，文法研究者にとって現状では唯一無二のコーパスであると言える。

7　コーパスの公開とインタフェース

7.1　コーパスの公開

NPCMJ は，国立国語研究所共同研究プロジェクト「統語・意味コーパスの開発と言語研究」により，2016 年 4 月から開発を開始した。2019 年 3 月の時点で，約 54 万語，約 3 万文のアノテーションをウェブ上 (http://npcmj.ninjal.ac.jp/) で公開している。テクストは，漢字かな交じり，ローマ字の両方で利用可能である。プロジェクトの終了する 2022 年 3 月までに毎年 1 万文を追加し，最終的に 6 万文のアノテーションを公開する予定である。アノテーションを施した言語データは，日本語教科書や文法書の例文，新聞記事，小説，ウィキペディア記事，ノンフィクション，様々な書籍，法律文や聖書に及んでいる。

7.2　ウェブ・インタフェース

ウェブ上で NPCMJ を検索利用するためのツールとして，NPCMJ ウェブサイトの「検索インターフェース一覧」のページから「概要とコンテキスト表示」,「パターン・ブラウザー」,「文字列検索」,「ツリー検索とテキスト解析」,「クエリ作成」を選んで利用できるようになっている。インタフェースは, 日本語, 英語の両方で利用可能である。

なお，全データをダウンロードすることが出来るので，ウェブ・インタフェースを使用せず，ローカルな環境で Tregex 等のツールを使って検索利用することも可能である。

7.2.1　概要とコンテキスト表示

「概要とコンテキスト表示」では，原テクストをそのまとまりごとに書誌情報とともに提示している。それぞれの文の統語解析木を表示することができる。すべての種類のインタフェースを通じて，ディフォールトとしての統語解析木の表示以外に，以下の表示モードを選ぶことができる。

インデクス表示 (indexed)

統語解析情報を自動意味解析することにより, 指示対象である個体やイベントのアイデンティティ情報をインデクスの形で表示する。これにより，複雑な構文やゼロ代名詞においても，述語の項構造が正確に示される。

依存関係表示 (dependency)

主要句 (head) を中心として，他の語句が果たしている意味格役割などの依存関係が示される。

意味表示 (formula)

統語解析情報から自動的に生成された一階述語論理式を表示する。

また，統語解析木の表示は単独の例文に限られず，その前後の複数の文のものを同時に表示できるので，文脈を確認しながら木構造を見ることも可能である。表示された文解析情報は，SVG 画像，ラベル付き括弧形式，XML 形式の 3 つの形式から選んでダウンロードすることが可能である。

7.2.2 パターン・ブラウザー

トップダウン的に事前準備されたメニューの中から選ぶという形式によって, 特定の品詞, 句カテゴリーおよび構文を持つ文を統語解析木とともに表示する。コーパス利用の初心者にとってはもっとも利用しやすいインタフェースの 1 つであるが, その分個々のユーザーの必要に対し柔軟に対応することはできない。

7.2.3 文字列検索

検索したい表現のセグメンテーションが分からない場合や, 単語の連鎖に対してどのようなアノテーションが与えられるかを知りたい場合に文字列検索は効力を発揮する。文字列に関するセグメンテーションの様々な可能性に応じて，Liberal, Character, Strict および Mine の 4 種類のオプションが用意されている。

また，検索表現の内部に別の文字が挿入されている例を検索できる「よくばり検索」も用意されている。この場合，検索表現は複数の単語へとセグメンテーションが行われていてもよい。挿入される文字数はユーザーにより指定される。

7.2.4 ツリー検索とテキスト解析

NPCMJ の本格的利用のためには，統語構造のパターンを入力して行うツリー検索が必要となる。ツリー検索は，テクストボックスに TGrep-lite，あるいは XPath のいずれかを使ったクエリを入力して行う。

TGrep-lite

TGrep-lite は，ツリーバンク中の統語構造を検索するための検索エンジン TGrep 1 (Pito 1993, 1994) にもとづいており，簡単でありながら，オンラインでコーパスを検索するのに適した十分な力を備えている。

TGrep-lite では，ノードおよびノード間の構造上の関係についてパターンを指定して検索を行う。ノード間の関係としては，直接・間接の支配関係，先行/後行関係，姉妹関係，およびそれらの否定が指定される。

XPath

XML で記述したデータベースに対しては効率的に検索を行う技術が確立されており，しかも各ノードに対して従来のように単一の文字列だけを与えるのでなく，レンマなど様々な種類の情報を持たせることが可能になる。そこで本プロジェクトでは，カッコ表示方式により構築したコーパスを XML のマークアップ方式に変換し，さらに必要に応じ様々な情報を付加して拡張していくことにしている。また, 最近になって, XPath (XML ドキュメントの特定部分の指定に用いられる言語で, 木構造に対するクエリとし

て使用される) など，ウェブインタフェースのための一般的な技術が普及しており，さらに統語アノテーションを操作したり検索するためのツールも開発されてきている。XML はこれらの技術やツールとも適合しており，したがって XML マークアップの利用によって本プロジェクトでもそれらにもとづくウェブ･インタフェースの構築が可能になる。

7.2.5 クエリ作成

NPCMJ の精緻なアノテーションを完全に生かすためには抽象的な検索パターンを駆使して検索利用することが必要である。しかし，そのためには形式統語理論等の言語の数理的側面に関する知識が不可欠であり，言語研究者の大部分にとっては敷居が高い。形式言語理論にも言語処理技術にも習熟しておらず，また当該のツリーバンクのアノテーションの詳細にも通じていないユーザーが統語パターンを使ってツリーバンクを検索することを可能にするウェブ上のユーザーフレンドリーな検索エンジンとして，Query Builder というインターフェースを提供している。ここでは，上記の 3 つのインターフェースを使って得られた文を元に，クエリが作成できる。クエリ作成は，ICECUP における Fuzzy Tree Fragment 機能 (Nelson, Wallis & Aarts 2002) に倣ったインターフェースで，既成の構造から得られる特徴を自由に選択したクエリの作成を支援する。文を自分自身で作例し，自動解析にかけて，その結果を出発点とすることもできる。これはルーヴァン大学の GrETEL (Augustinus 2016) と類似の機能である。

8 さらなる発展

竹内他 (2019) では，日本語の述語に対し概念フレームおよび「動作主」「経験者」「対象」等の意味役割を付与した述語項構造シソーラスを構築しているが，さらにその拡張として，Arg0, Arg1 のように数値化された意味役割を付与する PropBank (Palmer et al. 2010) の方式での意味役割も追加している。これにより，PropBank の一般性・抽象性と具体的な意味役割による個別的な概念識別の両方の長所を兼ね備えた意味役割情報の把握が出来るようになる。さらに，この方式で NPCMJ に対し概念フレームと意味役割を与える作業を進めている。これにより，NPCMJ のデータについて，より意味的側面に立ち入った利用が可能になる。また，よりデータ量が多く高度の意味情報を持つ述語項構造シソーラス開発への道を開く。

9 おわりに

これまで統語論や意味論等，文法理論に関わる研究は主として作例や少数のデータにもとづいて行われてきた。そのため，どこまで現実の言語使用を反映するのかという問題がつねにつきまとった。コーパスを利用した言語研究は，語彙や形態論を中心に行われてきた。アノテーションが不十分なことから，言語の理論文法的側面に関心を持つ研

究者にとっては利用価値が乏しかった。NPCMJ を検索利用することで，構文等，特定の統語的条件を満たす例文を短時間で集めることができる。その際，試行錯誤を頻繁に行えることから，研究者の関心を満たす言語データを得られる確実性が高くなる。NPCMJ は，従来別個に行われてきた，言語の質的研究と量的研究とを統合するものであると言うことができる。大量の言語データから正確な文法情報が得られやすくなることにより，これまでにない説得力を持つ日本語の理論研究や他言語との対照研究が進展することを期待している。

参考文献

Augustinus, Liesbeth. (2016) About GrETEL. http://Gretel.ccl.kuleuven.be/project/

Butler, Alastair (2015) *Linguistic Expressions and Semantic Processing: A Practical Approach*. Springer.

バトラー-アラステア・吉本啓・岸本秀樹・プラシャント-パルデシ (2016)「統語・意味解析情報付き日本語コーパスのアノテーション」，『言語処理学会第 22 回年次大会発表論文集』, pp. 589-592, 東北大学.

Den, Yasuharu, Junpei Nakamura, Toshinobu Ogiso & Hideki Ogura. (2008) “A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. ” *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (*LREC 2008*), pages 1019--1024. Marrakech, Morocco: European Language Resources Association (ELRA).

橋田浩一 (2005) 「GDA 日本語アノテーションマニュアル」

橋本進吉 (1934) 『国語法要説』岩波書店.

ホーン, スティーブン-ライト (2017) 「アノテーション方式とインタフェース」, パネルセッション「統語・意味解析情報をタグ付けした日本語コーパスの開発―アノテーションの方法と文法研究への応用―」,『日本語文法学会第 18 回大会発表予稿集』, 日本語文法学会.

Horn, Stephen Wright, Alastair Butler & 吉本啓 (2017) “Keyaki Treebank Segmentation and Part-of-Speech Labelling,”『言語処理学会第 23 回年次大会発表論文集』, pp. 414-417, 筑波大学.

Iida, Ryu, Mamoru Komachi, Kentaro Inui & Yuji Matsumoto. (2007) “Annotating a Japanese text corpus with predicate-argument and coreference relations,” *Proceedings of the Linguistic Annotation Workshop*, Association for Computational Linguistics, pp. 132-139.

飯田龍・小町守・井之上直也・乾健太郎・松本裕治 (2010) 「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』第 17 巻, 第 2 号, pp. 25-50.

河原大輔・黒橋禎夫・橋田浩一 (2002) 「『関係』タグ付きコーパスの作成」『言語処理学会第 8 回年次大会発表論文集』, pp. 495-498.

Kawahara, Daisuke, Sadao Kurohashi & Koiti Hasida. (2002) “Construction of a Japanese relevance tagged corpus,” LREC.

小町守・飯田龍 (2011) 「BCCWJ に対する述語項構造と照応関係のアノテーション」日本語コーパス平成 22 年度公開ワークショップ.

Kozawa, Shunsuke, Uchimoto Kiyotaka, & Yasuharu Den. (2014) 「BCCWJ に基づく長単位解析ツール Comainu」『言語処理学会第 20 回年次大会発表論文集』, pp. 582-585, 京都大学.

Kudo, Taku, Kaoru Yamamoto, & Yuji Matsumoto. (2004) “Applying conditional random fields to Japanese morphological analysis,” *Proc. of EMNLP*, pages 230-237.

Kurohashi, Sadao & Makoto Nagao. (2003) “Building a Japanese parsed corpus – while improving the parsing system,” A. Abeille. (ed.) *Treebanks: Building and Using Parsed Corpora.* Kluwer.

Levy, Roger, & Galen Andrew. (2006) “Tregex and tsurgeon: tools for querying and manipulating tree data structure,” 5th International Conference on Language Resources and Evaluation.

Maekawa, Kikuo, et al. (2014) “Balanced corpus of contemporary written Japanese,” *Language Resources and Evaluation* 48(2).

三上章 (1970) 『文法小論集』くろしお出版.

Nelson, Gerald, Wallis, Sean & Aarts, Bas. (2002) *Exploring Natural Language: Working with the British Component of the International Corpus of English.* John Benjamins.

Palmer, Martha, Daniel Gildea, & Nianwen Xue. (2010) *Semantic Role Labeling.* Morgan & Claypool Publishers.

Petrov, Slav & Dan Klein. (2007) “Improved inference for unlexicalized parsing,” *Proceedings of NAACL HLT 2007*, pp. 404-411.

Pito, Richard. (1993, 1994) TGrep (1). Copyright under grant from the Benjamin Franklin Institute.

Santorini, Beatrice. (2010) Annotation Manual for the Penn Historical Corpora and the PCEEC (Release 2). Tech. rep., Dep. of Computer and Information Science, University of Pennsylvania.

竹内孔一・アラステア-バトラー・長崎郁・スティーブン-ライト-ホーン (2019) 「PropBank スタイルの意味役割タグを導入した述語項構造シソーラスと NPCMJ への付与計画」『言語処理学会第 25 回年次大会発表論文集』, 名古屋大学.

van Noord, Gertjan et al. (2013) “Large Scale Syntactic Annotation of Written Dutch: Lassy,” Spyns, Peter, & Odijk, Jan. (eds.) *Essential Speech and Language Technology for*

Dutch: Resources, Tools and Applications. Springer.
吉本啓 (2018)「言語研究と統語・意味解析情報付きコーパス」日本英語学会第 36 回大会シンポジウム「ツリーバンク開発と言語理論」, *Conference Handbook* 36, pp. 242-247, 横浜国立大学, 横浜市.

Construction of a Japanese Corpus with Syntactic and Semantic Annotations

Prashant Pardeshi
NINJAL

Kei Yoshimoto
Tohoku University

Abstract

This paper introduces and discusses NINJAL Parsed Corpus of Modern Japanese (NPCMJ), the first full-fledged publicly available corpus for Japanese based on phrase structure, which is being built by a research project at National Institute for Japanese Language and Linguistics (Butler et al. 2016). The currently available corpora in Japan, such as Kyoto University Text Corpus (KTC; Kurohashi and Nagao 2003) and Balanced Corpus of Contemporary Written Japanese (BCCWJ; Maekawa et al. 2014), parse sentences into the phonological unit bunsetsu (accentual phrase) as opposed to adopting phrase structure that reflects building blocks of sentence meanings. Accordingly, they fall short of providing information that is useful to those interested in various grammatical phenomena. With elaborate syntactic and semantic information in NPCMJ, those engaged in research and education of the Japanese language can obtain linguistic data that match expressions they are interested in. Furthermore, it enables dynamic extraction and use of grammatical information according to the need and kind of data being studied.

In this paper, we first discuss the motivation of our corpus development. Following that, we give an account of our annotation policy and the corpus building procedure. Then, an illustration is given about the essential features of NPCMJ, especially concerning predicate-argument structure in comparison with other corpora for Japanese. We also give an overview of the interfaces available at our website. Then, we explain further development of our project attained by cooperation with related research areas. Lastly, our perspective on long-range significance of NPCMJ in the study of Japanese is discussed.