Construction and Analysis of Information-Structure Annotation of the "Balanced Corpus of Contemporary Written Japanese"

国立国語研究所論集 (NINJAL Research Papers) 16: 19–33 (2018)
ISSN: 2186-134X print/2186-1358 online

『現代日本語書き言葉均衡コーパス』への情報構造アノテーションとその分析

宮内拓也[a]　浅原正幸[b]　中川奈津子[c]　加藤　祥[d]

[a] 東京外国語大学大学院 博士後期課程／日本学術振興会 特別研究員／
国立国語研究所 共同研究員
[b] 国立国語研究所 コーパス開発センター
[c] 千葉大学人文科学研究院 特任研究員
[d] 国立国語研究所 コーパス開発センター 非常勤研究員

要旨

本稿では,『現代日本語書き言葉均衡コーパス』のテキスト（新聞（PN）コアデータ 16 サンプル）内の名詞句に対し, 情報構造に関係する文法情報のラベル（情報状態, 共有性, 定性, 特定性, 有生性, 有情性，動作主性）をアノテーションした結果を報告する。特に，本稿ではアノテーションの概要と基礎統計について述べる。ラベル間の対応を Kappa 値で評価した結果，先行研究で既にアノテーションされていた共参照情報を基にした情報状態と定性･特定性の間には中程度の一致（0.41 以上）が見られたのに対し，今回新たに付与した共有性と定性･特定性の間にはほとんど完璧な一致（0.81 以上）が見られた。冠詞選択に大きな影響を与える定性・特定性のアノテーションは，定性・特定性が話し手側により踏み込んだ概念であることから複雑で難度が高いため，他の文法情報で定性・特定性を推定する方がより容易であると考えられる。評価の結果は，定性・特定性の推定には，共参照情報を基にした情報状態だけでは十分でなく，聞き手／読み手の観点を考慮した共有性が重要であることを意味している。また，日本語では助詞「は」と「が」の使い分けについて，情報構造との関連が指摘されているが，付属語主辞とのラベルの関係を見ると，「が」「を」「に」は新情報が多く，「は」は若干旧情報が多いこと，「は」「の」に定性・特定のものが多く，「を」に不定・不特定のものが多いことがわかった *。

キーワード：情報構造，アノテーション，名詞句，冠詞選択，助詞

* 本稿の一部は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」（プロジェクトリーダー：浅原正幸）の研究成果である。加えて，本研究は JSPS 科研費「言語コーパスに対する読文時間付与とその利用」（課題番号：25284083，研究代表者：浅原正幸），「日本語歴史コーパスに対する統語・意味情報アノテーション」（課題番号：17H00917，研究代表者：浅原正幸），「顕在的な冠詞がない言語における名詞句の統語構造と意味解釈の研究」（課題番号：17J07534，研究代表者：宮内拓也）の助成を受けている。また，本稿の内容は以下の口頭発表に基づき，その内容を修正，増補したものである：

・「『現代日本語書き言葉均衡コーパス』への情報構造アノテーションの分析」，言語資源活用ワークショップ 2016，2017 年 3 月 8 日，国立国語研究所。
・「『現代日本語書き言葉均衡コーパス』への情報構造アノテーションの構築」, 言語処理学会第 23 回年次大会，2017 年 3 月 15 日，筑波大学。
・“Information-Structure Annotation of the “Balanced Corpus of Contemporary Written Japanese”,” The 15th International Conference of the Pacific Association for Computational Linguistics (PACLING 2017), 16 Aug. 2017, Sedona Hotel Yangon (Myanmar).

1. はじめに

本稿では，『現代日本語書き言葉均衡コーパス』（以下，BCCWJ; Maekawa et al. 2014）のテキスト内の名詞句に対して情報構造に関わる文法情報のアノテーションを行った結果を報告する。

日本語の情報構造に関する過去のアノテーションは，主としてテキスト中に出現する情報が談話中に既出であること，つまり情報状態（information status）を共参照情報として付与するもの（植田他 2015，浅原・大村 2016）であった。本研究は，情報状態のみならず，情報状態と深く関連のある定性，特定性，またトピックやフォーカスといった情報構造上の概念と深く関連のある有生性や動作主性などの項目でラベルを付与した。本稿で示す情報構造関連の文法情報は，冠詞のある言語では冠詞という文法カテゴリーによって表示され得るが，日本語は冠詞のない言語であるため，それらの情報は音声，形態的に発現しない。本稿では，定性，特定性などの顕在化しない文法情報間の分布を調べるため，ラベル間の対応を Kappa 値で評価した。ただし，日本語では助詞「は」と「が」の使い分けについて，情報構造との関連が指摘されている（三上 1963，野田 1996，庵 2010 など）。そのため，本稿では各ラベルとアノテーションされた名詞句が含まれる文節の付属語主辞の分布についても分析を行った。

本稿で示すアノテーションの工学的応用としては，機械翻訳における冠詞の推定等が考えられる。冠詞がない言語を母語とする者にとって，冠詞がある言語を習得する際の冠詞選択は難しいものである（Ionin et al. 2004, Tanaka 2013）。冠詞選択には，一般に定性（definiteness）や特定性（specificity）などの情報構造に関係する文法情報が大きな影響を与える。英語のように定性により定冠詞と不定冠詞を使い分ける言語もあれば，サモア語のように特定性により冠詞を使い分ける言語もある（Mosel and Hovdhaugen 1992）。さらには，コヴェ語[1]のように定性と特定性が共に冠詞選択に影響を与える言語もある（Sato 2013）。言語処理の分野では英語母語話者が産出した大量のテキストから，英語学習者の冠詞の誤りを検出する手法が提案されている（Nagata et al. 2005）。しかし，日本語母語話者が産出する他言語の冠詞選択を検討する場合，日本語における名詞句の情報構造を考慮する必要がある。さらに，機械翻訳において日本語文を冠詞のある言語に訳す際にも，日本語の情報構造が問題となってくる。加えて，本稿で示した情報構造アノテーションを視線計測のデータ（Asahara et al. 2016）と重ね合わせることにより，情報構造に対する読み時間のふるまい（Asahara 2017）について調査を行うことができ，情報呈示手法とリーダビリティの研究に資する。

以下，2 節では情報構造アノテーションに関する先行研究を示す。3 節は我々が提案するラベルとアノテーション基準について例を示しながら紹介する。4 節に基礎統計を示すとともに，各ラベル間の関係，付属語主辞との共起傾向を検証する。5 節にまとめと今後の課題を示す。

2. 関連研究

情報構造のアノテーションには，当該言語形式の情報構造をどのように決定するかという点で，

[1] コヴェ語は，パプアニューギニアのニューブリテン島で話されるオーストロネシア語族の言語の一つである。

二つのタイプがある。まず初めに，当該言語形式そのものに基づいて情報構造を決める研究がある。例えば, Calhoun et al.（2005）は, Vallduví and Vilkuna（1998）や Steedman（2000）に言及し，韻律を採用した。L+H*LH%[2] の韻律を持つ形式はテーマ（theme）[3] となり，H*L，H*LL%[4] の韻律を持つものはレーマ（rheme）[5] となる。そして, 彼らは当該の名詞句が以前に言及されたか否か，またそれが以前述べられた個体から言及可能か否かという点をもとに情報構造をアノテーションした。Hajičová et al.（2000）は語順を用いた情報構造のアノテーションを提案した。この研究は情報構造についてプラハ学派の伝統に触発されたものであり，それゆえに動詞より左にある言語形をトピックとする。これらのアノテーション基準は言語依存であり，日本語に適用可能なものではない[6]。

二つ目のタイプの研究では，言語学的なテストを採用する。Götze et al.（2007）は言語に依存せず, かつ特定の言語理論にもよらず, 情報状態[7] とトピック[8], フォーカス[9] をアノテーションするための基準を策定した。例えば，アバウトネストピックは以下（1）の手続きで決定される。

（1）　An NP X is the aboutness topic of a sentence S containing X if
「ある名詞句 X が, これを含む文 S のアバウトネストピックであるのは以下のときである：」

a. S would be a natural continuation to the announcement *Let me tell you something about X*
「『X について話させて』という予告の後に，S が自然に続き得るとき。」

b. S would be a good answer to the question *What about X?*
「S が，『X についてはどう？』という質問にふさわしい答えであるとき。」

c. S could be naturally transformed into the sentence *Concerning X, S'*, where S' differs from S only insofar as X has been replaced by a suitable pronoun.
「S が『X に関しては，S'』に自然に変形できるとき。ただし，S' は S における X が適切な代名詞に置き換わっている点のみで異なる。」　（Götze et al. 2007: 165）

本研究は Götze et al.（2007）に沿うものであるが, いくつかの点で彼らの研究とは大きく異なっている。まず，本研究ではトピックとフォーカスを直接アノテーションしない。これはトピック

[2] L+H*LH% の韻律では，低めに上昇した後，発話境界の上昇調が現れる（Steedman 2000: 654–655）。
[3] 概してトピック（topic）に対応する。
[4] H*L の韻律では，急速な高いピッチから始まりそして下がる。H*LL% の韻律は，H*L に発話末の低い音調が伴ったものである（Steedman 2000: 654–655）。
[5] 概してフォーカス（focus）に対応する。
[6] なお，Calhoun et al.（2005）は英語を，Hajičová et al.（2000）はチェコ語を対象にしている。
[7] 情報状態は，旧情報（given）/ 補完可能（accessible）/ 新情報（new）の三つが区別される。ある表現が，それ以前の談話で明示的に言及されている先行詞があれば旧情報であり，関連のあるものが言及されていれば補完可能であり，言及されていなければ新情報である（Götze et al. 2007: §3.2）。
[8] トピックについては，アバウトネストピック（aboutness topic）/ フレームセッティングトピック（frame setting topic）の二つが区別される。アバウトネストピックは（1）で説明する。フレームセッティングトピックとは，そのトピックを含む文の述語が解釈されるべきフレームであり，典型的には時間や場所の表現がこれにあたる（Götze et al. 2007: §4.2.3）。
[9] フォーカスに関しては，新情報フォーカス（new-information focus）/ 対比フォーカス（contrastive focus）の二つが区別される。新情報フォーカスとは，談話を進めるために欠けている新しい情報を提供する文の要素である。対比フォーカスとは，他の発話に対して対比を呼び起こす文の要素である（Götze et al. 2007: §5.2）。

やフォーカスがそれぞれに多次元であるためである（Nakagawa 2016）。実際に Götze et al.（2007: 163）では，例えば，指示的（referential）な名詞句，特定（specific）解釈や総称（generic）解釈を持つ不定（indefinite）の名詞句など，様々な種類のアバウトネストピックが区別されている。定性や特定性のような要因はトピックとは独立であると考え，そのようにアノテーションする方がより単純である。第 2 にトピックやフォーカスと相関すると知られている要因については，定性や特定性以外にも例えば，有生性（animacy）や動作主性（agentivity）（Givón 1976, Keenan 1976）など，より多くが考えられる。そのため，本研究では情報構造アノテーションの一環としてこれらについてもラベルを付与することとする。

3. ラベルとアノテーション基準

BCCWJ では，長単位と短単位という二つの単位が採用されているが，本研究では，短単位の名詞をアノテーション対象とする。ただし，複合語については，前部要素には指示性（referentiality）がないこと等を考慮して，前部要素まで含めて一つの名詞と捉える [10]。

本研究では，以下の（2）で示す項目についてラベルを設定した。

（2）
- a. 情報状態（information status）：「新情報（discourse-new）」/「旧情報（discourse-old）」
- b. 共有性（commonness）：「共有（hearer-old）」/「非共有（hearer-new）」/「想定可能」
- c. 定性（definiteness）：「定（definite）」/「不定（indefinite）」
- d. 特定性（specificity）：「特定（specific）」/「不特定（unspecific）」
- e. 有生性（animacy）：「有生（animate）」/「無生（inanimate）」
- f. 有情性（sentience）：「有情（sentient）」/「非情（insentient）」
- g. 動作主性（agentivity）：「動作主（agent）」/「被動作主（patient /theme）」

上に示したラベルは言語学の専門的な知識を持つものでないとアノテーションできないこと，研究の目的がわかっているものがアノテーションを行うのが効率が良いことを考慮して，本研究では著者の内の 1 人がアノテーションを行った。なお，基準については共著者間で相談しながら策定した。アノテータは BCCWJ-DepParaPAS（植田他 2015，浅原・大村 2016）に付与された共参照情報を確認しながら作業を行う。定性，特定性，有生性，有情性，動作主性 [11] については，与えられた文脈から判断できない場合に「どちらでもよい」というラベルを認めた。特定性，動作主性，共有性については，その概念が認めがたい場合に「どちらでもない」というラベルを認めた。

以下，実例と共にそれぞれのラベルのアノテーション基準を示す。

[10] つまり，短単位では 2 語以上の扱いを受けるものに関しても，1 語扱いになっている場合があることになる。これは BCCWJ への共参照アノテーションと同様の方針である。
[11] ただし，3.5 で見るように，動作主性については，「どちらでもよい」のラベルは主節から見た場合と従属節から見た場合で動作主性の値が異なる場合に付与される。

3.1 情報状態

（2a）の情報状態とは，いわゆる旧情報と新情報の区別である。ある談話において，新たな情報は「新情報」となり，聞き手 / 読み手が知っている情報は「旧情報」となる [12]。一つのテクスト全体を一つの談話と見なし [13]，アノテーションを行った。

（3）　a.　担任だった池田弘子先生は違った。
　　b.　スクールカウンセラーでもあった先生の授業は　（読売新聞 [BCCWJ: PN1c_00001]）

（3a）の下線部の名詞「池田弘子」はこのテクストで初出の名詞であるため，新情報ラベルが付与される。一方，（3b）の下線部の名詞「先生」は（3a）の「池田弘子」を指示しているため旧情報ラベルが付与される。これらの名詞は共参照関係にある。

3.2 共有性

（2b）の共有性は，情報を聞き手 / 読み手が既に知っていると話し手が想定しているか否かを示すパラメータである。聞き手 / 読み手が既に知っていると話し手が想定している情報は「共有」であり，知らないと想定している情報は「非共有」である。なお，この判断の際はアノテータの世界知識を使ってもよいこととし，「想定可能」というラベルも許す。このラベルは，ブリッジング（bridging）を起こしている際に付与される。

（4）　a キャンティ街道を抜け、b オリーブ畑に囲まれた田園地帯の c レストランで、
（読売新聞 [BCCWJ: PN4c_00001]）

（4）の下線部 a の名詞「キャンティ街道」は，世界遺産にも登録されている，ワインで有名な街道であり，アノテータは既にこの街道について知っていたため，共有のラベルが付与された。下線部 b の名詞「オリーブ畑」はこの記事からどんなオリーブ畑であるのか判断できないため，非共有のラベルが与えられる。下線部 c の名詞「レストラン」はキャンティ街道のレストランを指しており，ある種のブリッジングを起こしているため，想定可能のラベルが付与される。

3.3 定性，特定性

（2c）の定性とは，指示対象を聞き手 / 読み手が同定できるか否かを示すカテゴリーである [14]。指示対象を聞き手 / 読み手が同定できると話し手が想定していれば「定」であり，同定できないと想定していれば「不定」である。本研究では，スコープとして前後 3 文を見ることとする。

[12] 情報構造自体についての詳細は Kruijff-Korbayová and Steedman（2003）や Hinterwimmer（2011）などを参照のこと。
[13] 本研究では，BCCWJ の新聞（PN）のデータを用いたため，一つの記事が一つの談話であると見なしている。
[14] 定性そのものについては，Lyons（1999），Heim（2011）などを参照のこと。

(5)　高等部では自由な校風もあって、流行に乗ってかばんを薄くつぶしたり、ピアスをしたり。呼び出して注意する先生もいたが、二、三年時に担任だった池田弘子先生（七十五）は違った。「そんな薄い [a]かばんじゃ [b]遊び道具も入らないよ」
（読売新聞 [BCCWJ: PN1c_00001]）

(5) の下線部 a の名詞「かばん」はスコープである (5) の前 3 文以内に既出の名詞であり，ここでは具体的に聞き手 / 読み手の持ち物のかばんを指示している。話し手はこの「かばん」は聞き手 / 読み手により同定し得ると想定していると考えられるため，定のラベルが与えられる。(5) の下線部 b の名詞「遊び道具」は特に具体的な何らかの遊び道具を指示しているわけではないため，不定のラベルが付与される。

(2d) の特定性は，定性と少々似た概念であるが，話し手が特定の事物を想定しているか否かを示す意味論的カテゴリーである [15]。話し手が特定の事物を想定しているならば「特定」となり，想定していなければ「不特定」となる。定性と同様，特定性に関してもスコープとして前後 3 文を見ることとする。

(6)　米どころの同町では、降霜対策で農家による廃タイヤの野焼きが行われてきたが、ダイオキシン問題や交通妨害が指摘され、行き場を失った [a]廃タイヤがあぜ道や [b]納屋の横に放置されてきた。同町が昨秋行った調査では、廃タイヤは農家が抱えるものや不法投棄を含め約三万本に上るという。（北海道新聞 [BCCWJ: PN2e_00001]）

(6) の下線部 a の名詞「廃タイヤ」は，北海道鷹栖町に放置された約 30,000 本のタイヤを具体的に指しており，これは (6) の前後 3 文から読み取ることが可能であるため特定のラベルが付与される。(6) の下線部 b の名詞「納屋」は特定の納屋が想定されているわけではなく，不特定のラベルが与えられる。

3.4 有生性，有情性

(2e) の有生性とは，生きているか否かを示すカテゴリーである。有生性は情報構造上重要な概念であるトピックやフォーカスと相関すると知られている（Givón 1976, Keenan 1976 など）。生物（人間，動物など）は「有生」であり，無生物（植物を含む）は「無生」である。有生性は名詞句レベルのみで判断し，付与されるものとする。有生性と似た概念として (2f) の有情性がある。これは，情意があるか否かを示すパラメータである。自由意志による移動が可能な場合は「有情」となり，自由意志による移動がないなら「非情」となる。日本語については，有生 / 無性の区別よりも有情 / 非情の区別が重要であるとする先行研究もあり（山口 1985 など），また，有生性と有情性の値が異なる場合もあり得る [16] ことから，このパラメータの設定が必要となる。

[15] 特定性そのものに関しては，von Heusinger (2011) などを参照のこと。
[16] 例えば，ゾンビと幽霊は有生性と有情性の値の差により区別できる。ゾンビは腐敗した人体が自発的意思なく徘徊するため，有生，非情である。一方，幽霊とは死者の魂が未練や遺恨により現れたものであるため，無生，有情となる。

情意の有無は名詞句単体では判定できない場合があるため，有情性は述語－項レベルまで見た上で判断し，付与されるものとする。

(7)　オオクチバスなどの [a]ブラックバス類が、少なくとも四十三都道府県の七百六十一のため池や [b]湖沼に侵入し、　（読売新聞 [BCCWJ: PN4c_00001]）

(7) の下線部 a の名詞「ブラックバス」は生物であるため，有生のラベルが付与される。また，ブラックバスに情意があるか否かは判断が難しいが，その述語は「侵入する」となっており，これは意志的な動作，行為を表しているため，ここでの「ブラックバス」は有情のラベルが付与されることになる。(7) の下線部 b の名詞「湖沼」は無生物であり，情意もないと判断されるため，それぞれ，無生，非情のラベルが与えられる。

3.5 動作主性

(2g) の動作主性は，事態に関わる事物や人物がその事態で果たしている役割を示す。動作主性も情報構造上重要な概念であるトピックやフォーカスと相関すると知られている（Givón 1976, Keenan 1976 など）。行為を意図的に実現するものは「動作主」とし，行為によって変化を被るものを「被動作主」とする。このパラメータについては節レベルまで見て判断し，ラベルを付与することとする。その際，主節と従属節の両方を考慮する。また，「どちらでもよい」「どちらでもない」を許す。

(8)　a.　編み笠をかぶった人なつっこい笑顔を見るだけで、
　　b.　もみじの木にとまって仲良く寄り添う二羽のキジバト。
　　c.　独特な雰囲気の写真になりました。　（産経新聞 [BCCWJ: PN1d_00001]）

(8a) の下線部の名詞「笑顔」は，主節では被動作主であり従属節では動作主であると解釈できる。このような場合に「どちらでもよい」というラベルを付与する。(8b) の下線部の名詞「キジバト」は，それを含む文がこの名詞で終わる体言止めの文であるため主節では動作主性の判断ができないが，従属節では動作主であるため，「動作主」というラベルを付与する。(8c) の下線部の名詞「写真」は動作主でも被動作主でもないため，「どちらでもない」となる。

3.6 その他

固有名詞については，アノテーションの際，有名の度合いを考慮してよいこととし，アノテータの持つ世界知識を参照してもよいとする。(9a) の形式名詞や (9b) のような慣用表現は対象から外し，それぞれ「形式名詞」，「慣用表現」ラベルを付与する。なお，慣用表現であるか否かについてはアノテータによる揺れを許すこととする。

(9)　a.　様々な人がいるということが
　　b.　聞く耳を持たせてくれるんです。　（読売新聞 [BCCWJ: PN1c_00001]）

4. 基礎統計

4.1 ラベル間の関係

対象は BCCWJ の新聞（PN）コアデータ 16 サンプルに出現する名詞 2,023 件とした。サンプルの選択は BCCWJ-ANNOTATION-ORDER[17] に基づく。対象の基本的なデータとして，表 1 に対象となるテキストの総語数（短単位数），文節数，文数，名詞数を示す。また，BCCWJ-ANNOTATION-ORDER によるサンプルの選択とサンプル名の関係を示すため，アノテーション順も記す。

表 1　対象となるテキストのデータ

アノテーション順	サンプル名	短単位数	文節数	文数	名詞数
1	PN1c_00001	784	236	42	72
2	PN1d_00001	783	235	34	70
3	PN1e_00001	763	219	35	85
4	PN1f_00001	797	181	38	72
5	PN2e_00001	750	214	27	85
6	PN3b_00001	975	311	43	120
7	PN3g_00001	2,640	919	142	373
8	PN4a_00001	1,244	425	51	148
9	PN4b_00001	758	246	26	103
10	PN4c_00001	737	250	31	88
11	PN4f_00001	1,047	297	40	115
12	PN4g_00001	905	296	36	158
13	PN1a_00002	1,797	611	93	207
14	PN1b_00002	1,024	277	38	95
15	PN1d_00002	734	206	28	74
16	PN1e_00002	919	272	35	158
合計		16,657	5,195	739	2,023

表 2 にラベルの基礎統計を示す。情報状態のラベルは以前アノテーションされた共参照情報に基づいているが，情報状態と他の分布は異なっている。ゆえに，この差異は日本語からの翻訳の際の冠詞選択に影響を与えると考えられる。

[17] BCCWJ コアデータサンプルにおけるアノテーション優先順序である。

表 2　ラベルの基礎統計

情報状態	新情報	旧情報	—	—
	1,345	678	—	—
共有性	共有	非共有	想定可能	どちらでもない
	1,036	494	489	4
定性	定	不定	どちらでもよい	—
	1,122	899	2	—
特定性	特定	不特定	どちらでもよい	どちらでもない
	1,157	749	116	1
有生性	有生	無生	どちらでもよい	—
	342	1,680	1	—
有情性	有情	非情	どちらでもよい	—
	337	1,678	8	—
動作主性	動作主	被動作主	どちらでもよい	どちらでもない
	192	338	2	1,491

まず，共参照情報に基づく情報状態と定性・特定性の関係について示す。表 3 に情報状態と定性の分割表を示す。不定のラベルは新情報のラベルと共に現れることが多いが，定のラベルは新情報，旧情報のどちらのラベルとも現れ得るという傾向がある。これは共参照情報の冠詞選択への貢献が限定的であることを示しているといえる。「新情報↔不定」と「旧情報↔定」の対応の Kappa 係数は 0.47 と中程度の一致度であった。

表 3　情報状態と定性

	新情報	旧情報
定	497	625
不定	846	53
どちらでもよい	2	0

なお，Kappa 係数とは，データ間の一致度を評価する指標である。実測値と期待値の比較によって，偶然による一致の可能性を排除した上で算出される（Cohen 1960）。データの判定が実際に一致した割合を P_o とし，データ間の独立を仮定した上で偶然に一致が期待される割合を P_e とすると，Kappa 係数 κ は次式の通りに表される。

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

（Cohen 1960: 40）

これからわかるように，Kappa 係数は -1 から 1 の値を取る。負の値であれば偶然より一致度が低いことになり，0 であればデータ間に偶然の一致しかないことになり，0.81 以上であればほぼ

完全に一致しているということになる[18]。つまり，値が大きければ大きいほど一致度が高いと判定できる。

表 4 は情報状態と特定性の分割表である。これについても情報状態と定性のものと同様の分布を示している。「新情報↔不特定」と「旧情報↔特定」の対応の Kappa 係数は 0.43 と中程度の一致度であった。

表 4　情報状態と特定性

	新情報	旧情報
特定	531	626
不特定	705	44
どちらでもよい	108	8
どちらでもない	1	0

表 5 に定性と特定性の分割表を示す。不特定のものはすべて不定である。言い換えると，定であって不特定のものは存在しなかった。

表 5　定性と特定性

	定	不定	どちらでもよい
特定	1,120	36	1
不特定	0	749	0
どちらでもよい	2	113	1
どちらでもない	0	1	0

次に，聞き手 / 読み手における情報構造である共有性と定性・特定性の関係について示す。表 6 に情報状態と共有性の分割表を示す。談話上の情報状態が新情報である名詞句は，共有･非共有･想定可能に対して一様に分布している。

表 6　情報状態と共有性

	新情報	旧情報
共有	425	611
非共有	460	34
想定可能	456	33
どちらでもない	4	0

[18] Landis and Koch（1977: 165）では，Kappa 値を評価する基準として以下のものが挙げられている。

Kappa Statistic（Kappa 統計量）	Strength of Agreement（一致度の強さ）
< 0.00	Poor（低い一致）
0.00–0.20	Slight（ごく軽度の一致）
0.21–0.40	Fair（軽度の一致）
0.41–0.60	Moderate（中程度の一致）
0.61–0.80	Substantial（高度の一致）
0.81–1.00	Almost Perfect（ほとんど完璧な一致）

表 7 に共有性と定性の分割表を，表 8 に共有性と特定性の分割表を示す。非共有・想定可能を一つのカテゴリーとした場合に，「共有↔不定」と「非共有・想定可能↔定」の対応の Kappa 係数は 0.86 とほとんど完璧な一致を示した。「共有↔不特定」と「非共有・想定可能↔特定」の対応の Kappa 係数も 0.81 とほとんど完璧な一致を示した。

表 7　定性と共有性

	定	不定	どちらでもよい
共有	1,010	26	0
非共有	74	420	0
想定可能	37	450	2
どちらでもない	1	3	0

表 8　特定性と共有性

	特定	不特定	どちらでもよい	どちらでもない
共有	1,008	26	5	0
非共有	91	391	11	1
想定可能	57	332	100	0
どちらでもない	1	3	0	0

このことから，冠詞推定に必要な情報である定性・特定性の推定には，共参照情報に基づく談話上の情報状態よりも，聞き手 / 読み手の観点を考慮した共有性の方が重要であることがわかる。話し手側により踏み込んだ概念である定性・特定性のアノテーションは複雑で難度が高いため，他の文法情報で定性・特定性を推定する方がより容易であると考えられる。この結果は共参照情報よりも共有性の方が高い精度で定性・特定性の推定が可能であることを示している。

表 9 に，ラベルの対応を Kappa 値で評価したものを示す。評価において，「どちらでもない」「どちらでもよい」は排除した。動作主性は，動作主・被動作主の対について評価した。この結果は有生性・有情性が情報状態や共有性の判定に貢献しないことを示唆している。

表 9　ラベルの対応（Kappa 値）

		共有性	定性	特定性	有生性	有情性	動作主性
情報状態	旧情報 /（新情報）	0.51	0.47	0.43	0.1	0.1	0.24
共有性	共有 /（非共有･想定可能）		0.86	0.81	–0.04	–0.04	0.25
定性	定 /（不定）			0.96	0.04	0.04	0.37
特定性	特定 /（不特定）				0.02	0.02	0.36
有生性	有生 /（無生）					0.98	0.41
有情性	有情 /（非情）						0.41

4.2 ラベルと付属語主辞の関係

日本語では助詞「は」と「が」の使い分けについて，情報構造との関連が指摘されている（三上 1963，野田 1996，庵 2010 など）ため，次に名詞句が含まれる文節の付属語主辞[19]と各ラベルの分布について確認する。付属語主辞は係り受け解析器 CaboCha[20]（Kudo and Matsumoto 2002，工藤・松本 2002）に含まれる UniDic 主辞規則に基づくものである。

付属語主辞と情報状態の分布を表 10 に示す。付属語主辞が「、」や「。」となっているものは，格表示されない体言止めの表現を表している。

表 10　付属語主辞と情報状態

付属語主辞	新情報	旧情報
、	284	168
。	130	82
が	99	47
は	59	62
も	29	10
を	222	53
に	84	28
の	108	86
で	40	13
と	13	12

付属語主辞と情報状態の関係を見ると，「が」「を」「に」は新情報が多い。「は」については，旧情報の名詞句と共に使われることが若干多いが，新情報とさほど差があるわけではないことがわかる。これは Nakagawa（2016）の結果と一致している。

付属語主辞と定性の分布を表 11 に，付属語主辞と特定性の分布を表 12 に示す。

表 11　付属語主辞と定性

付属語主辞	定	不定
、	268	184
。	122	90
が	74	72
は	79	42
も	10	29
を	89	186
に	43	69
の	122	72
で	35	29
と	13	12

表 12　付属語主辞と特定性

付属語主辞	特定	不特定
、	270	171
。	126	77
が	74	55
は	82	33
も	10	25
を	96	155
に	46	54
の	123	61
で	36	12
と	14	8

[19] 付属語主辞とは，付属語のうち主要な要素を意味する。
[20] https://taku910.github.io/cabocha/ 参照のこと。

表 11, 表 12 を見ると,「は」「の」に定･特定のものが多い一方,「を」に不定･不特定のものが多い。「は」が定・特定の名詞句と共に使われる（傾向がある）という指摘は，本稿のアノテーションの結果からも支持されるものであることがわかる。

5. おわりに：まとめと今後の課題

本稿では，BCCWJ に対する情報構造のアノテーションデータについて紹介した。本研究では日本語の名詞句に対し，七つの情報構造に関する概念を導入した。本稿で示したアノテーションを行うことによって，日本語母語話者の冠詞選択や冠詞誤りの修正に必要な定性・特定性の推定には，BCCWJ-DepParaPAS に既に付与されていた共参照情報だけでは十分でなく，聞き手 / 読み手の観点を考慮した共有性が重要であることを明らかにした。話し手側により踏み込んだ概念であるため，定性・特定性のアノテーションは複雑で難度が高いが，アノテーションの難度がより低い別の文法情報で定性・特定性を推定する場合，共参照情報よりも共有性の方が高い精度で推定が可能である。例えば，冠詞推定についての言語処理のアプリケーションを構築，実装する上では，言語によって定性か特定性を推定することが必要であるが，情報状態よりも共有性の方がそれらの推定には役に立つと考えられる。また，付属語主辞とのラベルの関係を見ると，「が」「を」「に」は新情報が多く,「は」は旧情報が若干多いこと,「は」「の」に定･特定のものが多く,「を」に不定・不特定のものが多いことがわかった。

今後の課題は以下に示す通りである。

まず，情報構造をアノテーションする被験者実験を行う。本稿で示したラベルは言語学者が議論しながら検討しアノテーションしたものである。被験者実験のためには，言語学の専門知識を持たない人でも回答が可能なように，わかりやすい質問に落とし込む必要がある。非言語学者によって情報構造のラベルを判定できる質問を作成し，非言語学者にもわかるような言語学的なテストを設計する。これにより，アノテーションの数だけでなく，標的サンプルをも増やすことができる。

第 2 に，ラベルの再検討を行っていくことも考える。名詞の飽和 / 非飽和（西山 2003）や Löbner（1985, 2011）の名詞の 4 分類（Sortal（タイプ $\langle e, t\rangle$）/ Individual（タイプ e）/ Relational（タイプ $\langle e, \langle e, t\rangle\rangle$）/ Functional（タイプ $\langle e, e\rangle$））を付与することで，名詞の指示性と他のラベルの関係性がわかってくると考えられる。さらに，動作主性を動作主 / 被動作主のみでなく，より詳細な意味役割として，受益者，経験者，道具，場所などを付与することで，意味役割と他のラベルの関係性を突き止めることも考えたい。

第 3 に，BCCWJ の翻訳テキストとの対照分析を行う。今回アノテーションしたデータの一部は，英語・イタリア語・中国語・インドネシア語の人手による翻訳が公開されている。英語・イタリア語については，情報構造と冠詞の有無を対照することで，人手による翻訳においてどのように冠詞付与がなされているかを明らかにする。

最後に各情報構造のラベルと文中の位置との関係や係り受けの数との関係を統計的に分析する。旧情報が新情報の先に来るのかなどを人手によるアノテーションに基づいて分析を行う。

参照文献

Asahara, Masayuki (2017) Between reading time and information structure. *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation PACLIC 31*.

浅原正幸・大村舞（2016）「BCCWJ-DepParaPAS：『現代日本語書き言葉均衡コーパス』係り受け・並列構造と述語項構造・共参照アノテーションの重ね合わせと可視化」『言語処理学会第 22 回年次大会発表論文集』489–492.

Asahara, Masayuki, Hajime Ono and Edson T. Miyamoto (2016) Reading-time annotations for *Balanced Corpus of Contemporary Written Japanese*. *Proceedings of COLING-2016*, 684–694.

Calhoun, Sasha, Malvina Nissim, Mark Steedman and Jason Brenier (2005) A framework for annotating information structure in discourse. In: Adam Meyers (ed.) *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, 45–52. Ann Arbor: The Association for Computational Linguistics.

Cohen, Jacob (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37–46.

Givón, Talmy (1976) Topic, pronoun, and grammatical agreement. In: Charles N. Li (ed.) (1976), 149–187.

Götze, Michael, Thomas Weskott, Cornelia Endriss, Ines Fiedler, Stefan Hinterwimmer, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas and Ruben Stoel (2007) Information structure. In: Stefanie Dipper, Michael Götze and Stavros Skopeteas (eds.) *Information structure in cross-linguistic corpora: Annotation guidelines for phonology, morphology, syntax, semantics and information structure*, 147–187. Potsdam: Universitätsverlag Potsdam.

Hajičová, Eva, Jarmila Panevová and Petr Sgall (2000) *A manual for tectogrammatical tagging of the Prague Dependency Treebank*. *ÚFAL/CKL Technical Report TR-2000-09*. Prague: Charles University.

Heim, Irene (2011) Definiteness and indefiniteness. In: Klaus von Heusinger et al. (eds.) (2011), 996–1025.

von Heusinger, Klaus (2011) Specificity. In: Klaus von Heusinger et al. (eds.) (2011), 1058–1087.

von Heusinger, Klaus, Claudia Maienborn and Paul Portner (eds.) (2011) *Semantics: An international handbook of natural language meaning* Vol. 2. Berlin: Mouton de Gruyter.

Hinterwimmer, Stefan (2011) Information structure and truth-conditional semantics. In: Klaus von Heusinger et al. (eds.) (2011), 1875–1908.

Ionin, Tania, Heejeong Ko and Kenneth Wexler (2004) Article semantics in L2 acquisition: The role of specificity. *Language Acquisition* 12: 3–69.

庵功雄（2010）「産出のための日本語教育文法：「は」と「が」の使い分けを例として」『台湾日本語文学報』28: 40–55.

Keenan, Edward L. (1976) Towards a universal definition of subject. In: Charles N. Li (ed.) (1976), 303–334.

Kruijff-Korbayová, Ivana and Mark Steedman (2003) Discourse and information structure. *Journal of Logic, Language and Information* 12: 249–259.

Kudo, Taku and Yuji Matsumoto (2002) Japanese dependency analysis using cascaded chunking. *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, 63–69.

工藤拓・松本裕治（2002）「チャンキングの段階適用による日本語係り受け解析」『情報処理学会論文誌』43(6): 1834–1842.

Landis, J. Richard and Gary G. Koch (1977) The measurement of observer agreement for categorical data. *Biometrics* 33: 159–174.

Li, Charles N. (ed.) (1976) *Subject and topic*. New York: Academic Press.

Löbner, Sebastian (1985) Definites. *Journal of Semantics* 4: 279–326.

Löbner, Sebastian (2011) Concept types and determination. *Journal of Semantics* 28: 279–333.

Lyons, Christopher (1999) *Definiteness*. Cambridge: Cambridge University Press.

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka and Yasuharu Den (2014) Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation* 48: 345–371.

三上章（1963）『日本語の論理』東京：くろしお出版.

Mosel, Ulike and Even Hovdhaugen (1992) *Samoan reference grammar*. Oslo: Scandinavian University Press.

Nagata, Ryo, Tatsuya Iguchi, Fumito Masui, Atsuo Kawai and Naoki Isu (2005) A statistical model based on the three head words for detecting article errors. *IEICE TRANSACTINOS on Information and Systems*, 1700–1706.

Nakagawa, Natsuko (2016) Information structure in spoken Japanese: Particles, word order, and intonation. Ph.D. thesis, Kyoto University.

西山佑司（2003）『日本語名詞句の意味論と語用論：指示的名詞句と非指示的名詞句』東京：ひつじ書房.
野田尚史（1996）『新日本語文法選書 1「は」と「が」』東京：くろしお出版.
Sato, Hiroko (2013) Definiteness and specificity in Kove. *Proceedings of the International Workshop on Information Structure of Austronesian Languages*, 37–45.
Steedman, Mark (2000) Information structure and the syntax-phonology interface. *Linguistic Inquiry* 34: 649–689.
Tanaka, Junko (2013) A multivariate analysis of L2 English article use by article-less L1 learners. In: Erik Voss, Shu-Ju D. Tai and Zhi Li (eds.) *Selected Proceedings of the 2011 Second Language Research Forum*, 139–147. Somerville, MA: Cascadilla Press.
植田禎子・飯田龍・浅原正幸・松本裕治・徳永健伸（2015）「『現代日本語書き言葉均衡コーパス』に対する述語項構造・共参照関係アノテーション」『第 8 回コーパス日本語学ワークショップ予稿集』205–214.
Vallduví, Enric and Maria Vilkuna (1998) On rheme and kontrast. In: Peter Culicover and Louise McNally (eds.) *The limits of syntax*, 79–108. San Diego: Academic Press.
山口光（1985）「存在文と所有文」金田一春彦・林大・柴田武（編）『日本語大辞典』198–200.　東京：大修館書店.

関連 Web サイト

国立国語研究所『現代日本語書き言葉均衡コーパス』「中納言」https://chunagon.ninjal.ac.jp/

Construction and Analysis of Information-Structure Annotation of the "Balanced Corpus of Contemporary Written Japanese"

MIYAUCHI Takuya[a]　ASAHARA Masayuki[b]　NAKAGAWA Natsuko[c]　KATO Sachi[d]

[a]Ph.D. Student, Tokyo University of Foreign Studies / JSPS Research Fellow / Project Collaborator, NINJAL
[b]Center for Corpus Development, NINJAL
[c]Research Fellow, Graduate School of Humanities, Chiba University
[d]Adjunct Researcher, Center for Corpus Development, NINJAL

Abstract
This paper presents the information structure's annotation data (information status, commonness, definiteness, specificity, animacy, sentience, and agentivity) of the "Balanced Corpus of Contemporary Written Japanese." The annotation schema and statistics are displayed. Evaluation utilizing Kappa value indicates a moderate agreement (0.41≤) between the information status that is based on the already annotated co-reference information and definiteness/specificity. In addition, there is an almost perfect agreement (0.81≤) between commonness, which is recently annotated in this research, and definiteness/specificity. Thus, we conclude that commonness is more significant than information status to estimate definiteness and specificity, significantly affecting article selection in languages with articles. We investigate the relation between some particles and labels explained in this research since some researchers report that information structure is related to the distinction between the particles *wa* and *ga* in Japanese. Hence, the particles *ga*, *o*, and *ni* are usually employed with discourse-new noun phrases and *wa* with discourse-old ones. The particle *wa* is generally employed with definite and specific noun phrases, while *o* is employed with indefinite and unspecific ones.

Key words: information structure, annotation, noun phrase, article selection, particle