『現代日本語書き言葉均衡コーパス』に対する名詞述語文アノテーション

著者	今田 水穂
雑誌名	言語資源活用ワークショップ発表論文集
巻	3
ページ	382-398
発行年	2018
URL	http://doi.org/10.15084/00001673

『現代日本語書き言葉均衡コーパス』に対する名詞述語文アノテーション

今田 水穂 (文部科学省) *

Copular Sentence Annotation on 'Balanced Corpus of Contemporary Written Japanese'

Mizuho Imada (MEXT)

要旨

「現代日本語書き言葉均衡コーパス」コアデータに対して名詞述語文に関する文法情報付与を行った。付与したラベルの概数は名詞述語文の主語述語 13700 組、名詞述語の連体修飾節 3200、機能表現などの周辺的ラベル 4600 である。主語は名詞主語とノ節主語に分類し、前者は is_a など述語との意味関係を、後者は分裂文焦点に対する文法関係を付与した。述語は通常の名詞述語の他、「X は Y になる」のような補語も若干数付与した。周辺表現ラベルは「する」が省略された漢語動詞、名詞述語由来の機能表現、述語が省略された節などを含む。本稿では、データの設計と既存の述語項構造データとの違い、構築したデータの計量的概観について説明し、本データが名詞述語文の文法研究における諸問題とどのように関係するかについて論じる。

1. はじめに

日本語の名詞述語文研究では、現象の観点からは措定と指定の区別 (三上 1953)、ウナギ文 (奥津 1978)、述語名詞の文末表現化 (新屋 1989, 角田 2011) などが、理論的な観点からは名詞の指示性 (西山 2003) や文の情報構造 (砂川 2005) などが主要な話題として取り上げられてきた。主語と述語の意味関係という観点から見ると、多くの研究では is_a 関係や is_the 関係を表す文を名詞述語文の中心的事例と見なしており、ウナギ文などそれ以外の関係を表す文はしばしば名詞述語文の周辺的事例として扱われてきた。

表 1　be の多義性

文	意味	文	意味
吾輩は猫である	吾輩 is_a 猫	きゅうりは緑色だ	きゅうり is 緑色
田中が幹事だ	田中 is_the 幹事	子豚は 3 匹だ	子豚 amounts_to 3 匹
山田は愉快な性格だ	山田 has 愉快な性格	僕はウナギだ	僕 eats ウナギ

* imadamizuho.ac@google.com

しかしながら、名詞述語文にはウナギ文のように文脈に依存して関係が決定するようなものから、事物の属性や数量を叙述するような文脈にそれほど依存しなくても解釈が可能なものまで、is_a や is_the 以外の様々な関係を表す文があり、文の意味解釈を考える上で be の多義性を解決することは不可欠の課題である。

名詞述語文の意味を構成する主要な要素の 1 つは名詞の意味である。「猫」や「緑色」は形式意味論では 1 項述語だが、存在論的タイプが異なる。「幹事」は「誰」と「何」の 2 項を取る 2 項述語であり、「誰」をガ格、「何」をノ格で取る。別の 2 項述語には「高さ」があり、対象と数値を項に取るが、統語的な具現化は一様ではなく、「東京タワーの高さは 333m だ」「東京タワーは 333m の高さだ」などがある。名詞の項構造に関する最近の研究としては、庵 (2007)、竹内 (2015) などがある。

もう 1 つの主要な要素は名詞の意味を組み合わせて文全体の意味を作るための形式的な規則である。この規則は「東京タワーの高さは 333m だ」「東京タワーは 333m の高さだ」「東京タワーは 333m だ」などの異なる統語構造を共通の意味解釈へと結びつける必要があり、標準的な形式意味論の演算規則では必ずしも十分ではない。生成語彙論 (Pustejovsky 1998) は意味演算規則を拡張する有力な提案の 1 つであり、概念と概念の関係に関する知識 (クオリア構造) が句の意味形成において積極的な役割を果たす。形式意味論を用いた名詞述語文の最近の研究には郡司 (2015, 2016) が、生成語彙論を用いた研究には今田 (2012) がある。

本研究では、これらの意味論的研究を進めるための言語資源の整備を目的として、「現代日本語書き言葉均衡コーパス」(BCCWJ) コアデータに対する名詞述語文の述語項構造付与を行っている。同データに対する述語項構造データとしては既に BCCWJ-PAS(小町・飯田 2011) があるが、名詞述語は特にコピュラが省略された場合に述語か否かの判別が難しく、BCCWJ-PAS と本データの判定には異同がある。現段階において本研究が付与している述語項構造ラベルは、名詞述語文の主語と述語の関係を記述するものと、名詞述語が他の節の項に相当する場合にその関係を記述するものである。述語名詞が他の節の項に相当する場合とは、「肉を食べたのはトラです」のような分裂文と、「トラは肉を食べる動物です」のような連体修飾節である。他に「～は～が～だ」構文や属格のアノテーションも行いたいが、現時点では着手できていない。

意味論的情報の付与は整備したデータを用いた次の段階の研究課題だが、試験的に分裂文以外の名詞述語文について主語と述語の意味関係を is_a、is_the、has などのラベルで付与した。これらはより複雑な意味論的情報を付与するための予備的分類としての用途の他に、主語と述語の存在論的タイプが一致しない名詞述語文の数量的内訳の調査や、措定文、指定文などの名詞述語文類型についての記述的研究のための用例収集などの用途を想定している。また、名詞述語やコピュラ由来の機能表現など、名詞述語文のアノテーションの過程で見つかった周辺的表現についてもラベルを付与した。以下では、構築したデータの設計について説明し、付与したラベルの数量的内訳の報告と、記述的および理論的研究のための利活用についての展望を述べる。

2. データの設計

2.1 名詞述語文ラベル

名詞述語文に対して、主語と述語の関係を示す述語項構造ラベル (図 1 上段) と、節主語 (ノ節のみ) または連体修飾節と述語の関係を示す述語項構造ラベル (図 1 下段) を付与した(1)。

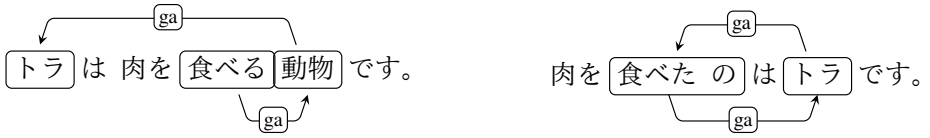

図 1　アノテーション対象とする述語項構造

ただし本データは名詞述語文のアノテーションを目的としているため、全ての情報を名詞述語を主要部とする形式に集約した。そのため、節主語や連体修飾節については、一般的な述語項構造ラベルとは矢印の向きが逆になっている。「〜は〜が〜だ」構文のように主語が複数ある場合や、連体修飾節が複数ある場合には、述語名詞に最も近いものに対してのみラベルを付与した。

図 2　本データにおけるアノテーション設計

この設計では、名詞述語文に付与するラベルは名詞述語ラベル (npred、ncomp)、主語ラベル (nsubj、csubj)、連体修飾節ラベル (acl) の 3 種類に分類される。名詞述語ラベルには npred と ncomp の 2 種類があり、npred は通常の名詞述語を表す。ncomp は「X が Y になる」「X を Y とする」など Y が節末の述語ではなく補語の位置に生起する場合で、実質的に X と Y の間に主語と述語の関係が成り立つ。

図 3　名詞述語

主語ラベルは nsubj と csubj の 2 種類がある。nsubj は名詞主語である。主語は述語に直接かかっているとは限らず、連体修飾節の被修飾語や先行文脈中の名詞句などの場合がある。テ

(1) データの構築は、既存のアノテーションデータ (BCCWJ、BCCWJ-PAS、および BCCWJ-CBL (丸山 2013) を参照してアノテーション対象文字列の候補をリストアップし、人手で修正する方法で実施した。なお、本データで使用するラベル名の一部は Universal Dependencies (http://universaldependencies.org) を参考にした。

キスト中に主語に相当する文字列がない場合 (外界参照など) は、BCCWJ-PAS の仕様に倣い、1 人称 (exo1)、2 人称 (exo2)、一般 (exog)、節参照 (ana_cla) のいずれかのラベルを付与した。

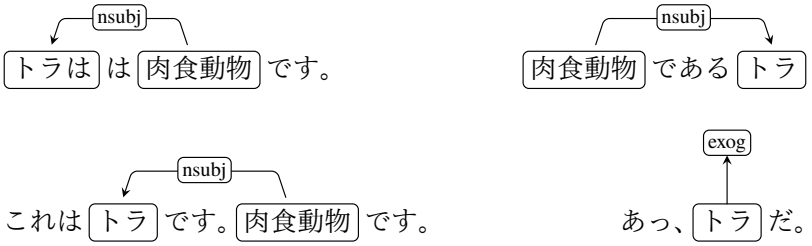

図 4　名詞主語

nsubj の下位分類として、主語と述語の関係を次の 8 種に分類し、いずれかのラベルを付与した。is_a と is_the は主語と述語の存在論的タイプが一致する。それ以外のラベルのうち、is_sth_like は隠喩に相当し、他は広義の換喩に相当する。

表 2　意味関係ラベル

ラベル	説明
is_a	主語が述語の下位クラスまたはインスタンス
is_the	主語と述語が同一
has	主語と述語が所有関係 (「X の Y」)
has_prop_of	主語と述語が所有関係 (「X の Y」と言いにくいもの)
means	定義文など
is_sth_like	比喩表現など
amounts_to	数量の叙述
other	その他

csubj は節主語である。ただし、csubj は分裂文を区別することが主な目的のため「の」節にのみ付与した。

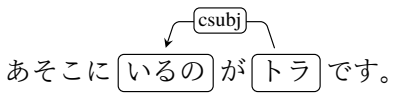

図 5　節主語

csubj の下位分類として、節主語の述語と名詞述語の文法関係を分類し、以下のいずれかのラベルを付与した。

表 3　文法関係ラベル

ラベル	説明	例
ga	ガ格	笹を食べるのはパンダだ (パンダが笹を食べる)
o	ヲ格	パンダが食べるのは笹だ (パンダが笹を食べる)
ni	ニ格	パンダがいるのは上野だ (上野にパンダがいる)
time	時間表現	パンダが来たのは先月だ (先月パンダが来た)
mod	その他の格と修飾語	パンダが来たのは中国からだ (中国からパンダが来た)
_mod	間接的修飾語	爪が鋭いのはパンダだ (パンダの爪は鋭い)
ext	外の関係	パンダが笹を食べるのは事実だ

連体修飾節ラベルは acl の 1 種類のみである。動詞、形容詞、名詞述語などの連体形の他、名詞述語の連体形に相当すると考えられる名詞 +「の」も連体修飾節として扱った。

図 6　連体修飾節

acl の下位分類として、連体修飾節の述語と名詞述語の文法関係を分類し、ラベルを付与した。ラベルは csubj の下位分類と同様だが、日本語記述文法研究会 (2008) に倣い、外の関係を次の 3 種類に分類した。

表 4　文法関係ラベル (acl 限定)

ラベル	説明	例
e1_内容補充	外の関係	パンダが笹を食べる事実
e2_相対名詞		パンダが笹を食べた後
e3_付随名詞		パンダが笹を食べた結果

2.2　機能表現ラベル等

機能表現ラベル等は、名詞やコピュラ由来の機能表現など、名詞述語文に関係する周辺的表現をマークアップしたものである。

表 5　機能表現ラベル等

ラベル	説明
vpred	動詞述語 (「する」が省略されたサ変動詞)
xpred	その他の述語
naux	名詞由来助動詞 (「はずだ」など)
nmark	名詞由来機能表現 (「中で」など)
cmark	コピュラ由来の機能表現 (「だが」など)
func	その他の機能表現 (「によって」など)
orphan	述語が省略された節末名詞句 (「トラに願いを。」)
concat	2 つ以上の文節が連結して 1 つの文節のようになっているもの
ambig	名詞述語か動詞述語か特定できないもの (「いざ調査開始！」)

vpred と xpred は名詞述語以外の述語である。vpred は名詞述語と同形の動詞述語であり、サ変名詞に後接する「する」が省略されたものである。xpred はその他の述語で 3 例あるが、いずれもイレギュラーな例であり、今後廃止する可能性がある。naux、nmark、cmark、func は名詞やコピュラに由来する機能表現である。func はそれ以外の機能表現であり、多くは格助詞 + 動詞の形式の複合助詞である。orphan は述語の省略された名詞句に付与した(2)。concat は元データである BCCWJ において、複数の文節に分割すべきものが単一の文節にまとめられている場合に付与した。ambig はラベルの判別が不可能なものに付与した。多くは、サ変名詞が述語として使用されているが、名詞に後接する「だ」ないし「する」が省略されており、かつ「だ」「する」のいずれも付加可能なため、npred か vpred かの判断ができないものである。

naux、nmark、cmark が名詞やコピュラに由来する周辺的表現であるのに対して、func、orphan、concat はいずれも名詞述語と直接関係しない言語要素だが、共通点がある。もともと名詞述語は「名詞 + で」と「ある」の 2 つの文節が結合して 1 つの文節になったものである(3)。「だ」「です」などは「である」と異なり 2 つの部分に分割することができないが、名詞述語が名詞句としての機能と述語としての機能を併せ持った文節であることは同様である。func は「を」「に」「と」などの格助詞と動詞が結合して機能語化したものであり、これも元は名詞句と述語が融合した文節と考えられる。orphan は述語を欠く名詞句だが、省略された述語はテキスト中に実体を持たないので、アノテーション仕様上は名詞句にゼロ形式の述語が包摂されたものとして扱う方法が考えられる。concat の一部も、名詞句と述語が結合して 1 つの文節になったものである。

(2) 本データの orphan は原則として文末の名詞句に限られるが、類似する構造は文中に生起することもある。1 つは「犬が肉を、猫が魚を食べた」のような部分並列構造である。部分並列構造をアノテーションしたコーパスはいくつか存在するが、コーパスによって扱いは異なる。浅原 (2013) など参照。もう 1 つは「～を目標に頑張る」のような副詞節で、「～を」のかかり先は「目標に」だが、活用する述語を欠く。

(3) 中学校の国語教科書では「だ」「です」は助動詞の一覧に含まれるが「である」は含まれないことが一般的である。この場合、「である」は「名詞 + で」と補助動詞「ある」の 2 つの文節に分割される。

表 6　結合された文節

ラベル	例		
npred	猫で + ある	→	猫である
func	猫に + よって	→	猫によって
orphan	願いを + ∅。	→	願いを。
concat	関係が + ない	→	関係がない

本研究は自動処理によりアノテーション対象の候補をリストアップし、人手で修正するという手続きを取った。これらの言語表現がリストアップされたのは名詞文節と述語文節の結合という構造的類似性によるものである。本研究の本来の目的からは外れる表現だが、構造的類似性という観点からは名詞述語の近傍の研究対象と考えることができるため、特にラベルを付与して残した。

3. データの計量的概観

3.1 概要

BCCWJ コアデータの書籍、雑誌、新聞、白書の 4 レジスタ (571 サンプル) に対してアノテーションを行い、延べ 35293 のラベルを付与した。ラベルの内訳を表 7 に示す。

表 7　ラベル数

ラベル			要素数	小計	合計
名詞述語文ラベル	述語	npred	13487	13730	30673
		ncomp	243		
	主語	nsubj	12617	13730	
		csubj	1113		
	連体修飾節	acl	3213	3213	
機能表現ラベル等	述語	vpred	1516	1519	4620
		xpred	3		
	機能表現	naux	806	1792	
		nmark	241		
		cmark	206		
		func	539		
	その他	orphan	744	1309	
		concat	493		
		ambig	72		

BCCWJ-PAS[(4)]との差異を確認する。ここでは、BCCWJ の「名詞」または「代名詞」の長単位単語のうち、長単位を構成する短単位のいずれかに PAS の述語タグが付与されているものを PAS における名詞述語と見なす。この数え方では、PAS の名詞述語は書籍、雑誌、新聞、白書の 4 レジスタで 11512 ある。本データの npred ラベルとの重なりを確認すると、本データと PAS の両方で名詞述語とするもの 8266 例、本データのみ名詞述語とするもの 5221 例、PAS のみ名詞述語とするもの 3246 例だった。PAS のみ名詞述語とするもの 3246 例について、本データで付与したタグの内訳を次の図に示す。nsubj、csubj、acl は npred ないし ncomp に付随するタグなので対応関係から除外する。remove はいずれのタグにも該当しないと見なしてタグを付与しなかったものである。

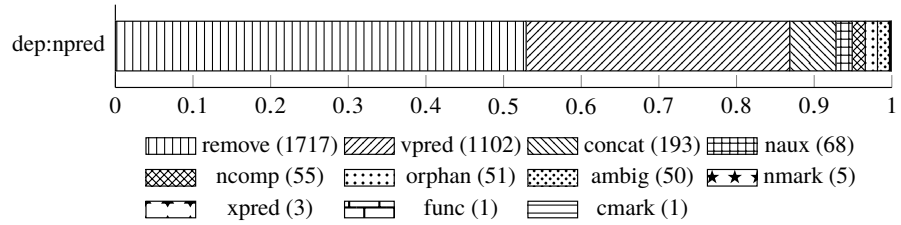

図 7　BCCWJ-PAS における名詞述語の本データにおける扱い

3246 例のうち約 5 割の 1717 例は、本データでは名詞述語文ではないと見なし、アノテーション対象から除外している。これらの多くは文全体が名詞句であるなど、文末に名詞が生起するが述語として機能していないものである。また、約 3 割の 1102 例は本データでは動詞述語 (vpred) としてアノテーションした。これらは文末にサ変名詞が生起しているが、「だ」ではなく「する」が省略されていると判断できるものである。

3.2　名詞述語文ラベル

3.2.1　名詞述語・補語

名詞述語は補語も含めて約 13700 語ある。このうち補語は 243 語と少ないが、今回のアノテーションは述語を中心に実施したので、補語は悉皆的に付与できていない。補語に後節する述語は「なる」が 141 例と最も多く、次いで「する」19 例、「いう」13 例などがあり、その他には「認める」「位置付ける」「期待する」「思う」などが見られる。名詞述語に準ずる補語については、今後、悉皆的なアノテーションを検討したい。

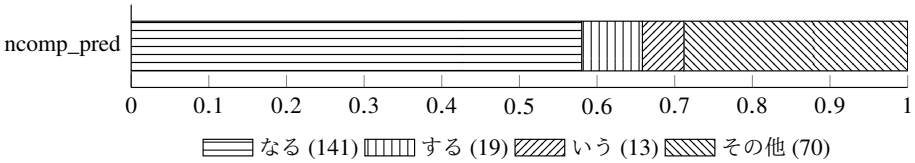

図 8　ncomp に後接する述語

(4) BCCWJ-DepParaPAS-3.3.0_1.2.0 を使用した。

3.2.2 名詞主語

主語は約 13800 語あるが、そのうち名詞主語は約 12700 語、節主語は約 1100 語である。ここで言う節主語は「の」節に限られるが、主語の大部分が名詞主語であることが分かる。

名詞主語の下位分類を確認する。名詞述語文は is_a や is_the を表すものが規範的であり、それ以外のものは周辺的である (特に語用論的な解釈を要求されるものがウナギ文と呼ばれる) と見なされる傾向がある。しかし実際には、is_a と is_the は合計しても名詞述語文全体の半分程度であり、それ以外の主語と述語の存在論的タイプが一致しないような名詞述語文が大きな割合を占めることが分かる。

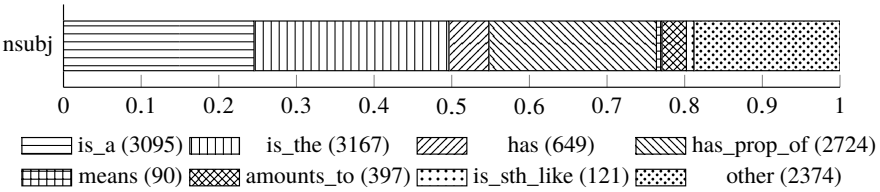

図 9　nsubj における意味関係ラベルの内訳

3.2.3 節主語

節主語の下位分類 (名詞述語との文法関係) の内訳を示す。時間表現も含めると、節主語内の要素が名詞述語として焦点化された分裂文とみなすことができるものが過半数を超えるが、分裂文ではない節主語も 4 割以上を占める。

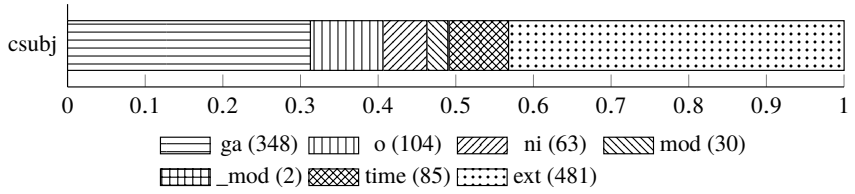

図 10　csubj における文法関係ラベルの内訳

分裂文ではない節主語が、どのような名詞述語を取るか確認する。「こと」「もの」などの形式名詞の他、「特徴」「狙い」「目的」など命題的情報を内容として持つ概念を表す名詞が多く、その内容を「〜のが特徴だ」のように節主語が具体的に示す。

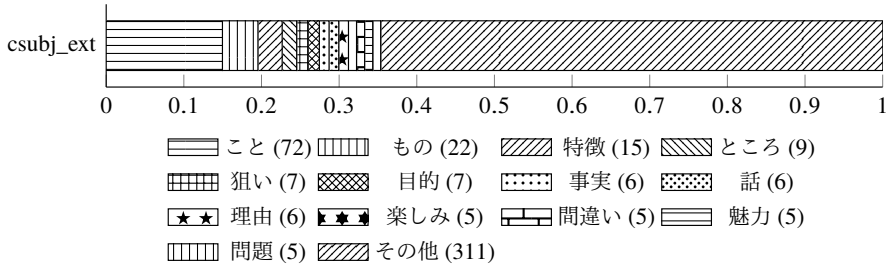

図 11　csubj(ext) における名詞述語の内訳

3.2.4 連体修飾節

連体修飾節は約 3200 例あり、名詞述語の約 23% が連体修飾節を持つ。連体修飾節の下位分類 (名詞述語との文法関係) の内訳を示す。節主語と比べると ga の割合が大きいことが分かる。また、節主語ではほとんど見られなかった_mod も、連体修飾節では比較的多く見られる。外の関係の連帯修飾節は、ほとんどが内容補充連体修飾節だった。

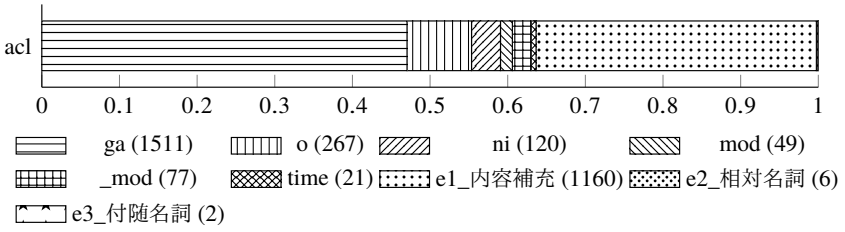

図 12　acl における文法関係ラベルの内訳

内容補充連体修飾節を取る名詞述語の内訳を確認する。分裂文ではない節主語の場合と同様、名詞述語は命題的情報を内容として持つ概念を表す名詞が多い。しかし個別の名詞を見ると「予定」「疑い」「方針」などが高頻度であり、節主語の場合とは語彙が異なる。

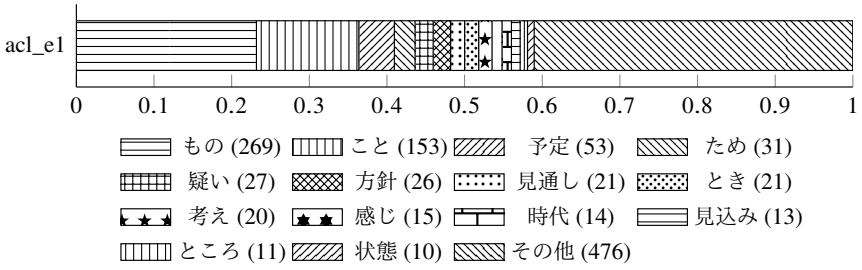

図 13　acl(e1) における名詞述語の内訳

3.3 機能表現ラベル

名詞述語文ラベル以外のラベルのうち、付与対象となる表現がある程度限られる naux、nmark、cmark、func の 4 種類の機能表現ラベルについて、主要な表現の内訳を示す。naux は名詞由来の助動詞相当表現で、ほとんどが形式名詞 +「だ」である。主なものを以下に示す。表記上の変種 (「はず」「筈」や「だ」「です」「である」など) はまとめて集計した (この節の他の図も同様)。

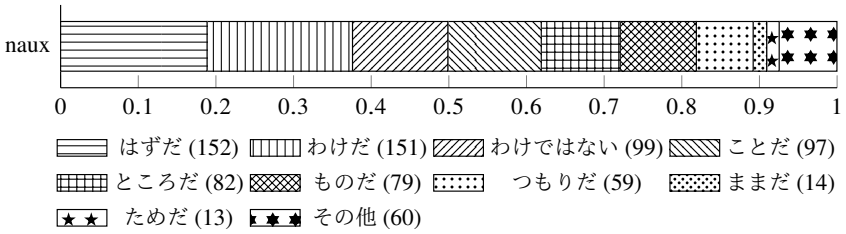

図 14　naux の内訳

nmark は名詞由来の助動詞以外の機能表現である。多くは形式名詞か、または形式名詞 + 「で」「に」の形態で、連用的な従属節を作る。主なものを以下に示す。

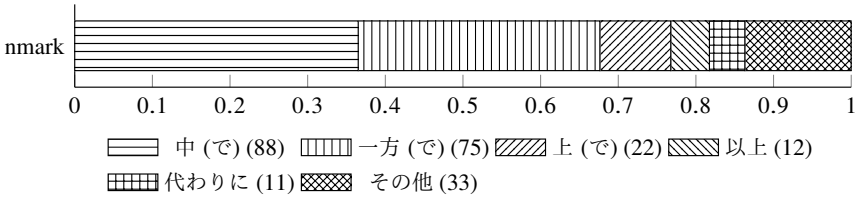

図 15　nmark の内訳

cmark は「だ」などのコピュラに由来する機能表現である。多くは、提題ないし取り立ての機能を持つもの (「だが」「だって」「だと」など) か、並列の機能を持つもの (「～だ～だ」「～だの～だの」「～だとか～だとか」) のいずれかに分類できる。主なものを以下に示す。

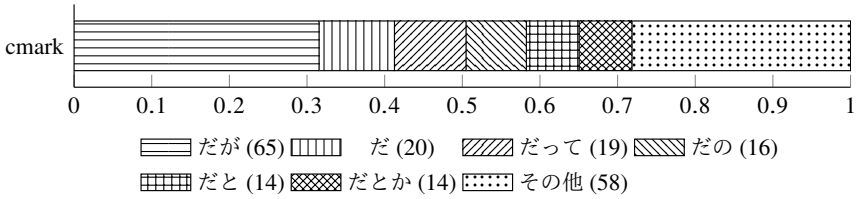

図 16　cmark の内訳

func はその他の機能表現であり、多くは格助詞 + 動詞に由来する複合助詞である。主な例を以下に示す。naux、nmark、cmark などが名詞やコピュラを由来とする名詞述語文の周辺的表現であるのに対して、func は基本的に名詞述語文とは関係がない。これらの表現が今回のアノテーションの副産物としてタグ付与された背景については、2.2 節で述べた通りである。

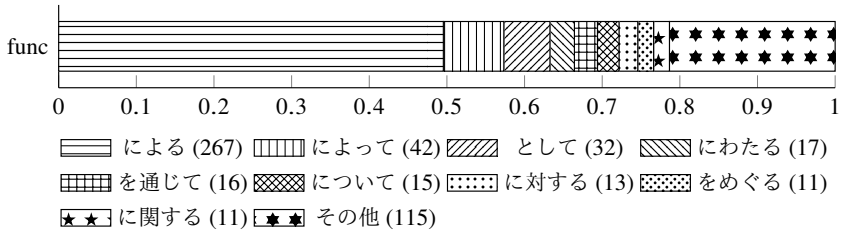

図 17　func の内訳

4. データの利活用

4.1 記述研究との接点

本データの意味関係ラベルはより複雑な意味論的情報を付与するための予備的分類として付与したものだが、一方で名詞述語文研究における主要な研究課題のために役立つことを念頭において設計している。また節主語や連体修飾節に付与した文法関係ラベルは BCCWJ-DepPara には含まれないガ、ヲ、ニ以外の文法関係もサポートしており、文法研究のために役立つ。個別の研究課題について、本データのどのラベルが関係するかについてまとめる。

■措定と指定　is_a と is_the は包摂関係と同一関係を区別するために設定したラベルであり、措定文と指定文の区別とある程度対応することを想定している。措定文、指定文の他に同一性文などの分類を設ける場合には、これも is_the に含まれる。これらの記述的分類は、名詞句の指示性 (西山 2003)、主題や焦点などの情報構造 (砂川 2005)、あるいは「指定する」「同定する」といった発話機能によって区別されるものであり、包摂関係や同一関係のような集合論的関係のみで区別されるものではない。しかし、より詳細な分析のための大まかな前処理としては役立つ。

■ウナギ文　is_a と is_the 以外の関係を表す名詞述語文は、しばしばウナギ文と総称される広範な領域に押し込められて、十分な整理が与えられてこなかった。本データでは、この領域に対して has、has_prop_of、amounts_to、other という下位区分を割り当てている。has は主語と述語の間に属格関係が成り立つ場合であり、述語が部分 (「彼はおかしな顔だ」)、属性 (「彼はおかしな性格だ」)、命題的概念 (「彼は～する予定だ」) などを表す文に適用することを想定している。他の 3 つは属格関係が成り立たない場合であり、has_prop_of は属性、amounts_to は数量、other はその他の様々な事物や概念との関係を記述することを想定している。

■隠喩文　is_sth_like は「人生は旅だ (のようなものだ)」などの比喩表現に付与するために設定したラベルである。ウナギ文がしばしば換喩と関連付けて論じられるのに対して、この構文は直喩や隠喩と関連する。特に隠喩は、形態的には通常の名詞述語文との区別がつかず、換喩と同様に意味論的な情報の付与が不可欠である。

■文末名詞文　文末名詞文は「～は～する予定だ」のように、名詞述語が文法化して文末表現のようになった文である。名詞述語は連体修飾節を持ち、名詞述語の文末表現化によって連体

修飾節が主節化する。連体修飾節が主節化するためには、主節主語が連体修飾節述語の項であり、かつ主節述語が項ではない (削除可能である) 必要があり (今田 2017)、従って連体修飾節は多くの場合、外の関係の連体修飾節である。本データでは acl の下位分類として文法関係ラベルを付与しており、このラベルは外の関係も含む。BCCWJ-DepPara の述語項構造データと組み合わせることで、文末名詞文のような複雑な構造的条件を満たす構文を効率的に抽出することができる。

■分裂文 「のは」「のが」で表示される主語は、形態的特徴から容易に抽出することが可能だが、本データでは csubj ラベルを付与し、さらに下位分類として文法関係ラベルを付与した。「のは」を主語とする名詞述語文には分裂文[(5)]とそれ以外のものがあるが、両者の判別のために文法関係ラベルが役立つ。また、同定文・提示文 (西山 2003) は主語が「のが」で表示される文を多く含む。

■定義文 means は「～とは～だ」のような定義文を記述するために設定したラベルである。主語は「とは」「というのは」で表示されるが、「は」の場合もある。述語は「～のことだ」の形式が多いが、「～という意味だ」にも means ラベルを付与した。分裂文と同様、形態的特徴から抽出しやすい構文だが、通常の名詞述語文と同形のものもあるため means ラベルが役立つ。定義文は、主語のメタ言語性、定義文と同定文の機能的類似性、定義文と分裂文の統語的類似性などの研究課題がある。

4.2 理論研究との接点

Jackendoff (2002) は文の意味構造をいくつかの層に分割して記述する。この枠組みを使うと、名詞述語文の意味構造は次のように記述することができる。

音韻/統語構造	吾輩 $_1$ は猫 $_2$ である
記述層	be(x_1, y_2)
指示層	1
情報構造層	Topic_1

本研究の意味分類ラベルは記述層の情報を記述するものであり、be の下位分類に相当する。従来の文法研究で問題とされてきた名詞述語文の意味論的特徴のいくつかについては、この構造の別の層で記述される。例えば、この構造の指示層は「吾輩」が指示的で「猫」が叙述的であることを表し、情報構造層は「吾輩」が主題であることを表す。

本データでは be の下位分類を 8 種類のラベルで表現したが、実際には be の解釈はより多様である。体系的で豊かな意味記述のために、意味関係ラベルを組織化されたより大規模な意味データベースで置き換えることを考えてみることにしよう。SUMO(Suggested Upper Merged Ontology)(Niles and Pease 2001) は述語論理をベースとしたオントロジー体系である。本データのラベルの少なくとも一部は、SUMO の述語で次のように置き換えることができる。

(5) 正確には英語の分裂文 “It is . . . that . . .” ではなく擬似分裂文 “What . . . is . . .” に相当する。

表 8　SUMO 表現

本データ	SUMO
x is_a y	instance(x, y) | subclass(x, y)
x is_the y	equal(x, y)
x has_prop_of y	attribute(x, y)
x amounts_to y	measure(x, y)
x means y	containsInformation(x, y)

しかし理論的により興味深いことは、これらの述語を be の下位分類として割り当てることよりも、様々な構文を横断して観察される同義性を一般的に記述するのに役立つということである。次の図は、BCCWJ の実例の意味解釈を SUMO 述語のネットワークで表現したものである(6)。

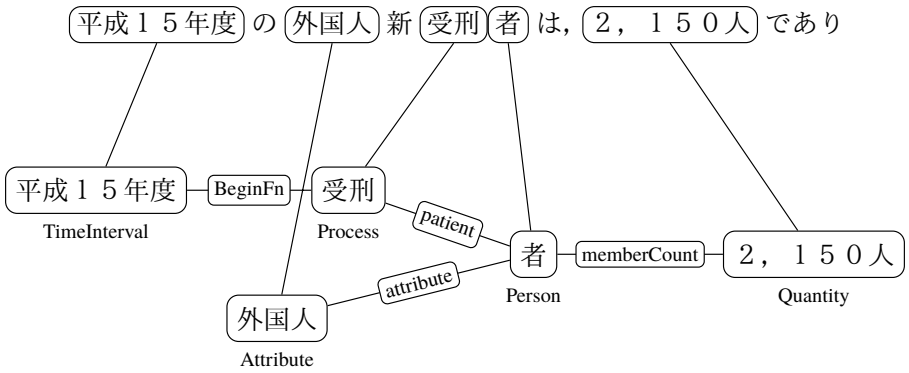

図 18　意味ネットワーク

同じ意味構成は「平成１５年の新受刑者は２，１５０人が外国人だ」「外国人新受刑者は平成１５年が２，１５０人だ」など、様々な構文で表現される。これらの構文が共有する意味情報は、統語規則、語彙規則、インターフェイス規則などによってそれぞれの構文にエンコードされるが、どの部門に多くの仕事を与えるかは理論によって異なる(7)。

統語と意味のギャップをどの部門で解消する立場を取るにせよ、運用論的な観点からは我々

(6) 簡単のために、SUMO 表現は簡略化してある。例えば「平成１５年度」と「受刑」の関係は、厳密には and(instance(x, TimeInterval), instance(y, Process), during(BeginFn(y), x)) と表現される。また、memberCount は実際には Person ではなく Collection のインスタンスを項として持つ。

(7) 主流派生成文法は統語論に多くの仕事を与える傾向があるが、別の理論、例えば Sympler Syntax(Culicover and Jackendoff 2009) では統語論の仕事を最小にしてインターフェイスや意味解釈に多くの仕事を割り当てる。名詞述語文研究では、同義性の問題は、特に指定文とウナギ文において中心的な話題の 1 つになっている。指定文については、多くの研究者が「幹事は田中だ」と「田中が幹事だ」ないし「幹事なのは田中だ」との同義性を共通の基底構造からの統語的派生によって説明することを提案している (西山 1985, 西垣内 2016)。ウナギ文については、「僕はウナギだ」をより整形的な構造からの派生として説明する多くの説明が提案されており (奥津 1978)、ウナギ文と整形文の間の同義性問題として捉えることができる。一方で、仁田 (1980) のようにウナギ文の解釈を統語論ではなく意味論の問題として扱う立場もある。

がいかにして表層的な言語表現からその意味解釈へ適切に到達できるかを説明することがより重要な課題となる[8]。標準的な方法の 1 つは形式意味論的な計算に基づく解釈の解決であり、郡司 (2015, 2016) は固有名詞、1 項述語、2 項述語に相当する名詞を含む名詞述語文の意味を形式意味論の手法で分析している。生成語彙論 (Pustejovsky 1998) は、標準的な計算プロセスから外れる多様な意味の生成を形式的に記述するために役立つ。今田 (2012) は、ウナギ文を含むいくつかのタイプの名詞述語文について、意味構成の手続きを生成語彙論を用いて分析している。

本データの意味関係ラベルは be の下位分類という浅いレベルの意味情報に相当するが、意味の理論は言語をいかにして図 18 のような深いレベルの意味情報と結び付け、同義性や推論の問題を解決するかを目標とすべきである。統語と意味の対応を研究することは、言語の複雑さを理論のどの部門に割り当てるかを研究することでもあり、また組み合わせのシステムが言語や認知のどの問題を処理するべきかを明らかにすることにも繋がる。

4.3 まとめ

BCCWJ に対する名詞述語文アノテーションの概要について説明し、記述研究や理論研究との繋がりについて論じた。このデータは名詞述語文に特化した述語項構造情報と意味関係情報、および名詞述語文の周辺的な機能表現に関する情報を含み、各種の文法研究に応用することが期待できる。一方で、本研究で付与した意味関係ラベルは粒度の荒い分類であり、文法研究の様々な目的を満たすためには意味構造の他の層の情報や、より深い解釈レベルの意味との結び付きを分析する必要がある。今後、データの利活用や情報の拡充のための研究を進めたい。

付録: 物理フォーマット

本文とラベル情報を分離したスタンドオフ形式の XML である。s は文、w はラベルを表す。

```
<sample>
  <s start="0">トラはネコ科の肉食動物であるはずだ。</s>
  <w id="0" type="npred" text="肉食動物" start="7" end="11"/>
  <w id="1" type="nsubj" text="トラ" start="0" end="2" head="0"
  ↪  rel="is_a"/>
  <w id="2" type="acl" text="ネコ科" start="3" end="6" head="0" rel="ga"/>
  <w id="3" type="naux" text="はずだ" start="14" end="17"/>
</sample>
```

各ノードの属性は以下の通りである。w 要素の属性のうち、text、start、end 属性はテキスト内に実体を持つノードにのみ付与される。target 属性は type が nsubj、csubj、acl で、テキスト内に実体を持たない (外界照応の) ノードにのみ付与される。head 属性は type が nsubj、csubj、

[8] 「僕はウナギだ」を「僕の注文はウナギだ」のような別の構造に還元するアプローチは、それ自体では意味解釈の説明としての効力を持たない。このような考え方は、統語と意味のインターフェースを簡潔にするためには役立つが、我々がいかにして与えられた入力から「僕の注文はウナギだ」という整形的な構造を復元するかという問題はそのまま残される。

acl のノードにのみ付与される。

表 9　各要素の属性

ノード	属性	説明
s	start	文頭位置 (サンプル頭からの文字数)
w	id	ラベル ID
	type	ラベルの種類 (npred | ncomp | . . .)
	text	ラベル付与範囲の文字列 (外界照応以外)
	start	ラベル開始位置 (サンプル頭からの文字数)
	end	ラベル終了位置 (サンプル頭からの文字数)
	target	外界照応タイプ (exo1 | exo2 | exog | ana_cla)。type = nsubj | csubj | acl のみ。
	head	述語ラベル ID。type = nsubj | csubj | acl のみ。

謝　辞

本研究は JSPS 科研費 17H00009 の助成を受けたものです。

文　献

浅原正幸 (2013) 「係り受けアノテーション基準の比較」,『第 3 回コーパス日本語学ワークショップ予稿集』, 81–90 頁.

Culicover, Peter W and Ray Jackendoff (2009) *Simpler Syntax*, Oxford: Oxford University Press, OCLC: 874574508.

郡司隆男 (2015) 「日本語のコピュラ文の形式意味論的分析」,『トークス= Theoretical and applied linguistics at Kobe Shoin : 神戸松蔭女子学院大学研究紀要言語科学研究所篇』, 第 18 号, 13–24 頁, 3 月.

——— (2016) 「項を 2 つとる名詞コピュラ文の形式意味論的分析」,『トークス= Theoretical and applied linguistics at Kobe Shoin : 神戸松蔭女子学院大学研究紀要言語科学研究所篇』, 第 19 号, 17–28 頁, 3 月.

今田水穂 (2012) 「名詞述語文の生成語彙論的解釈」,『文藝言語研究. 言語篇』, 第 61 号, 83–101 頁.

——— (2017)「外の関係の連体修飾節を伴う名詞述語について」,『言語資源活用ワークショップ 2017 発表論文集』, 74–83 頁.

庵功雄 (2007) 『日本語におけるテキストの結束性の研究』, くろしお出版.

Jackendoff, Ray S. (2002) *Foundations of Language: Brain, Meaning, Grammar, Evolution*: Oxford University Press.

日本語記述文法研究会（編） (2008) 『現代日本語文法 6 複文』, くろしお出版.

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予稿集』, 325–330 頁.

丸山岳彦 (2013) 「BCCWJ に対する節境界ラベルのアノテーション」,『言語処理学会第 19 回年次大会発表論文集』, 154–157 頁.

三上章 (1953) 『現代語法序説: シンタクスの試み』, 刀江書院.

Niles, Ian and Adam Pease (2001) "Towards a Standard Upper Ontology," in *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems*, pp. 2–9.

西垣内泰介 (2016) 「「指定文」および関連する構文の構造と派生」,『言語研究』, 第 150 巻, 137–171 頁.

西山佑司 (1985) 「措定文, 指定文, 同定文の区別をめぐって」,『慶應義塾大学言語文化研究所紀要』, 第 17 巻, 135–165 頁.

——— (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』, ひつじ書房.

仁田義雄 (1980) 『語彙論的統語論』, 明治書院.

奥津敬一郎 (1978) 『「ボクハウナギダ」の文法 : ダとノ』, くろしお出版.

Pustejovsky, James (1998) *The Generative Lexicon*: MIT Press.

新屋映子 (1989) 「"文末名詞"について」,『国語学』, 第 159 号, p88–75 頁.

砂川有里子 (2005) 『文法と談話の接点: 日本語の談話における主題展開機能の研究』, くろしお出版.

竹内孔一 (2015) 「名詞の項構造データの構築」,『第 8 回コーパス日本語学ワークショップ予稿集』, 233–236 頁.

角田太作 (2011) 「人魚構文 : 日本語学から一般言語学への貢献」,『国立国語研究所論集』, 第 1 巻, 53–75 頁.

関連 URL

コーパス検索アプリケーション『中納言』 https://chunagon.ninjal.ac.jp/