日本語名詞述語文への意味情報付与

著者	今田 水穂
雑誌名	国立国語研究所論集
号	8
ページ	51-76
発行年	2014-11
URL	http://doi.org/10.15084/00000542

国立国語研究所論集 (NINJAL Research Papers) 8: 51–76 (2014)
ISSN: 2186-134X print/2186-1358 online

日本語名詞述語文への意味情報付与

今田水穂

文部科学省 初等中等教育局／国立国語研究所 コーパス開発センター プロジェクト研究員［–2014.03］

要旨

日本語名詞述語文に関する既存の記述的研究の集約と共有可能な研究用言語資源の構築を目的として，京都大学テキストコーパスに含まれる名詞述語文に意味情報を付与した。このタスクは i）コーパスの XML 化, ii）4 種類の言語資源（拡張固有表現タグ付きコーパス, CRL 固有表現データ，日本語 WordNet，SUMO）による語義付与，iii）名詞述語文の抽出，iv）主語と述語の意味関係付与の 4 つの下位タスクを含む。アノテーションの結果に基づき，意味関係と語義の共起関係や，名詞述語文の構文的，意味的特徴について検討を行った *。

キーワード：名詞述語文，アノテーション，語義，意味関係，オントロジー

1. はじめに

日本語の名詞述語文にどのような種類のものがあるかについては，多様な観点から多くの研究が行われてきた。古典的な分類としては三上（1953）の措定，指定，端折りの 3 分類がある。

（1）　私は幹事です　　（措定）
（2）　幹事は私です（= 私が幹事です）（指定）
（3）　僕は紅茶だ（注文の場合）　　（端折り）　　（以上，三上 1953: 44–45）

措定と指定の区別は様々な意味論的観点から分析が試みられてきた。名詞の意味の観点からは西山（1988, 2003）の指示叙述理論による分析やメンタル・スペース理論を援用した坂原（1990）の分析がある。高橋（1984）は実例の分析に基づいて主語と述語の意味関係を詳細に記述している。機能論的観点からの研究としては前項焦点，後項焦点，全体焦点の 3 分法に類する分析が取られる場合が多く, 新屋（1994）, 天野（1995）, 砂川（1996, 2005）, 菊地（1997）, および益岡（2000）の叙述類型論における属性叙述，事象叙述，指定叙述の 3 分法などがあるが，分類の仕方は必ずしも一致しない。三上（1953）の端折りはいわゆるウナギ文に相当する文であり，奥津（1978）をはじめとして 1970 年代から 1980 年代にかけて活発に議論され，特に変形統語論的な観点から述語代用説や分裂文説など様々な提案がなされたが，現在でも換喩と関連付けた意味論的分析など研究が続けられている。他には野田（1982）の「カキ料理は広島が本場だ」のような 3 項型の

* 本論文は今田（2013b）および第 106 回 NINJAL サロン「日本語名詞述語文への意味情報付与と言語研究へのフィードバック」（2014 年 3 月）の内容の一部を発展させたものである。本研究の一部は国立国語研究所共同研究プロジェクト「コーパスアノテーションの基礎研究」（プロジェクトリーダー：前川喜久雄）および国立国語研究所コーパス開発センター「超大規模コーパス構築プロジェクト」によるものである。また，本研究は JSPS 科研費 23720225「Ruby と MSXML による日本語名詞述語文の実例調査とコーパス分析ツールの構築」(1)（研究代表者：今田水穂）の助成を受けている。

名詞述語文や，新屋（1989）が文末名詞文，角田（2011）が人魚構文と呼ぶ「太郎は名古屋に行く予定だ」のような構文の研究がある。

既存の研究で提案された様々な次元の意味論的属性を統合的に扱うために，今田（2010）はJackendoff（2002）の意味論モデルを援用した。このモデルでは文は音韻構造（Phonology），統語構造（Syntax），概念構造（Conceptual structure）の 3 部門を並列的に持ち，また概念構造は記述層，指示層，情報構造という 3 つの独立した層を含む。記述層（Descriptive tier），指示層（Referential tier），情報構造（Information structure）はそれぞれ述語項構造，指示と量化，主題や焦点に相当する情報を保持し，異なる部門や層の対応する要素には同一の指標が割り当てられる。全ての層を含む簡潔な記述例が Jackendoff（2002）に示されていないので，情報構造を欠く例を次に示す。

（4） Syntax/phonology: [S [NP a fox]$_1$ [VP ate [NP a grape]$_2$]]$_3$
Descriptive tier: [Event EAT([Object FOX]$_1$, [Object GRAPE]$_2$)]$_3$
Referential tier: 1 2 3 (Jackendoff 2002: 395)

このモデルを用いると「私は幹事です」「私が幹事です」という 2 つの文は記述層や指示層の構成は同じだが情報構造が異なる文として記述することができる。

（5） 統語 / 音韻：私は幹事です
記述層： [State BE([Object 私]$_1$, [Object 幹事]$_2$)]$_3$
指示層： 1 3
情報構造： Topic$_1$ Focus$_2$

（6） 統語 / 音韻：私が幹事です
記述層： [State BE([Object 私]$_1$, [Object 幹事]$_2$)]$_3$
指示層： 1 3
情報構造： Focus$_1$

「私が幹事です」はほぼ同等の機能を持つ「幹事は私です」という文に言い換えることができる。ここでは 2 つの文が何らかの共通の構造を共有する（例えば同一の基底構造から倒置などの統語操作によって派生される）という仮定はせず，両者は記述層においても情報構造においても異なる構成を持つものと考える。両者の同義性は解釈のレベルで保証される。

（7） 統語 / 音韻：幹事は私です
記述層： [State BE([Object 幹事]$_1$, [Object 私]$_2$)]$_3$
指示層： 2 3
情報構造： Topic$_1$ Focus$_2$

本研究の主たる目的は，これらの構造のうち記述層の情報に相当する記述を精緻化することである。Jackendoff（1983）はコピュラ文が表す関係として IS AN INSTANCE OF，IS TOKEN-IDENTICAL TO，IS INCLUDED IN を挙げている。

(8) A dog is a reptile. [IS INCLUDED IN (DOG, REPTILE)]
Clark Kent is Superman. [IS TOKEN-IDENTICAL TO (C. KENT, SUPERMAN)]
Max is a dog. [IS AN INSTANCE OF (MAX, DOG)] (Jackendoff 1983: 95)

Jackendoff（1983）は（理論的観点から）これらを区別する必要は無いという提案をしているが，本稿の関心はこれらを記述的に区別することである。日本語の名詞述語文に関する限り，Jackendoff（1983）が挙げた関係の他に少なくとも次のような関係がある。

(9) 幹事は私だ （主語がクラスで述語がインスタンス）
(10) 太郎は名古屋に行く予定だ （主語と述語が所有的関係）
(11) 僕は紅茶だ （いわゆるウナギ文）

今田（2010）では帰属関係，同一関係，役割・値関係，隣接関係という 4 種の類型を設定したが，十分に精緻な分析，記述とは言えず，より形式的，体系的な分析，記述が課題として残されていた。そのための基礎資料を構築するために，本研究は京都大学テキストコーパス[(2)]（黒橋・長尾 1997，以下京大コーパス）を対象として名詞述語文の悉皆調査と言語情報付与を行った。京大コーパスは毎日新聞 1995 年版のデータに形態論情報，構文情報などを付与した約 4 万文，100 万語規模のコーパスであり，新聞本文 1 日分または社説 1 か月分を 1 文書として，文書，記事，文，文節，語という階層構造を持つ。そのうち約 5 千文，13 万語のデータ（以下コアデータ）には格関係情報，照応・省略情報，共参照情報が付与されており，追加の言語情報を付与するためにタグ単位という要素が文節の下に追加されている。本研究ではこのコアデータを対象として，以下の 4 つのタスクを実施した。

(12) i. 物理フォーマットの変更
ii. 語義付与
iii. 構文抽出
iv. 意味関係付与

i）物理フォーマットの変更は京大コーパスのファイル形式をプログラムで処理しやすい XML 形式に変換する工程である。併せて，後続の処理で必要となる追加の形態論情報などの付与を行った。ii）語義付与はコーパス中の語彙に語義情報を付与する工程である。主語と述語の意味論的なタイプが一致するか否かは意味関係を分類する上で重要な基準であるため，この情報を付与した。iii）構文抽出はコーパス中の名詞述語文を特定し主語と述語に相当するタグ単位のペアを抽出する工程である。iv）意味関係付与は抽出した主語と述語のペアに対して意味関係情報を付与する工程である。次節以降，各工程別に方法の詳細と結果を示す。

2. 物理フォーマットの変更

2.1 方法

京大コーパスは日本語構文・格解析システム KNP[(3)]（笹野他 2013）の出力形式に準拠した物

理フォーマットを持つ。# が文，* が文節，+ がタグ単位を表し，それ以外の行は語を表す。

```
# S-ID:950101003-001 KNP:96/10/27 MOD:2005/03/08
* 0 26D
+ 0 1D
村山 むらやま * 名詞 人名 * *
富市 とみいち * 名詞 人名 * *
+ 1 37D <rel type="=" target="村山富市" sid="950101003-001" tag="0"/>
首相 しゅしょう * 名詞 普通名詞 * *
は は * 助詞 副助詞 * *
* 1 2D
+ 2 3D
年頭 ねんとう * 名詞 普通名詞 * *
に に * 助詞 格助詞 * *
* 2 6D
+ 3 10D <rel type="ニ" target="年頭" sid="950101003-001" tag="2"/><rel type="ガ"
target="不特定:状況"/>
あたり あたり あたる 動詞 * 子音動詞ラ行 基本連用形
```

これをプログラムで処理しやすい XML 形式に変換した。併せて，後続の処理で必要となる追加の形態論情報，言語単位情報の付与を実施した。

```
<?xml version="1.0" encoding="UTF-8"?>
<document id="950101">
  <article id="950101003">
    <sentence id="950101003-001" info="KNP:96/10/27 MOD:2005/03/08">
      <chunk id="0" link="26" rel="D">
        <tag id="0" link="1" rel="D" head_base="村山富市" head_range="0:1">
          <tok id="0" read="むらやま" base="村山" pos="名詞-人名" ctype="*"
                  cform="*" cat="N">村山</tok>
          <tok id="1" read="とみいち" base="富市" pos="名詞-人名" ctype="*"
                  cform="*" cat="N">富市</tok>
        </tag>
        <tag id="1" link="37" rel="D" head_base="首相" head_range="2:2" func_
                base="は" func_range="3:3">
          <rel type="=" target="村山富市" sid="950101003-001" tag="0"/>
          <tok id="2" read="しゅしょう" base="首相" pos="名詞-普通名詞" ctype="*"
                  cform="*" cat="N">首相</tok>
          <tok id="3" read="は" base="は" pos="助詞-副助詞" ctype="*" cform="*"
                  cat="P">は</tok>
```

```
    </tag>
  </chunk>
```

コーパス中の言語単位と KNP 形式，XML 形式における要素の対応を以下に示す。

表 1　KNP 形式と XML 形式の対応

言語単位	KNP 形式	XML 形式
文書	S-ID の 1–6 桁目で表現	document
記事	S-ID の 7–9 桁目で表現	article
文	# で始まる行	sentence
文節	* で始まる行	chunk
タグ単位	+ で始まる行	tag
語	単語で始まる行	tok

tok 要素には，品詞の類別に関する情報を追加付与した。品詞類別は山崎（2014）に倣い N（名詞類），V（動詞類），M（形容詞類），I（接続詞等），P（助詞・助動詞），O（その他）の 6 種類とし，tok 要素の cat 属性として付与した。ただし本研究における品詞類別付与は，後述する語義付与の工程で使用する検索用見出し語の取得を主目的とするため，サ変名詞に後接する「する」を V ではなく P に分類するなど独自の仕様を含む。本研究における品詞類別と京大コーパスの品詞名の対応を以下に示す。

表 2　品詞類別の定義

品詞類別	京大コーパスの品詞名
N	名詞，未定義語，指示詞（名詞形態），特殊（記号），接尾辞（名詞性）
V	動詞
M	形容詞，副詞，連体詞，指示詞（副詞形態，連体詞形態），接頭辞，接尾辞（形容詞性名詞接尾辞）
I	感動詞，接続詞
P	判定詞，助動詞，助詞，(VN-) する，接尾辞（動詞性接尾辞，形容詞性述語接尾辞）
O	特殊（句読点，括弧）

tag 要素には要素内の実質語，付属語範囲に関する情報と，構文関係に関する情報を付与した。実質語に関しては上記の品詞類別で N，V，M，I を実質語構成要素と見なし，付属語に関しては tag 要素内で最初に現れる P 要素以降の N，V，M，I，P を付属語構成要素と見なした。以下に詳細を示す。

表 3　tag 要素の追加属性

tag 要素の追加属性	説明
head_base	実質語範囲の語の出現形を連結し末尾の語を基本形にしたもの
head_range	実質語範囲の tok 要素の id（開始位置：終了位置）
func_base	付属語範囲の語の出現形を連結し末尾の語を基本形にしたもの
func_range	付属語範囲の tok 要素の id（開始位置：終了位置）

2.2 結果

一連の処理により，京大コーパスおよびコアデータの全体を XML 形式に変換した。要素数を以下に示す。

表 4　京都大学テキストコーパスの構造と要素数

要素	京大コーパス	コアデータ
文書	28	6
記事	2927	555
文	38400	5127
文節	372130	50242
タグ単位	–	66186
語	972894	132327

このうち本研究で使用するのはコアデータのみである。コアデータ約 13 万語について品詞類別を集計した結果を以下に示す。

表 5　コアデータの品詞類別

品詞類別	N	V	M	I	P	O	合計
頻度	59002	9679	6519	384	40669	16074	132327

またコアデータに含まれる 66186 タグ単位のうち，66172 要素に head_base 属性と head_range 属性を，35228 要素に func_base 属性と func_range 属性を付与した。

3. 語義付与

3.1 方法

コアデータに含まれる全てのタグ単位を対象として，複数の言語資源を使用して語義付与を実施した。使用した言語資源は拡張固有表現タグ付きコーパス [4]（橋本・中村 2010），CRL 固有表現データ [5]（Sekine and Isahara 2000），日本語 WordNet[6]（Bond et al. 2009），SUMO[7]（Niles and Pease 2001）の 4 種類である。拡張固有表現タグ付きコーパスと CRL 固有表現データは人名などの固有表現に対して語義を付与するために使用した。日本語 WordNet は固有表現以外の一般語彙に対して語義を付与するために使用した。SUMO は前述の 3 種類の言語資源によって得られた語義を統合するために使用した。

3.1.1 拡張固有表現タグ付きコーパスと CRL 固有表現データ

拡張固有表現タグ付きコーパスは毎日新聞などのテキストに対して関根の拡張固有表現階層を付与したものである。固有表現の種類は 243 種類で最大 4 階層まで階層化されている。コアデータの範囲（毎日新聞 95 年 1 月 1 日から 7 日まで）を概ねカバーしているが，一部にデータの欠落が見られる。

CRL 固有表現データは評価型ワークショップ IREX で作成された，毎日新聞 95 年 1 月 1 日から 10 日までの全記事，約 1 万文に対して固有表現をタグ付けしたデータである。固有表現の種類は組織名，人名，地名，固有物名，日付表現，時間表現，金額表現，割合表現の 8 種類である。コアデータの範囲全体をカバーしているが，金額や割合以外の数値表現は含まれていない。

この 2 種類の固有表現データを，tok 要素の ene 属性（拡張固有表現）および ne 属性（CRL 固有表現）としてコアデータに重ね合わせた。構文解析システム CaboCha[8] の出力形式に倣って固有表現の先頭の要素に B-，それ以外の要素に I- という接頭辞を付加した。

```
<tok id="0" read="むらやま" base … ene="B-Person" ne="B-PERSON">村山</tok>
<tok id="1" read="とみいち" base … ene="I-Person" ne="I-PERSON">富市</tok>
```

3.1.2 日本語 WordNet

日本語 WordNet は Princeton WordNet に日本語の語彙を対応付けた概念辞書で，57238 概念（synset），93834 単語（word），158058 語義（概念と単語のペア）が収録されている。また，後述する SUMO 概念名への対応テーブルを含む。この辞書を用いて tag 要素の head_base 属性の値を検索し，得られた概念名を tag 要素の下に wordnet 要素として付与した。wordnet 要素には word（検索語），synset（synset の固有 ID），name（synset 名），sumo（対応する SUMO 概念名），rel（対応する SUMO 概念との関係）を属性として付与した。複数の概念名が得られた場合にはそれらを全て列挙した。

```
<wordnet word="美術" synset="02743547-n" name="fine_art" rel="=" sumo="ArtWork"/>
<wordnet word="美術" synset="06156968-n" name="fine_arts" rel="∈"
     sumo="FieldOfStudy"/>
```

3.1.3 SUMO への統合

上述の 3 種の言語資源から得られた語義情報を SUMO（Suggested Upper Merged Ontology）の概念体系を用いて統合した。SUMO は既存のオントロジーを統合するために開発された上位オントロジーで 25000 程度の概念を含む。拡張固有表現，CRL 固有表現，日本語 WordNet の優先順位で，いずれか 1 つの語義情報を入力として SUMO 概念名を取得し，情報を付与した。

```
<tag id="0" link="1" rel="D" head_base=" … sumo="Entity|Object|Physical">
  <tok id="0" read=" むらやま " ba … ene="B-Person" ne="B-PERSON"> 村山 </tok>
  <tok id="1" read=" とみいち " ba … ene="I-Person" ne="I-PERSON"> 富市 </tok>
  <sumo src="ene" class_name="Human" ancestors="Entity|Object|Physical"/>
</tag>
```

```
<tag id="7" link=" … sumo="Abstract|Entity|Object|Physical|Proposition">
  <rel type=" 修飾 " target=" 装飾 " sid="950101068-003" tag="6"/>
  <tok id="11" read=" びじゅつ " base=" 美術 " pos=" 名詞 - 普通名詞 " … > 美術 </tok>
  <tok id="12" read=" と " base=" と " pos=" 助詞 - 格助詞 " ctype="*" … > と </tok>
  <wordnet word=" 美術 " … name="fine_art" rel="=" sumo="ArtWork"/>
  <wordnet word=" 美術 " … name="fine_arts" rel="∈" sumo="FieldOfStudy"/>
  <sumo src="wn" class_name="ArtWork" type="class" ancestors="Entity|Object|Phy
      sical"/>
  <sumo src="wn" class_name="FieldOfStudy" type="instance" ancestors="Abstract|
      Entity|Proposition"/>
</tag>
```

拡張固有表現と CRL 固有表現については独自に作成した対応テーブル（今田 2014）を用いて SUMO 概念名（クラス名）を取得し，tag 要素の下に sumo 要素として付与した。日本語 WordNet は SUMO へのリンクを収録しているのでこれを用いて SUMO 概念名（クラス名またはインスタンス名）を取得し，tag 要素の下に sumo 要素として付与した。sumo 要素には以下の属性を付与した。

表 6　sumo 要素の属性

sumo 要素の属性	説明	
src	入力情報の種類（ne: CRL 固有表現 ; ene: 拡張固有表現 ; wn: WordNet）	
class_name	SUMO クラス名	
instance_name	SUMO インスタンス名	※入力が WordNet の場合のみ
type	概念のタイプ（クラスかインスタンスか）	※入力が WordNet の場合のみ
ancestors	主要な上位クラス名	

SUMO の概念にはクラスとインスタンスの区別がある。概念がインスタンスの場合，インスタンスに名前が与えられている場合は instance_name 属性を付与し，与えられていない場合は帰属するクラス名を class_name 属性で付与した。ancestors 属性については図 1 に示す 9 種類のクラス名のうち該当するもののみを付与した。

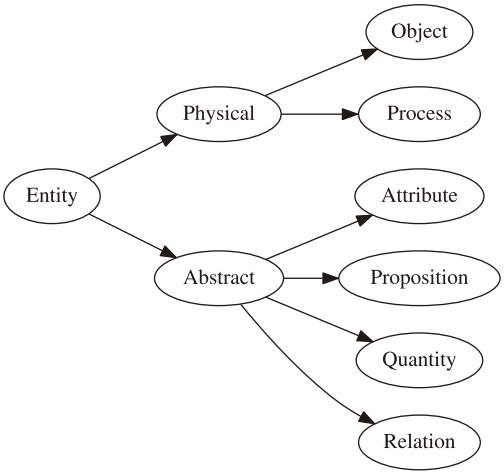

図 1　SUMO の上位クラス名

また，tag 要素中の全ての sumo 要素の ancestors 属性の和集合を，tag 要素の sumo 属性として付与した。

3.2 結果

一連の処理により，9177 件の固有表現情報と 14054 件の拡張固有表現情報をコアデータに付与した。また，コアデータに含まれる 66186 タグ単位のうち 42679 要素に（1 つまたは複数の）wordnet 要素を付与した。これらの語義情報を用いて 66186 タグ単位のうち 53699 要素に（1 つまたは複数の）sumo 要素を付与し，53625 要素に上位クラス名を集約した sumo 属性を付与した。以下に sumo 属性の値の集計結果を示す。多義性などの要因により，上位クラスの頻度と下位クラスの頻度の合計は一致しない。

表 7　タグ単位の持つ意味タイプ

要素数	Enti.	Phys.	Obje.	Proc.	Abst.	Attr.	Prop.	Quan.	Rela.
66186	53625	37651	20000	20349	31313	17566	2119	9184	14069

4. 構文抽出と意味関係付与

4.1 方法

京大コーパスの形態論情報，構文情報，省略情報を利用して名詞述語文の抽出を実施し，抽出した主語と述語のペアに対して意味関係の判定，付与を実施した。

4.1.1 構文抽出

名詞述語文の抽出については，まず述語要素の特定を行い，次に各述語と対応する主語要素の特定を行った。述語要素は tag 要素の下に copula 要素を付与することによって表現し，主語要素

は copula 要素の下に arg 要素を付与することによって表現した。

```
<tag id="2" link="-1" rel="D" head_base="亥" head_range="5:5" conj="N">
  <rel type="ガ" target="えと" sid="950101019-001" tag="1"/>
  <memo>CO</memo>
  <tok id="5" read="い" base="亥" pos="名詞-普通名詞" ctype="*"  … >亥</tok>
  <tok id="6" read="。" base="。" pos="特殊-句点" ctype="*" cfor … >。</tok>
  <copula copula_omitted="true" head_base="亥">
    <arg tag="1" type="ガ" head_base="えと"/>
  </copula>
</tag>
```

述語要素は規範的には判定詞「だ」「です」「である」を伴うが, 実例では伴わない場合も多い。京大コーパスのコアデータには判定詞の省略情報（<memo>CO</memo>）が付与されているため, 判定詞を含むか，または判定詞の省略情報を含むタグ単位を述語要素として特定した。ただし省略情報 CO はナ形容詞，ナノ形容詞の語尾の「だ」が省略されている場合にも付与されているため，実質語範囲の末尾が形容詞語幹であるものについては抽出対象から除外した。述語要素を表す copula 要素には以下の属性を付与した。

表 8　copula 要素の属性

copula 要素の属性	説明
copula_omitted	判定詞の省略
head_base	実質語範囲の基本形

主語要素についてはコアデータの格関係情報（<rel …/>）を使用して特定を行った。抽出する主語は述語要素と対応するガ格のタグ単位とし，該当するタグ単位が述語と同一文中に含まれる場合のみを抽出対象とした。京大コーパスの格関係情報ではガ格に相当する格関係ラベルとして「判ガ」「ガ」「ガ２」の 3 種類が設定されている。「判ガ」はサ変名詞＋判定詞などのパターンに対して使用されるラベルで，サ変名詞のガ格が「ガ」で，名詞＋判定詞のガ格が「判ガ」で表示される。「ガ２」はガ格要素が 2 つある場合に使用されるラベルで, 述語から遠くにある要素が「ガ２」で表示される。

(13)　朗読は女優の仕事である。（判ガ：朗読，ガ：女優）
(14)　男子は十二位以内が安全圏だ。（ガ２：男子，ガ：十二位以内）

本研究では「判ガ」がある場合にはこれを主語要素とし,「判ガ」がない場合には「ガ」「ガ２」を主語要素とした。主語要素を表す arg 要素には以下の属性を付与した。

表 9　arg 要素の属性

arg 要素の属性	説明
tag	対応するタグ単位の id
type	「判ガ」「ガ」「ガ２」の区別
head_base	実質語範囲の基本形

4.1.2 意味関係付与

抽出した主語と述語のペアに対して両者の意味関係を人手で評定し，arg 要素の属性として付与した。

```
<copula base="仕事" ancestors="Abstract|Attribute|ChangeOfPossessio … ">
  <arg tag="0" type="判ガ" base="朗読" … const="NORMAL" rel="SORT-SUCC"/>
</copula>
```

意味関係の分類は評定の難度を鑑み，まず構文パターンをいくつかに分類した上で，意味関係の分類を複数の階層に分けて設定した。構文パターンとしては，通常の文，分裂文，除外事例の 3 種を区別し，const 属性として付与した。便宜的に，分裂文の分類には主語がノ節であるものを全て含めた。

表 10　構文パターンの類別

構文パターン	説明
NORMAL	通常の文（分裂文以外の文）
CLEFT	分裂文など（主語がノ節の文）
IGNORED	除外事例（名詞述語文ではないもの）

次に一般の文について，主語と述語の意味関係を rel 属性として付与した。上位分類として，同じタイプの概念の上位下位関係などを表しているものと，それ以外のものを区別した。

表 11　意味関係の類別

意味関係	説明
SORT	上位下位関係，内包外延関係，同一関係など同じタイプの概念間の関係
NSORT	SORT ではない関係

次に SORT の下位分類として，主語と述語のどちらが上位概念かに基づく分類を付与した。

表 12　意味関係の類別（SORT の下位分類）

意味関係	説明	例
SORT-SUCC	主語が下位概念または外延（A≻B）	吾輩は猫である（吾輩 ≻ 猫）
SORT-PREC	主語が上位概念または内包（A≺B）	今年の干支は亥（今年の干支 ≺ 亥）
SORT-EQ	主語と述語が同一関係（A=B）	ジキルはハイドだ（ジキル = ハイド）

NSORTについては主語と述語主名詞の間に属格関係が成立するか否かで下位分類を付与した。

表 13　意味関係の類別（NSORT の下位分類）

意味関係	説明	例
NSORT-GEN	主語と述語が属格関係（A の B）	私は外出する予定だ（私の予定）
		塀は 3m の高さだ（塀の高さ）
NSORT-NGEN	主語と述語が非属格関係	僕はウナギだ（# 僕のウナギ）
		塀は 3m だ（* 塀の 3m）

4.2 結果

一連の処理により，コアデータに含まれる 66186 タグ単位のうち 1912 要素に copula 要素を付与した。このうち 990 要素は判定詞を伴い，922 要素は判定詞が省略されていた。これらの copula 要素は 0 〜 5 個の arg 要素を持っていた。内訳を以下に示す。

表 14　copula 要素数（arg 要素数別集計）

arg 要素数	0	1	2	3	4	5	合計
頻度	545	1316	44	5	1	1	1912

arg 要素の総数は 1428 だった。cleft 属性, rel 属性, type 属性の値別に集計した結果を以下に示す。ガ 2 は全ての事例を IGNORED としたが，大半は「曙は琴の若が相手」（cf. 琴の若が曙の相手）の「曙は」など主語か述語の属格修飾語が主題化したものだった。

表 15　arg 要素数（属性別集計）

const: rel ＼ type	ガ	ガ 2	判ガ	合計
NORMAL:	1071	0	57	1128
SORT-SUCC	406	0	12	418
SORT-PREC	210	0	21	231
SORT-EQ	5	0	0	5
NSORT-GEN	82	0	5	87
NSORT-NGEN	368	0	19	387
CLEFT:	177	0	22	199
IGNORED:	82	19	0	101
合計	1330	19	79	1428

5. 考察

5.1 量的分析

5.1.1 語の多義性

日本語 WordNet を使用した語義付与では曖昧性の解消を行わず得られた語義を全て列挙したため，結果として SUMO クラス名も，理論的には相互排他的な複数のクラス名が列挙されている場合がある。どのクラス名の間に曖昧性が生じやすいかを調べるため，SUMO の第 3 レベルの 6 つのクラス名について共起関係を集計した。また，それらの共起関係の強さを評価するために MI 値を計算した。MI 値は $\log_2$（実測値 / 期待値）で求められる。一般的には次の式を用いる。

$$MI = \log_2 \frac{f(x,y) \times N}{f(x) \times f(y)}$$

ここでは N= 全タグ単位数（要素数），f(x)= 語義 x を持つタグ単位数，f(y)= 語義 y を持つタグ単位数，f(x,y)= 語義 x と y を持つタグ単位数とする。N，f(x)，f(y) は表 7 の数値を用い，f(x,y) はここで集計した頻度を用いる。結果を以下に示す。

表 16　同一タグ単位における語義の共起関係［頻度（MI 値）］

	Process	Attribute	Proposition	Quantity	Relation
Object	2940 (–1.06)	3755 (–0.50)	647 (0.01)	991 (–1.49)	2905 (–0.55)
Process		7871 (0.54)	1023 (0.65)	970 (–1.54)	8877 (1.04)
Attribute			983 (0.81)	1804 (–0.43)	7969 (1.09)
Proposition				80 (–1.88)	889 (0.98)
Quantity					1732 (–0.17)

これを見ると，Process，Attribute，Proposition，Relation は相互に共起する場合の MI 値が 0.54 ～ 1.09 と比較的高く，これらのクラスの間で曖昧性が生じやすいことが分かる。これに対して，Object や Quantity は他のクラスと共起する場合の MI 値は大半が負の値を取るなど比較的低く，他のクラスとの間で曖昧性を生じることは少ないことが分かる。

5.1.2 主語 / 述語と意味タイプ

以下では構文パターンが NORMAL である主語と述語のペア 1128 例を対象として，意味関係と語義に関する統計量を示し，考察を行う。最初に主語と述語の意味関係と，主語または述語が持つ意味タイプを交差集計し，その MI 値を計算した。N= 全タグ単位数，f(x)= 語義 x を持つタグ単位数，f(y)= 意味関係 y の名詞述語文数，f(x,y)= 主語 / 述語が語義 x を持つ意味関係 y の名詞述語文数とする。N，f(x) は表 7，f(y) は表 15 の「合計」，f(x,y) はここで集計した頻度を用いる。結果を以下に示す。

表 17　主語の意味タイプ［頻度（MI 値）］

主語	Enti.	Phys.	Obje.	Proc.	Abst.	Attr.	Prop.	Quan.	Rela.
SUCC	366	250	131	131	168	93	10	49	70
	(0.11)	(0.07)	(0.05)	(0.03)	(−0.24)	(−0.25)	(−0.42)	(−0.24)	(−0.34)
PREC	193	109	65	70	180	81	12	78	104
	(0.04)	(−0.27)	(−0.1)	(−0.02)	(0.72)	(0.4)	(0.7)	(1.28)	(1.08)
EQ	4	3	1	2	1	1	0	0	0
	(−0.02)	(0.08)	(−0.6)	(0.38)	(−1.24)	(−0.41)	(−)	(−)	(−)
GEN	74	48	39	13	33	16	1	8	11
	(0.07)	(−0.04)	(0.57)	(−1.04)	(−0.32)	(−0.53)	(−1.48)	(−0.59)	(−0.75)
NGEN	318	214	142	90	174	81	11	58	71
	(0.02)	(−0.04)	(0.28)	(−0.4)	(−0.07)	(−0.34)	(−0.17)	(0.11)	(−0.21)
合計	955	624	378	306	556	272	34	193	256
	(0.06)	(−0.04)	(0.15)	(−0.18)	(0.06)	(−0.14)	(−0.09)	(0.3)	(0.09)

表 18　述語の意味タイプ［頻度（MI 値）］

述語	Enti.	Phys.	Obje.	Proc.	Abst.	Attr.	Prop.	Quan.	Rela.
SUCC	371	243	181	105	301	208	24	91	109
	(0.13)	(0.03)	(0.52)	(−0.29)	(0.61)	(0.91)	(0.84)	(0.65)	(0.29)
PREC	207	94	62	40	148	38	6	107	34
	(0.15)	(−0.48)	(−0.17)	(−0.83)	(0.44)	(−0.69)	(−0.3)	(1.74)	(−0.53)
EQ	3	2	1	1	1	1	0	0	0
	(−0.43)	(−0.51)	(−0.6)	(−0.62)	(−1.24)	(−0.41)	(−)	(−)	(−)
GEN	83	61	41	41	78	55	36	8	45
	(0.24)	(0.3)	(0.64)	(0.62)	(0.92)	(1.25)	(3.69)	(−0.59)	(1.28)
NGEN	268	91	61	42	227	44	4	167	40
	(−0.23)	(−1.27)	(−0.94)	(−1.5)	(0.31)	(−1.22)	(−1.63)	(1.64)	(−1.04)
合計	932	491	346	229	755	346	70	373	228
	(0.03)	(−0.39)	(0.02)	(−0.6)	(0.5)	(0.21)	(0.95)	(1.25)	(−0.07)

各数値の解釈については 5.2 節で検討する。

5.1.3 主語の意味タイプと述語の意味タイプ

意味関係別に，主語と述語の意味タイプを交差集計し，その MI 値を計算した。N= 意味関係 z の名詞述語文数，f(x)= 主語が語義 x を含む意味関係 z の名詞述語文数，f(y)= 述語が語義 y を含む意味関係 z の名詞述語文数，f(x,y)= 主語が語義 x，述語が語義 y を含む意味関係 z の名詞述語文数とする。N は表 15 の「合計」，f(x) は表 17，f(y) は表 18，f(x,y) はここで集計した頻度を用いる。結果を以下に示す（SORT-EQ はサンプル数が少なく適切な MI 値が得られないことが予想されたため頻度のみを示す）。

表 19　主語と述語の意味タイプ（SORT-SUCC: 主語が下位概念または外延）［頻度（MI 値）］

主＼述	Enti.	Phys.	Obje.	Proc.	Abst.	Attr.	Prop.	Quan.	Rela.
Enti.	330	212	161	91	270	185	21	80	99
	(0.02)	(–0.01)	(0.02)	(–0.01)	(0.03)	(0.02)	(0)	(0.01)	(0.05)
Phys.	226	170	126	78	183	134	17	47	75
	(0.03)	(0.23)	(0.22)	(0.31)	(0.02)	(0.11)	(0.24)	(–0.21)	(0.2)
Obje.	122	86	81	21	92	70	5	17	29
	(0.07)	(0.18)	(0.51)	(–0.65)	(–0.04)	(0.1)	(–0.59)	(–0.75)	(–0.24)
Proc.	115	92	52	61	100	71	12	32	50
	(–0.02)	(0.27)	(–0.13)	(0.89)	(0.08)	(0.12)	(0.67)	(0.17)	(0.55)
Abst.	148	95	65	52	129	80	10	50	60
	(–0.01)	(–0.04)	(–0.16)	(0.3)	(0.09)	(–0.06)	(0.05)	(0.45)	(0.45)
Attr.	83	58	41	31	74	54	6	20	39
	(0.01)	(0.1)	(0.03)	(0.41)	(0.14)	(0.22)	(0.17)	(–0.02)	(0.69)
Prop.	9	6	6	0	9	8	1	6	1
	(0.02)	(0.05)	(0.47)	(–)	(0.32)	(0.68)	(0.8)	(1.46)	(–1.38)
Quan.	45	18	14	7	37	16	2	24	13
	(0.05)	(–0.66)	(–0.6)	(–0.81)	(0.07)	(–0.61)	(–0.49)	(1.17)	(0.02)
Rela.	59	48	27	31	56	38	3	18	33
	(–0.07)	(0.24)	(–0.17)	(0.82)	(0.15)	(0.13)	(–0.42)	(0.24)	(0.85)

表 20　主語と述語の意味タイプ（SORT-PREC: 主語が上位概念または内包）［頻度（MI 値）］

主＼述	Enti.	Phys.	Obje.	Proc.	Abst.	Attr.	Prop.	Quan.	Rela.
Enti.	180	84	53	37	129	36	6	92	31
	(0.06)	(0.1)	(0.03)	(0.15)	(0.06)	(0.18)	(0.26)	(0.04)	(0.13)
Phys.	102	59	38	25	71	27	5	41	26
	(0.06)	(0.41)	(0.38)	(0.41)	(0.02)	(0.59)	(0.82)	(–0.3)	(0.7)
Obje.	63	42	29	16	39	18	2	19	17
	(0.11)	(0.67)	(0.73)	(0.51)	(–0.09)	(0.75)	(0.24)	(–0.66)	(0.83)
Proc.	65	38	23	17	44	17	3	26	16
	(0.05)	(0.42)	(0.29)	(0.49)	(–0.03)	(0.56)	(0.72)	(–0.32)	(0.64)
Abst.	169	74	46	33	123	32	4	92	26
	(0.07)	(0.01)	(–0.07)	(0.08)	(0.09)	(0.11)	(–0.23)	(0.14)	(–0.03)
Attr.	75	45	26	22	47	22	0	27	17
	(0.05)	(0.45)	(0.26)	(0.65)	(–0.14)	(0.72)	(–)	(–0.47)	(0.51)
Prop.	11	9	4	5	8	5	1	3	4
	(0.03)	(0.88)	(0.31)	(1.27)	(0.06)	(1.34)	(1.68)	(–0.89)	(1.18)
Quan.	75	7	3	6	70	4	1	67	4
	(0.1)	(–2.18)	(–2.8)	(–1.17)	(0.49)	(–1.68)	(–1.02)	(0.89)	(–1.52)
Rela.	98	51	36	20	68	20	2	52	17
	(0.07)	(0.27)	(0.37)	(0.15)	(0.03)	(0.23)	(–0.43)	(0.11)	(0.15)

表 21　主語と述語の意味タイプ（SORT-EQ: 主語と述語が同一関係）［頻度のみ］

主＼述	Enti.	Phys.	Obje.	Proc.	Abst.	Attr.	Prop.	Quan.	Rela.
Enti.	3	2	1	1	1	1	0	0	0
Phys.	2	2	1	1	1	1	0	0	0
Obje.	1	1	1	0	0	0	0	0	0
Proc.	1	1	0	1	1	1	0	0	0
Abst.	1	1	0	1	1	1	0	0	0
Attr.	1	1	0	1	1	1	0	0	0
Prop.	0	0	0	0	0	0	0	0	0
Quan.	0	0	0	0	0	0	0	0	0
Rela.	0	0	0	0	0	0	0	0	0

表 22　主語と述語の意味タイプ（NSORT-GEN: 主語と述語が属格関係）［頻度（MI 値）］

主＼述	Enti.	Phys.	Obje.	Proc.	Abst.	Attr.	Prop.	Quan.	Rela.
Enti.	71	51	34	36	66	45	32	6	35
	(0.01)	(–0.02)	(–0.04)	(0.05)	(–0.01)	(–0.06)	(0.06)	(–0.18)	(–0.13)
Phys.	46	34	23	24	43	30	23	4	22
	(0.01)	(0.01)	(0.02)	(0.09)	(0)	(–0.02)	(0.21)	(–0.14)	(–0.17)
Obje.	38	28	20	20	36	24	18	4	18
	(0.03)	(0.03)	(0.12)	(0.12)	(0.04)	(–0.04)	(0.16)	(0.16)	(–0.16)
Proc.	12	8	4	5	10	7	5	3	6
	(–0.05)	(–0.19)	(–0.62)	(–0.29)	(–0.22)	(–0.23)	(–0.11)	(1.33)	(–0.16)
Abst.	30	19	10	12	25	18	9	4	16
	(–0.07)	(–0.28)	(–0.64)	(–0.37)	(–0.24)	(–0.21)	(–0.6)	(0.4)	(–0.09)
Attr.	16	9	6	5	14	9	4	3	9
	(0.07)	(–0.32)	(–0.33)	(–0.59)	(–0.04)	(–0.17)	(–0.73)	(1.03)	(0.12)
Prop.	0	0	0	0	0	0	0	0	0
	(–)	(–)	(–)	(–)	(–)	(–)	(–)	(–)	(–)
Quan.	7	5	2	3	7	6	2	1	5
	(–0.12)	(–0.17)	(–0.91)	(–0.33)	(–0.04)	(0.25)	(–0.73)	(0.44)	(0.27)
Rela.	10	7	3	5	6	5	4	1	3
	(–0.07)	(–0.14)	(–0.79)	(–0.05)	(–0.72)	(–0.48)	(–0.19)	(–0.02)	(–0.92)

表 23　主語と述語の意味タイプ（NSORT-NGEN: 主語と述語が非属格関係）［頻度（MI 値）］

主＼述	Enti.	Phys.	Obje.	Proc.	Abst.	Attr.	Prop.	Quan.	Rela.
Enti.	225 (0.03)	74 (–0.02)	47 (–0.09)	35 (0.02)	188 (0.01)	33 (–0.13)	3 (–0.13)	145 (0.08)	26 (–0.34)
Phys.	160 (0.11)	48 (–0.07)	33 (–0.03)	22 (–0.08)	138 (0.14)	25 (0.04)	2 (–0.15)	107 (0.21)	16 (–0.47)
Obje.	110 (0.16)	32 (–0.06)	23 (0.04)	14 (–0.14)	97 (0.22)	17 (0.07)	1 (–0.55)	74 (0.27)	11 (–0.42)
Proc.	63 (0.02)	17 (–0.32)	11 (–0.37)	7 (–0.48)	52 (–0.02)	7 (–0.55)	1 (0.1)	45 (0.21)	5 (–0.9)
Abst.	119 (–0.02)	40 (–0.03)	26 (–0.08)	17 (–0.15)	96 (–0.09)	11 (–0.85)	1 (–0.85)	76 (0.02)	17 (–0.08)
Attr.	53 (–0.08)	20 (0.07)	14 (0.13)	9 (0.03)	45 (–0.08)	9 (–0.03)	1 (0.26)	28 (–0.32)	13 (0.63)
Prop.	8 (0.07)	4 (0.63)	4 (1.21)	0 (–)	8 (0.31)	1 (–0.32)	0 (–)	4 (–0.25)	6 (2.4)
Quan.	41 (0.03)	16 (0.23)	8 (–0.19)	8 (0.35)	30 (–0.18)	1 (–2.72)	0 (–)	27 (0.11)	2 (–1.58)
Rela.	42 (–0.23)	11 (–0.6)	6 (–0.9)	6 (–0.36)	36 (–0.21)	3 (–1.43)	0 (–)	30 (–0.03)	5 (–0.55)

各数値の解釈については 5.2 節で検討する。

5.2 質的分析

5.2.1 SORT-SUCC

SORT-SUCC は主語が述語の表すクラスに属する関係である。主語は述語が表すクラスのインスタンスである場合と下位クラスである場合があるがここでは区別しない。

（15）　吾輩は猫である　　（インスタンスとクラス）［作例］
（16）　猫は哺乳動物である　　（下位クラスと上位クラス）［作例］

述語はクラス概念であり，通常の述語論理では 1 項述語に相当する。事物の種類以外の性質（色や形など）を表す 1 項述語と区別して sortal property と呼ばれる場合もある。述語の主要部は 1 項述語相当の場合と 2 項以上の述語相当の場合がある。主要部が 1 項述語相当の場合，(15)の「猫」のように単独で用いられる場合もあるが，クラスをさらに限定する連体修飾句や節を伴う場合も多い。以下，特に断りのない限り例文はコーパスからの引用である。

（17）　赤米は赤褐色のうるち米　　赤褐色の（修飾句）＋うるち米（1 項述語）
（18）　この化合物は土中に残留する物質で　　土中に残留する（修飾節）＋物質（1 項述語）

主要部が 2 項以上の述語相当の場合は連体修飾句により項の一部が充足され 1 項述語相当の名詞句が構成される。連体修飾句の代わりに主題や文脈によって項が充足される場合もある。

(19) 景洪はタイ人の故郷だ　タイ人の（修飾句）＋故郷（2 項述語）
(20) 曙は琴の若が相手　曙の（主題）＋相手（2 項述語）
(21) 師である楠部彌弌　今井さんの（文脈）＋師（2 項述語）

主要部は複合語である場合も多く，その場合も 1 項述語相当の場合や 2 項以上の述語相当の場合がある。1 項述語相当の複合語は「修飾語＋ 1 項述語」「項＋ 2 項述語」などのパターンがある。

(22)「アメリカ号」は木造船で　木造（修飾語）＋船（1 項述語）
(23) インドは英国領土　英国（項）＋領土（2 項述語）

2 項述語相当の複合語としては「2 項述語＋ 1 項述語」というパターンが見られる。

(24) 中国の中央銀行である中国人民銀行　中央（2 項述語）＋銀行（1 項述語）
(25) 欧州連合の議長国であるフランス　議長（2 項述語）＋国（1 項述語）
(26) 最大政党である自民党　最大（2 項述語）＋政党（1 項述語）

表 17 と表 18 を見ると，主語の意味タイプについては Proposition の MI 値がやや低い以外に顕著な特徴は見られないが，述語の意味タイプは Object，Attribute，Proposition，Quantity の MI 値が高いことが分かる。この分布が意味するところは明確でないが，Quantity については「〜の一つ」「〜の一人」などを SORT-SUCC に分類したことが影響しているものと考えられる。

また表 19 で SUMO の第 3 レベルのクラス（Object，Process，Attribute，Proposition，Quantity，Relation）を見ると，同じクラス名が交差する部分の MI 値は Attribute を除いて 0.5 以上と高い数値を示しており，Attribute も 0.22 と正の値を示している。これは SORT 型は主語と述語の意味タイプが一致するという予測を支持する。それ以外の組み合わせでは Process，Attribute，Proposition，Relation などの間で MI 値が高くなっている部分が多い。これらのクラスは 5.1.1 節で示したように相互に曖昧性を持つ場合が多く，それが主語と述語の MI 値に影響しているものと考えられる。ただし Proposition と Relation の組み合わせの MI 値は低く，検討が必要である。

それ以外に MI 値が高い組み合わせとしては主語が Proposition で述語が Quantity という組み合わせがある。この組み合わせは 6 例あるがそのうち 5 例は述語が「もの」であり，日本語 WordNet が「もの」に割り当てる語義の中に one という概念が含まれていることが影響しているものと考えられる。

5.2.2 SORT-PREC

SORT-PREC は述語が主語の表すクラスに属する関係である。主語は 1 項述語に相当するが「猫」のような単純な範疇語が単独で生起することはほとんどなく，多くの場合は 2 項以上の述語相当の語を主要部とする名詞句が生起する。項の一部が修飾句，主題，文脈などによって補われ，1 項述語相当の名詞句が構成される。

(27) 地震の規模はマグニチュード３・９　地震の（修飾句）＋規模（2 項述語）

(28) ジュネーブは人口わずか１６万人　　ジュネーブの（主題）＋人口（2 項述語）
(29) 震源地は宮城県沖で　　この地震の（文脈）＋震源地（2 項述語）

次の例では主語の主要部が 1 項述語で，連体修飾節による限定が加えられている。このような例では修飾部に「最も」「最大の」などの語句が含まれることが多いように思われるが，本研究の範囲では十分なデータが得られていない。

(30) 最も初期から香港に居住する外国人はインド人で
最も初期から香港に居住する（修飾節）＋外国人（1 項述語）

表 17 を見ると，主語の意味タイプは Attribute，Proposition，Quantity，Relation など抽象概念の MI 値が高く，SORT-SUCC の述語に抽象概念が多かったのと対照的である。一方，表 18 からは述語の意味タイプとして Process，Attribute，Relation などの MI 値が低く，Quantity の MI 値が高いことが分かる。

表 20 を見ると，主語と述語がどちらも Object，Process，Attribute，Proposition，Quantity のいずれかを含む場合の MI 値は 0.49 以上と高く, Relation 同士の場合も 0.15 と正の値を示している。SORT-SUCC の場合と同様，主語と述語に同じ意味クラスの語が生起する傾向が強いと言える。それ以外の組み合わせでは Process，Attribute，Proposition，Relation に加えて Object も他のクラスと共起する際の MI 値が比較的高い。この理由は現在のところよく分かっていない。Quantity は他のどのクラスとの MI 値も低い。

5.2.3 SORT-EQ

SORT-EQ は主語と述語が同一概念という関係である。現在のところ，この類型を適用するための厳密な基準は定まっておらず，SORT-SUCC と SORT-PREC のいずれにも分類しがたいものをこの類型に含めている。そのため，適用事例は 5 例のみと極端に少ない。この類型に含まれる事例には次の種類のものがある。

(31) インスタンス同一型
a.「フジタ工業」でない，神奈川県平塚市をホームタウンとする「ベルマーレ平塚」
(32) クラス同一型
a. インサイダー取引が何なのか知らないまま
b. 簡体字ではない繁体字
(33) トートロジー型
a. 連立政権は政権，選挙は選挙だ（2 例含む）

この類型の適用基準とこれらの事例の扱いについては今後の検討課題とする。

5.2.4 NSORT-GEN

NSORT-GEN は主語 N_1 と述語 N_2 を「N_1 の N_2」と言い換え可能なタイプである。述語 N_2 は通常，連体修飾要素 M を伴うが，M と N_2 は内の関係である（N_2 が M の項である）場合と外の関係である（N_2 が M の項ではない）場合がある。

（34） 内の関係

a. うちはそれほど高いレベルではない （レベルが高い）
b. 優勝戦に匹敵する内容の好試合 （内容が優勝戦に匹敵する）
c. 内部資料は五〇年十月の日付 （日付が五〇年十月だ）
d. 簡易電話は１日に約７０人の利用者 （利用者が１日に約７０人だ）

（35） 外の関係

a. 日弁連も実態調査に乗り出す方針 （* 方針が実態調査に乗り出す）
b. 容疑者は山下さんを殺した疑い （* 疑いが山下さんを殺した）
c. メンバーは四日に下山する予定だった （* 予定が四日に下山する）

本研究は変形文法的な立場を取るものではないが，便宜的に変形的な説明をすると，内の関係の文は [$_M$ ...N_1 の N_2...] という文の N_1 が主題化，N_2 が非連体修飾語化したものに相当し，外の関係の文は [$_M$ ...N_1...] が内容補充修飾節として N_2 にかかり，N_1 が主題化したものに相当する。

（36） [$_M$ うち $_1$ のレベル $_2$ はそれほど高く] ない
→ うち $_1$ は [$_M$ それほど高い] レベル $_2$ ではない

（37） [$_M$ 日弁連 $_1$ が実態調査に乗り出す]
→日弁連 $_1$ も [$_M$ 実態調査に乗り出す] 方針 $_2$

表 17 を見ると主語が持つ意味タイプの MI 値は Object が 0.57 と比較的高く，第 3 レベルの他の意味クラスはいずれも -0.53 以下と低い。また表 18 を見ると述語が持つ意味タイプの MI 値は Proposition が 3.69 と特に高く，Object，Process，Attribute，Relation も 0.62 〜 1.28 と比較的高いが，Quantity は -0.59 と低い。

SORT 型では主語と述語の意味クラスが一致することが予測されるのに対して，NSORT 型は主語と述語の意味クラスは（偶然一致する場合もあるが）基本的に独立であることが予測される。表 22 を見ると，主語と述語がいずれも第 3 レベルの同じ意味クラスを持つ場合の MI 値は Quantity を除いて -0.92 〜 0.12 と比較的低く，予測を支持する結果が得られた。Quantity については主語と述語がいずれも Quantity の場合の MI 値が 0.44 と比較的高く，また主語が Process か Attribute で述語が Quantity の場合の MI 値もそれぞれ 1.33，1.03 と高いが，すぐ上で述べたようにそもそも述語の意味クラスに Quantity が含まれる場合が少ないため，数値の信頼性に疑問が残る（MI 値はデータ数が少ないと数値が跳ね上がる傾向がある）。第 3 レベルの他の組み合わせの MI 値を見ても -0.91 から 0.4 と比較的低く大半は負の値であり，一部に強い負の相関が見

られる以外は概ね主語と述語の意味タイプは独立であると考えられる。

5.2.5 NSORT-NGEN

NSORT-NGEN は SORT 型ではなく（主語と述語の意味クラスが一致せず），NSORT-GEN でもない（主語と述語が属格的関係にない），いわばその他の類型である。この類型は多様な事例を含むため今後さらに整理する必要があるが，例としては述語が範疇以外の属性（non-sortal property）であるもの，述語が数量であるもの，主語と述語がウナギ文の関係にあるものなどが含まれる。

(38) 湯は茶褐色 （範疇以外の属性）
(39) 天下りした課長級以上の官僚は２０７人 （数量）
(40) 事件では，二三年は関東大震災 （ウナギ文：日付と事件）

(38)の「茶褐色」は通常の述語論理では1項述語として扱われる(= 茶褐色(湯))。(39)の「２０７人」や（40）の「関東大震災」は述語論理では述語として扱われないので，主語と述語を関係付ける何らかの述語 P が解釈によって補われるという扱いになる（=P（官僚，２０７人），P（二三年，関東大震災））。P がどのような関係を表しているかの解釈を決定する入力はいくつかの種類のものが考えられる。1 つは主語と述語の意味，およびそれに関連する百科事典的知識であり，例えば主語が事物で述語が個数であれば，標準的な解釈は主語の存在量を述語が表しているというものである。述語が個数ではなく長さや年齢であれば，それに応じた解釈が与えられる。別の場合は文脈や状況などが解釈を決定する。例えば「僕はウナギだ」という文が客と注文の関係を表しているということは，この文が料理を注文する場面で発話されるという状況に依存して解釈される（今田（2012）では生成語彙論の考え方を利用して解釈の一部を解決する方法を検討している）。

表 17 と表 18 を見ると主語の意味クラスについては顕著な特徴が見られないが，述語の意味クラスは Quantity の MI 値が 1.64 と高く，他の第 3 レベルの意味クラスは -1.63 〜 -0.94 と低い。すなわち，得られたデータとしては事物や概念の数量を述べるタイプの文が多い。また，Quantity のみが多いという分布は Quantity のみが少ないという NSORT-GEN の分布と対照的である。

また表 23 を見ると主語と述語が同じ意味クラスを含む場合の MI 値は -0.55 〜 0.11 と比較的低く，NSORT-GEN の場合と同様，主語と述語の意味クラスは一致しない傾向がある。それ以外の組み合わせでは主語が Proposition で述語が Object か Relation である場合の MI 値がそれぞれ 1.21，2.4 と高いが，そもそも主語の意味クラスに Proposition が含まれる場合が多くないので信頼性に疑問が残る。

6. おわりに

本論文では京大コーパスに対する語義付与，名詞述語文抽出，意味関係付与の結果について

報告し，名詞述語文の主語と述語の語義と意味関係の相関について検討した。語義については Process，Attribute，Proposition，Relation の間で曖昧性が生じる場合が多かった。主語と述語の意味関係については意味関係が SORT の場合は主語と述語が同じ意味クラスの語義を共有する場合が多く，NSORT の場合はその傾向が見られなかった。本研究の最終的な目的である名詞述語文の意味構造の形式的，体系的記述は十分に果たされたと言えないが，本研究で構築した言語資料とその過程で蓄積したコーパス構築のノウハウは，今後の名詞述語文研究に資する。

今後の課題としては i）意味情報の精緻化，ii）他の次元の言語情報の付与，iii）他のコーパスへの意味付与が挙げられる。本稿では語義と意味関係という 2 種類の意味情報を扱った。語義について本稿では曖昧性解消を行わず可能な解釈を全て列挙するという方策を取ったが，今後は語義曖昧性解消，語義の多相性の扱い（「男性」「学生」など属性を表す語の一部はその属性を持つ事物のクラスも表し得るなど），名詞句や複合語の意味計算などの問題を検討する必要がある。語義の多相性や意味計算については現状では一部の事例について分析しているが（本稿 5.2 節，および今田（2012, 2013a）参照），悉皆的に付与，調査するためにはクラスとインスタンスの区別，アリティ（項数），関数の定義域と値域のクラスなどの情報が必要である。これらの情報は部分的には SUMO から収集可能であるが十分ではないため，収集可能な情報の利用方法や情報の収集・自動推定方法について今後検討したい。意味関係については SORT と NSORT という 2 つの類型を分けたが，その下位分類のより詳しい検討と形式化，文全体の意味演算の手続き等について検討したい。

他の次元の言語情報の付与については，統語・形態論的情報や情報構造の付与が挙げられる。本稿では構文類型として通常の文（NORMAL）と分裂文（CLEFT）を区別したが，他に「～とは～だ」文，「～かが～だ」文，「～は～が～だ」文など特徴的な構文を区別しておいた方が意味情報付与の手続きを単純化できる可能性がある。他に主語のマーカーの区別（「は」「も」「が」および無助詞など）や連体修飾構造などが名詞述語文の統語・形態論的特徴として考えられる。情報構造については 1 節で述べたように本研究で扱う意味情報とは別の独立した層として想定しており，より統合的な意味情報付きコーパスを構築するために別途付与して重ね合わせを行う必要がある。

他のコーパスへの意味付与について，本稿では京大コーパス（新聞テキスト）を付与対象としたが，今後は対象を現代日本語書き言葉均衡コーパス [9]（BCCWJ）に換えて研究を継続したいと考えている。BCCWJ は京大コーパスより大規模なコーパスで新聞を含む様々なレジスタのテキストを含む。そのため本研究で収集した事例より広範な事例の収集や，レジスタ間の差異の観察を期待することができる。また BCCWJ をベースとした述語項構造・照応関係アノテーション（小町・飯田 2011），語義曖昧性解消（奥村・白井 2009），日本語フレームネット意味アノテーション（小原・加藤・斎藤 2011），否定の焦点アノテーション（松吉・大槻・福本 2013）など各種の意味情報付与に関する研究が行われており，本研究の成果とそれらの研究の成果を組み合わせることによって新たな研究課題，手法の開拓を期待することができる。

BCCWJ を用いた名詞述語文研究としては，最近，佐藤（2013, 2014）が発表されている。佐

藤（2013, 2014）は検索アプリケーション「中納言」[10] を用いて名詞述語文の用例を収集し，高橋（1984）の枠組みに基づいて分類し計量的な分析を行ったものである。高橋（1984）は実例に基づく悉皆性の高い分類であり，また意味論と機能論という次元の異なる情報の区別を明確に意識している点などで特筆すべき研究であるが，理論的研究への応用や資料の利活用，分類の再現性などを考慮するとより形式化，体系化された枠組みの開発が必要である。また，佐藤（2013, 2014）は用例収集にあたって「名詞は／が…名詞だ」というパターンを検索対象としているが，この条件で検索される名詞述語文はある程度限定されるため悉皆的ではない。これらの問題を解消する観点からも，共有可能で客観性の高い言語資料の構築とそれを利用する技術の開発が俟たれる。

参照文献

天野みどり（1995）「後項焦点の「A が B だ」文」『人文科学研究』89: 1–24．新潟大学人文学部.

Bond, Francis, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki (2009) Enhancing the Japanese WordNet. *Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009*: 1–8.

橋本泰一・中村俊一（2010）「拡張固有表現タグ付きコーパスの構築：白書，書籍，Yahoo! 知恵袋コアデータ」『言語処理学会第 16 回年次大会発表論文集』916–919.

今田水穂（2010）「日本語名詞述語文の意味論的・機能論的分析」博士論文，筑波大学.

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』61: 83–101．筑波大学大学院人文社会科学研究科文芸・言語専攻.

今田水穂（2013a）「オントロジー体系を用いた名詞述語文の意味記述」『日本言語学会第 146 回大会予稿集』156–161.

今田水穂（2013b）「日本語名詞述語文の意味関係アノテーション」『第 4 回コーパス日本語学ワークショップ予稿集』257–266.

今田水穂（2014）「拡張固有表現階層から SUMO への対応表」『第 6 回コーパス日本語学ワークショップ予稿集』183–192.

Jackendoff, Ray (1983) *Semantics and cognition*. Cambridge, Mass.: MIT Press.

Jackendoff, Ray (2002) *Foundations of language*. Oxford, New York: Oxford University Press.

菊地康人（1997）「「が」の用法の概観」川端善明・仁田義雄（編）『日本語文法：体系と方法』101–123．東京：ひつじ書房.

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』325–330.

黒橋禎夫・長尾眞（1997）「京都大学テキストコーパス・プロジェクト」『言語処理学会第 3 回年次大会発表論文集』115–118.

益岡隆志（2000）『日本語文法の諸相』東京：くろしお出版.

松吉俊・大槻諒・福本文代（2013）「日本語における否定の焦点アノテーション」『第 3 回コーパス日本語学ワークショップ予稿集』425–434.

三上章（1953）『現代語法序説：シンタクスの試み』東京：刀江書院.

Niles, Ian and Adam Pease (2001) Towards a standard upper ontology. *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*: 2–9.

西山佑司（1988）「指示的名詞句と非指示的名詞句」『慶応義塾大学言語文化研究所紀要』20: 115–136.

西山佑司（2003）『日本語名詞句の意味論と語用論』東京：ひつじ書房.

野田尚史（1982）「「カキ料理は広島が本場だ」構文について」『待兼山論叢 日本語学篇』15: 45–66．大阪大学文学部.

小原京子・加藤淳也・斎藤博昭（2011）「日本語フレームネットにおける BCCWJ への意味アノテーション」『日本語コーパス平成 22 年度公開ワークショップ』513–518.

奥村学・白井清昭（2009）「BCCWJ を用いた新しい語義曖昧性解消タスク」『言語処理学会第 15 回年次大会

発表論文集』380–383.
奥津敬一郎（1978）『「ボクハウナギダ」の文法：ダとノ』東京：くろしお出版.
坂原茂（1990）「役割，ガ・ハ，ウナギ文」日本認知科学会（編）『認知科学の発展 3』29–66. 東京：講談社.
笹野遼平・河原大輔・黒橋禎夫・奥村学（2013）「構文・述語項構造解析システム KNP の解析の流れと特徴」『言語処理学会第 19 回年次大会発表論文集』110–113.
佐藤雄一（2013）「名詞述語文「A は B だ」の種類と使用比率」『共立国際研究 : 共立女子大学国際学部紀要』30: 161–177.
佐藤雄一（2014）「名詞述語文「A が B だ」の使用率：意味と構造の面から」『共立国際研究：共立女子大学国際学部紀要』31: 133–147.
Sekine, Satoshi and Hitoshi Isahara (2000) IREX: IR and IE Evaluation project in Japanese. *Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC 2000)*: 1475–1480.
新屋映子（1989）「“文末名詞”について」『国語学』159: 1–14.
新屋映子（1994）「意味構造から見た平叙文分類の試み」『日本語学科年報』15: 1–15. 東京外国語大学.
砂川有里子（1996）「日本語コピュラ文の談話機能と語順の原理:「A が B だ」と「A のが B だ」構文をめぐって」『文藝言語研究 言語篇』30: 53–71. 筑波大学文芸・言語学系.
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』東京：くろしお出版.
高橋太郎（1984）「名詞述語文における主語と述語の意味的な関係」『日本語学』3(12): 1–17.
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』1: 53–75.
山崎誠（2014）「言語単位と文の長さが品詞比率に与える影響」『第 5 回コーパス日本語学ワークショップ予稿集』233–242.

関連 Web サイト

(1) Ruby と MSXML による日本語名詞述語文の実例調査とコーパス分析ツールの構築　https://sites.google.com/site/kaken23720225/
(2) 京都大学テキストコーパス 4.0　http://nlp.ist.i.kyoto-u.ac.jp/index.php? 京都大学テキストコーパス
(3) KNP　http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
(4) 拡張固有表現タグ付きコーパス　http://www.gsk.or.jp/catalog/gsk2013-b/
(5) IREX　http://nlp.cs.nyu.edu/irex/index-j.html
(6) 日本語 WordNet　http://nlpwww.nict.go.jp/wn-ja/
(7) Suggested Upper Merged Ontology (SUMO)　http://www.ontologyportal.org/
(8) CaboCha　https://code.google.com/p/cabocha/
(9) 現代日本語書き言葉均衡コーパス　http://www.ninjal.ac.jp/corpus_center/bccwj/
(10) 中納言　https://chunagon.ninjal.ac.jp/

【付録】作成したコーパスの DTD

```
<?xml encoding="UTF-8"?>

<!ELEMENT document (article)+>
<!ATTLIST document
  xmlns CDATA #FIXED ''
  id CDATA #REQUIRED>

<!ELEMENT article (sentence)+>
<!ATTLIST article
  xmlns CDATA #FIXED ''
  id CDATA #REQUIRED>

<!ELEMENT sentence (chunk)+>
<!ATTLIST sentence
  xmlns CDATA #FIXED ''
  id NMTOKEN #REQUIRED
  info CDATA #REQUIRED>

<!ELEMENT chunk (tag)+>
<!ATTLIST chunk
  xmlns CDATA #FIXED ''
  id CDATA #REQUIRED
  link CDATA #REQUIRED
  rel NMTOKEN #REQUIRED>

<!ELEMENT tag ((memo|rel)*,mode*,tok+
,wordnet*,sumo*,copula?)>
<!ATTLIST tag
  xmlns CDATA #FIXED ''
  conj NMTOKEN #IMPLIED
  func_base CDATA #IMPLIED
  func_range NMTOKEN #IMPLIED
  head_base CDATA #IMPLIED
  head_range NMTOKEN #IMPLIED
  id CDATA #REQUIRED
  link CDATA #REQUIRED
  rel NMTOKEN #REQUIRED
  sumo CDATA #IMPLIED>

<!ELEMENT memo (#PCDATA)>
<!ATTLIST memo
  xmlns CDATA #FIXED ''>

<!ELEMENT rel EMPTY>
<!ATTLIST rel
  xmlns CDATA #FIXED ''
  sid NMTOKEN #IMPLIED
  tag CDATA #IMPLIED
  target CDATA #REQUIRED
  type CDATA #REQUIRED>

<!ELEMENT mode (#PCDATA)>
<!ATTLIST mode
  xmlns CDATA #FIXED ''
  rel CDATA #REQUIRED>

<!ELEMENT tok (#PCDATA)>
<!ATTLIST tok
  xmlns CDATA #FIXED ''
  base CDATA #REQUIRED
  cat NMTOKEN #REQUIRED
  cform CDATA #REQUIRED
  conj NMTOKEN #REQUIRED
  ctype CDATA #REQUIRED
  ene NMTOKEN #IMPLIED
  id CDATA #REQUIRED
  ne CDATA #IMPLIED
  pos NMTOKEN #REQUIRED
  read CDATA #REQUIRED>

<!ELEMENT wordnet EMPTY>
<!ATTLIST wordnet
  xmlns CDATA #FIXED ''
  name CDATA #REQUIRED
  rel CDATA #REQUIRED
  sumo CDATA #REQUIRED
  synset NMTOKEN #REQUIRED
  word CDATA #REQUIRED>

<!ELEMENT sumo EMPTY>
<!ATTLIST sumo
  xmlns CDATA #FIXED ''
  ancestors CDATA #IMPLIED
  class_name CDATA #REQUIRED
  instance_name NMTOKEN #IMPLIED
  src NMTOKEN #REQUIRED
  type NMTOKEN #IMPLIED>

<!ELEMENT copula (arg)*>
<!ATTLIST copula
  xmlns CDATA #FIXED ''
  copula_omitted CDATA #IMPLIED
  head_base CDATA #REQUIRED>

<!ELEMENT arg EMPTY>
<!ATTLIST arg
  xmlns CDATA #FIXED ''
  const NMTOKEN #REQUIRED
  dep CDATA #REQUIRED
  func_base NMTOKEN #IMPLIED
  head_base CDATA #REQUIRED
  rel NMTOKEN #REQUIRED
  tag CDATA #REQUIRED
  type CDATA #REQUIRED>
```

Semantic Annotation for Japanese Copular Sentences

IMADA Mizuho

Elementary and Secondary Education Bureau, MEXT/
Postdoctoral Research Fellow, Center for Corpus Development, NINJAL [–2014.03]

Abstract
I annotated semantic information on copular sentences contained in the Kyoto University Text Corpus for the purpose of consolidating descriptive research and constructing a sharable resource for linguistic research. This task had four subtasks: i) converting the corpus into XML format, ii) annotating word meanings using four semantic resources (Extended Named Entity Annotated Corpora, CRL named entity data, Japanese WordNet, and SUMO), iii) extracting copular sentences in the corpus, and iv) annotating a semantic relation between the subject and predicate of copular sentences. Following the results of the annotation, I investigated the semantic relationships between the subject and predicate and their word meanings, and the syntactic and semantic features of Japanese copular sentences.

Key words: Japanese copular sentences, annotation, word meaning, semantic relation, ontology