科学研究費助成事業　　研究成果報告書

平成　２６　年　　４　月　２２　日現在

機関番号：　６２６１８
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７２０２２５

研究課題名（和文）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

研究課題名（英文）Survey of Japanese copular sentences and development of corpus analyzing tools with Ruby and MSXML

研究代表者
今田　水穂（IMADA, Mizuho）

大学共同利用機関法人人間文化研究機構国立国語研究所・コーパス開発センター・プロジェクトPDフェロー

研究者番号：１０５７９０５６

交付決定額（研究期間全体）：（直接経費）　　1,700,000円、（間接経費）　　510,000円

研究成果の概要（和文）：京都大学テキストコーパスに含まれる日本語名詞述語文に対して、語義情報や意味関係情報などの言語情報の付与を行った。言語情報の付与はコンピュータ・プログラムや概念辞書(日本語WordNetおよびSUMO)を使用した自動処理と、人手による修正作業を併せて実施した。作成したデータとツールはインターネット上で公開した。

研究成果の概要（英文）：I annotated word meanings, semantic relations between two nouns, and other linguistic informations on Japanese copular sentences in Kyoto University Text Corpus. These tasks are partially automated by using computer programs and semantic dictionaries, and partially operated by human annotators. I released the data and tools on the Web.

研究分野：　言語学

科研費の分科・細目：　言語学・日本語学

キーワード：　名詞述語文　コーパス言語学

１．研究開始当初の背景

(1) 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記述的研究がなされていたが，諸説林立の状態にあり理論的観点からの体系化が十分になされていなかった。また，大量の言語資料を悉皆的に調査した研究はごく一部に留まり，名詞述語文の使用実態の全体を十分にカバーしているとは言えない状況にあった。

(2) 近年，言語情報つき大規模コーパス(研究利用可能な電子化テキスト)や概念辞書の開発，構築が進んでおり，形態論情報，構文情報，意味情報などを利用した大規模調査の可能性が開けるとともに，そのような研究方法の開発が期待される状況があった。

２．研究の目的

(1) 大規模コーパスを使用した名詞述語文の悉皆調査と意味情報付与，およびそれを通じた名詞述語文研究の改訂と精緻化

(2) コーパスを利用した言語研究の手法開発と，研究利用可能なコーパスおよびツールの構築・公開

３．研究の方法

京都大学テキストコーパス(毎日新聞テキストに言語情報を付与したもの)に収録されている関係タグ付きコーパスに対して概念辞書を利用した語義情報付与や名詞述語文（および主語名詞，述語名詞）の特定，抽出などの自動処理を行い，さらに人手による意味情報の付与，修正，補完などの作業を実施した。自動処理は主に Ruby というプログラミング言語を使用して実施した。人手による作業は3～5 名程度の言語学を専攻する大学院生に依頼して実施した他，その結果に基づく再作業を研究代表者自身により実施した。

４．研究成果

(1) 京都大学テキストコーパス全体(約 97 万語)および関係タグ付きコーパス(約 13 万語)について，プログラミング言語で処理しやすい XML 形式に変換するプログラムを作成し，変換処理を実施した。

図 1．京都大学テキストコーパス

```
# S-ID:950101003-001 KNP:96/10/27 MOD:...
* 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"...
首相 しゅしょう * 名詞 普通名詞 * *
は は * 助詞 副助詞 * *
```

図 2．XML 形式コーパス

```
<?xml version="1.0" encoding="UTF-8"?>
<document id="950101">
  <article id="950101003">
    <sentence id="950101003-001" info...
      <chunk id="0" link="26" rel="D">
        <tag id="0" link="1" rel="D">
          <tok id="0" read="むらやま" ...
          <tok id="1" read="とみいち" ...
        </tag>
        <tag id="1" link="37" rel="D">
          <rel type="=" target="村山...
          <tok id="2" read="しゅしょ...
          <tok id="3" read="は" base="...
        </tag>
      </chunk>
```

(2) (1)で作成した XML コーパスのうち関係タグ付きコーパスについて，京大コーパスに収録されている形態論情報(「人名」などの品詞情報)を使用した語義付与，および 2 種類の概念辞書(日本語 WordNet、SUMO)を使用した語義付与の自動処理を実施した。語義情報を付与するために lex という名前の XML 要素を追加し，品詞情報による語義情報を ne 属性で，概念辞書による語義情報を dic 属性で表示した。

図 3．語義付与

```
<tag id="0" link="1" rel="D">
  <tok id="0" read="むらやま" base="村...
  <tok id="1" read="とみいち" base="富...
  <lex base="村山富市" ... ne="Human"/>
</tag>
```

この結果，関係タグ付きコーパスに含まれる 66186 タグ単位(言語情報を付与するための，文節と同じかそれより小さな単位)のうち 49539 タグ単位に対して語義情報を付与した。以下は概念辞書を使用した語義付与の結果である。単一の文節に複数の語義が付与される場合があるため，文節数と語義数の合計は一致しない。

表 1．概念辞書による語義付与

クラス名	頻度
Abstract	669
Agent	3449
Attribute	16315
ContentBearingObject	2348
ContentBearingPhysical	4573
ContentBearingProcess	6370
Entity	1675
GeopoliticalArea	1629
Group	1322
Human	2388
Object	8119
Organization	2894
Physical	392

Process	20062
Proposition	2255
Quantity	3582
Region	3975
Relation	1038
SocialRole	3596
TimeMeasure	3083
Agent,Region	33
Attribute,Human	3
ContentBearingPhysical,Object	3
GeopoliticalArea,Human	41
合計	**89814**

(3) (2)で作成した語義情報付きコーパスについて，京大コーパスに収録されている形態論情報，省略情報，格関係情報を使用して名詞述語文を特定する自動処理を実施した。この結果，主語名詞と述語名詞に相当するタグ単位のペア 1431 対を特定した。これらのタグ単位を表示するため，述語名詞を含むタグ単位に copula という名前の XML 要素を追加し，さらにその要素の下に対応する主語名詞に関する情報を記述するための arg という名前の要素を付加した。

図 4．構文情報付与

```
<sentence id="950101019-001" info="K...>
  <chunk id="0" link="1" rel="D">
    <tag id="0" link="1" rel="D">
      <tok ...>一九九五</tok>
      <tok ...>年</tok>
      <tok ...>の</tok>
      <lex base="一九九五年" .../>
    </tag>
  </chunk>
  <chunk id="1" link="2" rel="D">
    <tag id="1" link="2" rel="D">
      <rel .../>
      <tok ...>えと</tok>
      <tok ...>は</tok>
      <lex base="えと" .../>
    </tag>
  </chunk>
  <chunk id="2" link="-1" rel="D">
    <tag id="2" link="-1" rel="D">
      <rel .../>
      <memo>CO</memo>
      <tok ...>亥</tok>
      <tok ...>。</tok>
      <lex base="亥" begin="5" end="5"/>
      <copula ...>
        <arg tag="1" rel="ガ" .../>
      </copula>
    </tag>
  </chunk>
</sentence>
```

(4) (3)で抽出したタグ単位のペア 1431 対の全てに対して人手により意味関係情報を付与した。意味関係の種類は taxonomic(上下関係，内包外延関係，同一関係など同一のタイプの概念間の関係)，non-taxonomic(所有関係など異なるタイプの概念間の関係，たまたま同じタイプの概念同士の場合もある)，cleft(分裂文)の 3 種類とし，さらにそれらを下位分類した。また通常の名詞述語文として扱うことが妥当ではない事例については reject という分類名をつけ参考として残した。これらの情報は arg 要素の sem_rel 属性として付与した。結果を以下に示す。

表 2．意味関係付与

関係名	頻度
taxonomic	697
non-taxonomic	491
cleft	144
reject	99
合計	1431

(5) 作成したコーパスおよび自動処理のためのツール(プログラム)について，近日中にウェブ上で公開する予定である。

表 3．公開するツール一覧

ツール名	説明
kc2xml.rb	京大コーパスを XML 形式に変換
morph.rb	形態論情報による語義付与他
wordnet.rb	日本語 WordNet による語義付与
sumo.rb	SUMO による語義付与
copula.rb	名詞述語文の抽出

なお(1-4)の数値はいずれも最終年度 3 月末時点のものであり，今後若干の修正がある場合がある。

(6) コーパスへの意味情報付与に関する研究としては動詞やサ変名詞の述語項構造情報を解析，記述するものが多くあるが，名詞述語文の述語項構造に相当する情報を記述する研究は稀有である。また，名詞述語文研究の中である程度の規模の言語資料やコーパスを悉皆的に調査したものは存在するが，作成したデータや作成(再現)のためのツールを共有可能な言語資源として公開したものは他に類が無い。今後は本研究で蓄積したノウハウを活用して現代日本語書き言葉均衡コーパス(BCCWJ)で同様の意味情報付与を試みると伴に，作成したデータを利用した統計的研究や，情報構造など他の種類の言語情報を付与したマルチレベルコーパス，名詞述語文以外の構文も対象としたより包括的な言語情報付きコーパスの研究などを推し進めたい。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）

① 今田水穂. 2012.「名詞述語文の精製後異

論的解釈」『文藝言語研究　言語篇』61. pp.83-101. 筑波大学. 査読有

② 今田水穂. 2011.「日本語名詞述語文の類型と主語の意味分類について: 京都大学テキストコーパスと分類語彙表を用いた調査・分析」『文藝言語研究　言語篇』60. pp.25-48. 筑波大学. 査読有

〔学会発表〕（計２件）

① 今田水穂．2013．「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．

② 今田水穂．2013．「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．

〔図書〕（計０件）

〔産業財産権〕

○出願状況（計０件）

○取得状況（計０件）

〔その他〕

ホームページ等

https://sites.google.com/site/kaken23720225/

６．研究組織

(1)研究代表者

今田　水穂（IMADA, Mizuho）

国立国語研究所コーパス開発センター・プロジェクト PD フェロー

研究者番号：１０５７９０５６

(2)研究分担者

なし

(3)連携研究者

なし