コーパスからの単文データの抽出

橋本三奈子 †　　斉藤初江 †　　井佐原均‡

†情報処理振興事業協会(IPA) 技術センター　　‡通信総合研究所 関西先端研究センター

1　はじめに

「ＩＰＡコーパス」および「RWCテキストデータベース」を用いて、実例中に現われる体言と、動詞や形容詞との係受け関係を抽出し、その情報を元に復元した単文を含むデータベースを作成した。以下では、データの概要、作成手順、抽出した単文、データベースファイルの内容、活用方法について解説する。

2　データの概要

コーパスとして用いたのは、「ＩＰＡコーパス」と「RWCテキストデータベース」である。それぞれのデータの概要は、次の通りである。

（1）ＩＰＡコーパス
- （a）コーパスの内容
  - 形態素情報付きのテキストデータベース
  - ・ＩＰＡＬ辞書に収められている意味記述文および文例
  - ・外国人用の日本語教科書の例文
  - ・岩波ジュニア新書７冊、岩波新書１３冊分
- （b）作成したデータベースファイルの規模
  - ・原文数　５７，４５６
  - ・延べ単文／用言数　１０８，２６２
  - ・異なり単文数　９３，５５４
  - ・異なり用言数　５，１６１
  - ・多出現用言(上位１３語 表記別)
    なる,ある,する,できる,ない,いう, いる, 見る,もつ,考える,言う,出る,つくる

（2）RWCテキストデータベース
- （a）データベースの内容
  - 形態素情報付きのテキストデータベース
  - ・毎日新聞（９４年度版）３０００記事
- （b）作成したデータベースファイルの規模
  - ・原文数　４３，４３０
  - ・延べ単文／用言数　５９，９３９
  - ・異なり単文数　５６，０３０
  - ・異なり用言数　３，２４３
  - ・多出現用言(上位１３語 表記別)
    なる,ある,する,ない,行う,よる,出る, 受ける,求める,見る,示す,入る,言う

3　作成手順

データは次のような手順で作成した。

（1）ＧＵＩツールを用いた人手による係り受け関係抽出
- （a）コーパスファイルを読み込む
- （b）文ＩＤ、単語番号付きファイルを生成する
- （c）一文を一行一単語で、品詞別に色分けして表示する
- （d）作業者が体言部分の末尾の形態素をクリックして選択する
- （e）作業者が助詞をクリックして選択する（連体修飾関係の場合は不要）
- （f）作業者が用言部分の先頭の形態素をクリックして選択する
- （g）体言と用言をラインで結び、横に助詞を表示する
- （h）必要があれば、作業者が助詞を正しい格助詞に変更する
- （i）（d）～（h）まで作業を繰り返す
- （j）体言・助詞・用言の単語番号、変更した格助詞を中間ファイルに書き込む
- （k）（c）～（j）まで繰り返す
- （l）中間ファイルを係受け関係ファイルに書き込む

（2）計算機によるデータベース作成
- （a）コーパスファイル、係受け関係ファイルを読み込む
- （b）単語番号と品詞タグを元に用言を復元する
- （c）単語番号と品詞タグを元に体言を復元する
- （d）体言に係る修飾部を品詞タグを元に復元する
- （e）一用言に係る体言をつなぎ単文を生成する
- （f）格助詞ごとに各項目を配置する
- （g）各項目をデータベースファイルに書き込む

4　抽出した文例

例えば、以下の二つの文は実例であるが、「（援用）する」「考える」「出す」「行なう」「（診療）する」「いる」「当たる」「（高齢化）する」という動詞が出現している。

例１：そこで種々の経験的な知識を援用することが考えられ、できるだけ早く解答を出すための工夫が行なわれた。

例２：高齢で元気に診療している医師もいるが、休日当番医に当たる開業医は次第に高齢化している。

例１の「考える」では、「られる」が付加され、この用言に係る体言「こと」の後には、格助詞「が」が現われているが、体言「こと」と用言「考える」を組み合わせて形成される単文においては、両者は「ヲ」で結びつけられるはずである。このような場

合、先に述べた（１ｈ）の作業で、「が」は「を」と交替する、という情報を入力する。

また、例２の「いる」では、この用言に係る体言「医師」の後には、係助詞「も」が現われているが、これは格助詞に置き換えれば、「が」になるものである。このような場合も、「も」が「が」と交替する、という情報を入力する。

また、「当たる」は連体用法で用いられている。このような場合は、被修飾語の「開業医」を抽出した後、「開業医」と「（休日当番医に）当たる」とを組み合わせて形成される単文において現われるはずの格助詞「が」を入力する。

さらに、例１の「出す」では、「早く」という形容詞連用形が連用修飾語として現われている。このような連用修飾語も、参考情報のために拾った。格助詞には便宜的に「φ」を入力した。

このような情報を元にして作成した文を下に示す。下の例の文ａは、作業者によって入力された格助詞を用いて作成したものである。文ｂは、実例中の助詞を用いて、「られる」「ている」や助動詞を付加したもので、原文に即した文である。「当たる」のように、連体修飾用法で用いられている場合には、文ａでは、文が自然になるように、被修飾語に「その」を付加し、文ｂでは被修飾語を用言の後ろに置いた。なお、「.」は単語の切れ目を、「_」は修飾要素と体言との切れ目を示す。「＜Ｓ＞」は、その部分に文相当の表現が入っていたことを示すものである。

（１）文ａ：種々.の.経験.的.な_知識を 援用する
　　文ｂ：種々.の.経験.的.な_知識を 援用する
（２）文ａ：<S>_ことを 考える
　　文ｂ：<S>_ことが 考え_られる
（３）文ａ：早くφ 解答を 出す
　　文ｂ：早くφ 解答を 出す
（４）文ａ：<S>.ため.の_工夫を 行なう
　　文ｂ：<S>.ため.の_工夫が 行なわ_れる
（５）文ａ：その医師が 高齢で 診療する
　　文ｂ：高齢で 診療し_て.いる 医師
（６）文ａ：<S>_医師が いる
　　文ｂ：<S>_医師も いる
（７）文ａ：その開業医が 休日.当番医に 当たる
　　文ｂ：休日.当番医に 当たる 開業医
（８）文ａ：<S>_開業医が 高齢化する
　　文ｂ：<S>_開業医は 高齢化し_て.いる

5　データベースファイルの内容

データベースファイルは、表にあげる項目の順序で１レコードとした、カンマ引用符形式のファイルである。以下で「〈 〉」で括るのは項目名を指す。

5.1　用言

用言として抽出したのは、動詞と形容詞である。いわゆる漢語サ変動詞については、〈用言〉には「する」のみを入れ、語幹部分は別項目の〈体言０〉に入れた。さらに、用言が連用形の場合には、その用言の直後に存在していた動詞を〈後用言〉という項目に入れた。これは複合動詞の可能性を示すものである。また、用言の直後に存在していた「（ら）れる」「（さ）せる」「ている」「てある」などは〈接尾〉に入れた。また、用言直後の「た」「ない」などの助動詞は〈助動詞〉に入れた。これらの要素は、用言が取り得る体言や格助詞を変更させる要因になる場合があるため、データとして残した。

5.2　助詞

単文形成時にあてはまる格助詞として、「が，を，に，から，と，より，で，へ，φ，に／へ，が／に，が／で，が／を，まで」を入力し、これらを〈格助詞ガ〉から〈格助詞マデ〉の欄に格納した。原文の助詞がここにあげた格助詞と交替できない場合のために〈格助詞他１〉から〈格助詞他１０〉も用意した（以下では「ガ」～「他１０」をｘで代用する）。

助詞として原文から抽出したのは、格助詞の「が，を，に，から，と，より，で，へ」、格助詞相当の連語の「として，と共に，に当たって，に当たり，において，にかけて，に関し，に際し（て），に従い，に従って，に対し（て），について，につけ（て），につれ（て），にとって，によって，により，にわたって，をもって，を通じて，を通して」、および係助詞の「は，も，こそ，さえ，しか，すら」である。これらは〈格助詞出現形ｘ〉に格納した。原文中で格助詞が脱落している場合には「φ」を、連体被修飾語の場合には「＊」を入れた。

ところで、「に」には必須格と随意格（副詞句ともいえる）とがある。後者の「に」は格助詞ではなく、助動詞「だ」の連用形、あるいは副詞化の助詞などといわれることもあるが、区別する基準を明確に設けるのは難しい。そのためこのような基準作りのデータとしても活用できるように、「～に」はなるべく拾った。ただし、格助詞や係助詞以外の品詞タグが付加されているものはツールの制限で拾うことができなかった。したがって、文ｂでは、この部分が「φ」に復元されてしまっているので、注意が必要である。このような「に」については、「格助詞ニ２」「体言ニ２」という項目に格納してある。

（9）原文：ある 時期 を 愉快 に 過ごし て 、…
文a：ある.時期を 愉快に 過ごす
文b：ある.時期を 愉快φ 過ごし_て

5.3 体言

体言として抽出したのは、「名詞」「固有名詞」「代名詞」「数」「いわゆる形式名詞」「名詞性の接尾」を表わす品詞タグがついた形態素であり、体言列の一番末尾を抽出した。したがって、抽出した要素の直前に、今あげた品詞タグを持つ形態素が続く場合には、複合語であると仮定し、それらを含めて各格助詞に対応する〈体言x〉に入れた。また、引用の格助詞「と」の場合を考えて、記号の「閉じ括弧」も抽出した。その場合には、「開き括弧」が現われるまでを一つの体言として格納した。したがって、格助詞「と」が括弧つきでない文相当の表現を受ける場合には、末尾が活用語の見出し形であるため、ト格そのものを拾えなかった。また、一つの体言は、複数の用言に係ることを許した。

5.4 先行

例えば、体言「種」には、「植物で、発芽のもととなるもの」という語義と、「ある感情を引き起こすもととなるもの」という二つの語義がある。「太郎が種を蒔く」という組み合わせだけでは、このような多義性を解消することはできない。実例中には、多義性を解消する要素が、「ひまわりの種を蒔く」や「争いの種を蒔く」のように、体言の修飾要素として現われていることが多い。そのため、体言の修飾要素も格納することにした。ただし、ここでは、直前の「の」「な」や活用語の連体形は直後の要素に係る、というような簡単な原則にのっとって、品詞タグだけを手がかりに復元した。

このとき、並立助詞（「や，と，とか」など）の扱いに問題があった。抽出した体言や修飾要素中の体言の直前に並立助詞が現われたら、それも〈先行x〉として復元したが、（10）のような場合には正しい復元となるが、（11）では「や」は直後の単語に係るわけではないので、単純に復元してはいけないものである。

（10）原文：膨大 な 量 の データ や 情報 を 記憶 する こと が…
文a：膨大.な.量.の.データ.や._情報を 記憶する
（11）原文：言語 理論 や 言語 学者 の 書く 文法 は 文 の 基本 的 な 構造 を 明らか に する こと に 目的 が あり…
文a：言語.理論.や._言語.学者が その文法を 書く
文b：言語.理論.や._言語.学者の 書く 文法

しかし並立助詞を復元するのをやめると（12）のように本来必要な情報まで落ちてしまうことになる。

（12）原文：関係 する データ や 情報 を 相互 に 結合 し
文a：情報を 相互に 結合する

分析した結果、「並立助詞」の復元に失敗するのは、（11）のように元の文が連体修飾用法になっている場合に顕著であることがわかった。そこで、連体修飾の場合に限って、並立助詞の復元をやめることにした。したがって、（11）は次のようになる。

（11'）文a：言語.学者が その文法を 書く
文b：言語.学者の 書く 文法

ただし、連体修飾でない場合でも、並立助詞が直後の体言に係らない場合もある。そのような場合にはデータから削除することを簡便に行なえるように、並立助詞の後の区切りは、「..」「._」で示し、他の単語の列とは異なった表示をしてある。

6 おわりに

最後にこのデータベースの特徴と活用方法を示す。このデータベースは、原文に出現する助詞と単文形成時の格助詞との変化に着目して作成してあるため、以下のようなことを調査し、ルール化するためのデータとして活用することができる。

（a） 用言に「られる」「させる」「てある」「がたい」「づらい」等が接続した場合に、どのような格助詞交替が起きるのか
（b） 用言に「ない」「て」「ている」等が接続した場合に、格要素や修飾要素の語彙がどのように変化するのか
（c） 実例中の「は」「も」などの係助詞はどの格助詞として解釈できるのか
（d） 実例中で脱落する格助詞には、どのような傾向があるのか
（e） 実例中で被修飾語になるのは、どのような格の体言なのか

さらに、体言と用言との係受け関係を記載してあるので、次のような言語処理アプリケーションのデータや学習用データとして利用することができる。

（f） 体言と用言の共起関係を用いたクラスタリング
（g） 統計的手法を用いた係受け解析
（h） 格関係辞書の自動生成
（i） コローケーション辞書作成支援

表 データベースファイルの仕様

Ｎｏ	ファイルごとのレコードのシーケンシャルＮｏ。
実例ＩＤ	コーパスの文ＩＤ。
格パタン	格助詞の組み合わせ。「+」でつなぐ。 検索の便宜を考えて、一定の順序（〈格助詞ｘ〉の#にあげた格助詞の順）につないである。
用言	動詞あるいは形容詞の見出し形。複合サ変動詞の場合は「する」のみ。
用言出現形	動詞あるいは形容詞の出現形。 複合サ変動詞の場合は「する」の出現形。
後用言	複合動詞の場合の後項要素の見出し形。
接尾	用言に続く「（ら）れる」「（さ）せる」「ている」などの見出し形。 形態素解析上の単語の切れ目は「.」で示す。
助動詞	用言に続く助動詞の見出し形。
格助詞ｘ	格助詞「ｘ」。単文形成時にあてはまる格助詞として入力したもの。 # 「ｘ」として、ガ、カラ、ヘ、ト、ヨリ、デ、ϕ、ヲ、ニ、ガ２、ニ／ヘ、ガ／ヲ、ガ／ニ、ガ／デ、マデ、ニ２、他１～他１０の欄がある。
格助詞出現形ｘ	格助詞の出現形。係助詞を含む。格助詞脱落の場合は「ϕ」、連体被修飾語の場合は「*」。
素性ｘ	意味素性のための欄。現在は「＊＊＊」。
先行ｘ	格助詞「ｘ」の位置にたつ体言を連体修飾する部分。 形態素解析上の単語の切れ目は「.」で示す。 並立助詞の後は、「..」で示す。
体言ｘ	格助詞「ｘ」の位置にたつ体言。 形態素解析上の単語の切れ目は「.」で示す。
体言Ｈｘ	格助詞「ｘ」の位置にたつ体言の、形態素解析上末尾にくる語。
格助０	空欄。
素性０	意味素性のための欄。現在は「＊＊＊」。
先行０	サ変接続の名詞接尾（例「化」）に先行する部分。
体言０	いわゆるサ変動詞語幹。
副詞	現在は空欄。
文ａ	先行ｘ、名詞ｘ、格助詞ｘ（入力した格助詞）、用言を用いて作成した文。これがいわゆる「単文」である。 先行ｘと名詞ｘとの切れ目は、「_」あるいは「._」で示す。実例での出現順に助詞をつなげてある（以下の文ｂ、文ｃでも同様）。
文ｂ	先行ｘ、名詞ｘ、格助詞出現形ｘ、用言出現形、後用言、接尾、助動詞を用いて作成した文。 連体修飾用法の場合はそのような形で復元した。原文に一番近い形の文。
文ｃ	先行ｘ、名詞ｘ、格助詞ｘ（入力した格助詞）、用言出現形、後用言、接尾、助動詞を用いて作成した文。 「ている」「ない」などが続かないと文として成立しない場合があるので、そのために作成した。 「られる」「される」などが接続した場合には、入力した格助詞を使っているため、かえって変な文になってしまうこともある。
備考	コメントのための欄。現在は空欄。
その他	コメントのための欄。現在は空欄。

図 データベースファイルの例


```
"00003610","00018340:0007","が+に","当たる","当たる","","","","が","＊","＊＊＊","","開業医","開業医",
"","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","",
"","","","","","に","に","＊＊＊","","休日.当番医","当番医","","","","","","","","","","","","","","","","","",
"","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","",
"","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","","",
"","","","","","","","","","","","","","","","","その開業医が 休日.当番医に 当たる","休日.当番医に 当たる
開業医","休日.当番医に 当たる 開業医","",""
```


[謝辞]このデータベースは、青山文啓氏、荻野紫穂氏、桑畑和佳子氏、徳永健伸氏、元吉文男氏、ＩＰＡ日本語グループの委員諸氏との議論を通じて作成したものです。データの利用を許可して下さった岩波書店、筑波大学、長尾真先生、毎日新聞社に感謝申し上げます。また、作業やプログラミングを担当して下さった黒木龍房氏、佐藤智子氏、鈴木悟子氏、田中啓介氏、玉井陽子氏真柄麻樹氏に感謝の意を表します。