動詞結合価データ作成および格認定の諸問題

荻野孝野[12]、大久保佳子[1]、小林正博[1]、井佐原均[23]
{ogino,ookubo,kobayasi} @jsa.co.jp; isahara@crl.cp.jp
1 日本システムアプリケーション
2 神戸大学大学院自然科学研究科
3 通信総合研究所

1. はじめに

大量の係り受け関係のついた言語データである EDR コーパスおよび EDR 共起辞書を用いて、「日本語動詞の結合価データ」を作成した。本データ作成の背景や作成手順、本データの形式などについて述べるとともに、作成途上のデータ整理段階で出現した格認定の諸問題について報告する。

２．作成の背景

①車が走る　②電車が走る　③人が走る　④亀裂が走る　⑤虫唾が走る

①②③は、人間や車が移動の一つの形態として早い速度をもって移動すること、④は形として線状の変形が起こること、⑤は「虫唾」という言葉で表現される不快な感情が気持ちの内部に流れることである。

このように、同じ「走る」でも語義の違いによって、動詞にかかることができる名詞側の単語に意味的な特徴がある。これは、動詞側からの名詞側に要求される語彙の意味的制約で、「選択制限」とか「共起関係」と呼ばれるものである。日本語を母国語として使っているものにはおそらく意識することなく、格助詞の使い分けがなされているものと思われるが、あらかじめそれらの知識を習得していない日本語学習者や、言語知識を蓄えておかなければ稼動しない言語処理関係のシステムなどでは、これらの関係を整理し、知識として蓄えておくことが必要である。

上記のような背景からＥＤＲ電子化辞書の大量コーパスを基に、日本語学習者や自然言語処理システムで使える結合価データを、「日本語動詞の結合価」[1]としてまとめた。

３．データ量

作成された結合価データは、およそ以下の規模からできている。

動詞概念にして	異なり	約　１２,４００概念
事例にして	延べ	約１６０,０００例
文数にして	延べ	約１５６,０００文

ここで上記に示した「動詞概念」とは、品詞が動詞（サ変動詞も含む）のもので「見出しの表記+概念ＩＤ」の組み合わせで限定される一概念を単位としたものである。

表１によって、その関係を説明すると次のようになる。①②は、概念ＩＤが同じだが異なる表記で用いられている。③④は同じ表記だが異なる概念がついている。このいずれのケースも「見出しの表記＋概念ＩＤ」レベルで異なるレコードとなり、ここで示す概念数とは表１に示す１行分を１概念と数えたものである。

表１　見出しの表記と概念ＩＤの関係

	見出しの表記	概念ＩＤ	概念説明
①	歩く	3cefe6	足を使って歩く
②	歩む	3cefe6	足を使って歩く
③	合う	1fa2c9	ぴったり一致する
④	合う	3ce9d3	釣り合いがとれる

４．元になったデータ

「日本語動詞の結合価」[1]の元になったデータは、①品詞などの形態的な情報、②構文的上の係り受け関係を示す構文的な情報、③着目の単語が該当する見出しの持つ複数概念のどれに対応しているのかを示す概念情報　などがつけられたＥＤＲ日本語コーパス、およびその日本語コーパスの構文情報で係り受けの２項関係を抽出したＥＤＲ共起辞書[2]である。

５．結合価データの形式

表１で示した動詞概念単位で１つの表にまとめている。その表は、該当動詞と共起関係にあった体言を表層格で整理し、実例文の１文単位を１行としたレコードで構成されている。

各表の上部には以下の情報が表示されている。

・動詞の単語表記　　例：食べる[タベ・ル]
・当該動詞の概念識別子　　例：3bc6f0
・当該動詞の概念説明　　例：食物をとる

データの部分は、１３列からなる表形式になっており、左から１１列は係り側情報で、係り受けの関係にあった係り側の用語が記述される。残りの右２列が例文情報エリアである。

格助詞部分は、左から、「は」、「が」、「を」、「に」、「へ」、「から」、「より」、「まで」、「で」、「と」、「その他」の順に表層格の列に定めている。「は」は他の格助詞に代替できない係助詞を、「その他」は、格助詞と同じ形の助詞や語尾で格助

詞に該当しないもの、あるいは格助詞であっても述語に束縛される格関係ではないものを配置する部分である。
係り受け関係にあった用語(句)の情報として、該当する表層格の位置に以下の情報を記述している。

・実際の出現形による格助詞あるいは係助詞
・係り受け関係にあった用語(句)の表層文字列
・係り受け関係にあった用語(句)の品詞

表２　結合価データの表示形式

食べる　タベル　　3bc6f0　　食物をとる											例文数:100文
は	が	を	に	へ	から	より	まで	で	と	その他	例文	コーパスRID
	が(牛(名詞))	→(エサ(名詞))									牛がうまそうに[[食べ]]ているエサは、シラカバの木のこま切れ。	JCO0112245
		を(サンドイッチ(名詞))	は(昼食(名詞))								あまりおなかがすいていなかったので、昼食は、軽くサンドイッチを[[食べ]]た。	JCO0013286
	が(バージニアさん(複合語))	を(これ(名詞))								か(いくつ(名詞))	毒入りチョコは、バレンタインデー前夜に、無記名で届けられ、これを同判事の妻、バージニアさんが、いくつか[[食べ]]たとたんに苦しみだし倒れた。	JCO0190386
	が(妹(名詞))	を(カップめん(名詞))									前日に買ったカップめんを、妹が[[食べ]]たと思った。	JCO0166432

６．結合価データの作成

作成手順の概要は以下の通りである。

(1)共起辞書から機械的処理によって結合価データの初期データを作成する。
(1.1) ＥＤＲ共起辞書から用言を係り先とした共起データを抽出する。
(1.2)同一例文の同一単語に関わる共起データごとに分類して同じ場所の述語にかかる共起データを統合する。
(1.3) ＥＤＲ結合価初期データを出力する。
(2)　(1)で作成した結合価の初期データを手作業によって整備する。
(2.1)結合価初期データを手作業によってチェックし、エラーや態変換の種類の区分分けを行うとともに、訂正内容そのものを記入する。
(2.2) (2.1)のエラーおよび態変換の区分と指示などを元にプログラムで一括訂正できる部分の修正を行う。
(2.3) (2.2)で行った自動訂正結果をチェックするとともに手作業で対応する部分の修正を行う。

６．１ 共起辞書から機械的処理による結合価の初期データの作成--共起データの統合

４に示した共起辞書を出発点として「表記　読み　品詞　概念ＩＤ」が一致したものを１グループとして集める。さらにそのグループの中で例文の位置関係情報（テキスト番号、文番号、単語の出現位置）が一致したものを、１レコード分の結合価データとする。
例１に「書（カ・ク)」のうち概念「0e910d」に相当する共起情報で、位置関係情報が一致するものをまとめたデータを示す。

例１

```
JCC0187586    構文情報:  紙(名詞)[3c1038]      =>[ に ]=> 書(動詞)[0e910d]
              意味情報:  書(0e910d) =>[ place ]=>          紙(3c1038)
              例文情報:  00100000c5bb-17-12/"<紙>に…(書)く
JCC1050271    構文情報:  文字(名詞)[1f5057]    =>[ を ]=> 書(動詞)[0e910d]
              意味情報:  書(0e910d) =>[ object ]=>         文字(1f5057)
              例文情報:  00100000c5bb-17-14/"<文字>を(書)く
```

例文情報のテキスト番号部分が 00100000c5bb で一致していて、このテキスト番号のコーパスに出現する「紙」、「文字」と「書く」による構成であることが分かる。これは、「紙」が共起関係子「に（助詞)」を介して「書（カ・ク)」に係り、「文字」が共起関係子「を（助詞)」を介して「書（カ・ク)」に係っていることを示している。また、共通のオフセット値から１７文字目に「書く」が存在することもわかる。

この共起辞書における同一の位置と係り受けの関係にある共起関係を統合することによって、２項関係にある共起関係から同じ動詞にかかる格関係をとりまとめ、初期データとする。

６．２　結合価データの整備

６．１でまとめた初期データでは、元データの表現形態は、受身、使役、状態を示すものなど様々である。また、共起辞書の係り受け関係から格助詞を手がかりに自動的に所定の列に移しただけでは、コーパスレベルで係り受け関係認定そのものが間違っている場合など、誤った係り受け関係のまま、格関係を所定の列に配置することになってしまう。そこで本データをまとめるにあたって、元々付与されていた構文的な係り受け関係を用いて自動的な配列を行なったあと、

①受身、使役など表層上の格を実質格の位置に移動
②係り受け関係などの誤りにより、不適切な述語にかかっている格助詞部分を適切な関係に修正

という手作業のよる調整を行った。

修正対象となった部分は大きくわけて以下のような部分である。

(1)格の移動に関する訂正
- (1.1)態の変換
- (1.2)連体句の中の格の変換
- (1.3)係助詞の変換
- (1.4)格助詞かどうかの判定がゆれる部分の調整

(2)初期データの不備の訂正
- (2.1)複合語や体言句の抽出範囲の補正
- (2.2)コーパスにおける動詞概念選択そのものの誤り部分の移動
- (2.3)コーパスにおける動詞への係り受け関係付けの不備の訂正

ここで、本データの使い方にもかかわる部分として、上記のいくつかについてエラーの内容や訂正内容について説明する。

(１)格の移動に関する訂正

(1.1)態の変換

該当する動詞が例文中で受身や使役などの場合は、態を変換して基本形にし、本来入るべき列に配置する。格助詞そのものの表現は変えない。これは、基本形の格パターンの検討にも、態が変換した場合の格の移動パターンの検討にも使えるようにしたものである。

(1.2)連体句の中の動詞と体言の格関係

連体句のため格を示す部分が「の」に変わっている場合や、連体句の修飾先に内側の述語の格関係がある場合は、着目している述語を本来とるべき格助詞の列に置き換えて配列する。

(1.3)係助詞の変換

「は、も、すら、さえ」など係助詞、副助詞は実体を示す格の列に配置する。ただし、移動した結果、同じ格が二つになって他の格に置きかえられない係助詞、副助詞は、そのまま「は」の列に表示する。

参考までに、表３で、上記に述べた理由による｢格の移動｣に該当する番号をつけたものを例示する。「＊」のついた部分は原文通りの配置である。

表３　表層の格表示と実質的な格の配置の関係

	書く　カク		0e910d	(文字や符号などを)書き記す									例文数:100文
格の移動の種類	は	が	を	に	へ	から	より	まで	で	と	その他	例文	コーパスRID
(1.1)			が(あて名(名詞))						で(文字(名詞))			しかし、あて名がどんな文字で[[書]]かれているか、までは読み取れない。	JCO0035065
＊		が(男の子(名詞))	を(作文(名詞))									小学5年生の男の子がこんな作文を[[書]]いている。	JCO0151405
(1.2)		が(方(名詞))	→(戦記(名詞))									中学生のころから戦争に興味を抱き始め、いろいろな方が[[書]]かれた戦記を読んできた。	JCO0176715
(1.3)		が(側(名詞))	は(草稿(名詞))								によれば(高官(名詞))	この西独高官によれば、共同声明の草稿は、西独側が[[書]]いた。	JCO0026635

（2）初期データの不備の訂正

いわゆる抽出エラーや元データの問題で、作業手順で述べたいわゆる共起データの統合だけでは不適切な結合価データができてしまう事例である。表４にその一部を示す。

表４　元データで体言の範囲指定や係り受け関係に問題がある場合の対応

エラーの内容	訂正時の対応	例
単語の切り誤り	複合語も検討対象とする。 ただし、概念を持つ語構成部分がとれていれば、複合語の部分でも良い（結合価を書くにあたって、語構成部分から用言に密接にかかわる意味マーカ部分が抽出できれば良い）。 明らかな区切り間違いのみ誤りとする。	誤りとして単語の範囲を変える例： ・どこ－かで ・あたたか－み ・寒－さ 誤りとしない例： ；人を指すことが分かる 「一般－市民」 「編集長－室」
慣用表現や複合表現がまとまって新たな意味を持つ場合の体言部分認定が不適切	修飾語句がないと意味が分からないものや不自然なものでも、体言部分から意味分類が特定できるものであれば、エラー指摘対象としない。 ただし、複数の単語や構文要素が合成されて新しい意味が生じている場合は、体言として取り出す範囲を変える。	誤りとする例： （湯の）花
格要素になるものと副詞･形容動詞の語尾との混同	副詞・形容動詞の語尾相当のものでも、格助詞に認定されてしまっていれば格助詞の列に配置されてしまう。また、「に」など前が体言で格助詞であること自体、誤りでなくても情況化にあたる格助詞など、その動詞にとって、格要素と考える必要がないと判断できるものは、その他の列に入れる。時を示す体言につく「ニ」(EX.３時に)も不要とする。もし格助詞「に」の列に配置されていれば、[その他]に移動する。	[その他]に移動する例： 最終的に、一般的に、 前に、一緒に－会う、 久しぶりに－会う
係り受け関係が誤って認定され、不適切な位置に配置されたもの	同じ格助詞が出現し、不適切な係り受け関係で余計なものが抽出されているために、本来の正しい格が入る場所がない。	「両親と私が靖国神社の社頭で兄と会う」 ；ト格の列に「両親と」の部分が認定されていたため、本来「会う」がとるべき「兄と」の部分が外れてしまっている。 ト(両親)× → ガ（両親と私）○ ト（兄）○

７．結合価データ作成過程での諸問題

共起データを格助詞によって自動的に所定の位置に配置した後、手作業チェックを経て訂正の種類によって訂正区分を記入した後、訂正種類によって訂正指示を自動変換、その後再度チェックを施して、本データを作成したものである。さらに最終的なチェック段階でいくつかの問題が出てきたので、今後の問題として以下に記載しておく。

（1）　受身形で使われるのが一般的な動詞

ここでは基本形で格関係を収録するのを基本としたため、受身形を基本形に直して収録したが、実際のデータでは受身形で出現するのが一般的で、基本形に訂正すること自体が無理というタイプのものがいくつか見られた。ここでは当初方針通り、基本形に変換して記載したが、本来なら受身形の形のまま見出しとして作成するのが妥当かもしれない。

例２　被害に悩まされる

例２の事例が「被害が○○を悩ます」という表現で使われることは一般的でない。こういった事例まで基本形に直すこと自体、利用という立場からは賢明ではないと思われる。

（2）　同じ格が一つの文に２個以上出てしまうケース

格助詞は、並列のケースを除いて一つという前提に立って、一つのセルに一つの格を想定したが、実際は並列以外のケースでも同じ格助詞が２個以上出ている事例もある。

例３　両親は、結婚二十周年記念【に】、ヨーロッパ旅行【に】出かけました。

通常は同じ格助詞を２度使うことを避けて、前の【に】の部分を【として】などに換えるところと思うが、ここでは２度使っている。ただし、意味的には上記の「に」は異なる機能で使われているので、表層上同じ形態の「に」が二重に出てくること自体はありうることである。

このように、並列以外でも表層的に同じ格助詞が出てくる場合があるが、ここでは並列形式と同じ記号で表現している。例３は異なる役割の格を同じ格助詞で表現しているわけで、本来の並列表示とは表示の意味を異にするものである。

文献

[1]荻野孝野･小林正博･井佐原　均,「日本語動詞の結合価」,三省堂(2003.12)
[2] http://www2.crl.go.jp/kk/e416/EDR/J_index.html