

## 結合価情報に基づく日本語語順の推定

徳永 健伸 (東京工業大学)

田中 穂積 (東京工業大学)

ディスクリプタ： 語順 結合価情報 意味素性

I P A L 動詞辞書

### 1. はじめに

日本語は語順が比較的自由的な言語であるといわれているが、語順にまったく制限がないわけではない[1]。特に述部の語順に関しては言語学者による多くの研究があり、かなり厳格な語順規則があることが明らかにされている[9]。一方、述部にかかる要素の語順については述部の語順ほど厳格な規則は見出されていない。

児玉は依存文法の立場から言語に普遍的な語順を規定する原則をまとめており、その一例として日本語をとりあげている[5]。児玉は語順を決定する要因として、

- 依存関係(主要語と修飾語)
- 品詞(名詞, 動詞, 前/後置詞)
- 文法機能(主語, 目的語, 補語, 述語)
- 形態(屈折, 接辞などの語形態と語連鎖の重み)
- 意味(語用論, 認知構造)

の5つを挙げ、中でも依存関係を重要視している。児玉の研究は言語普遍性を重視しているために、日本語固有の問題については深く立ち回っていない。

久野は文を越えた談話という観点から、情報の新旧、話題、焦点などが語順に及ぼす影響について論じている[2]。本論文では単文を考察対象とし、文脈情報は扱わない。

佐伯は小説 67 ページ中の文を手作業で分析し、補語、すなわち、名詞句と格助詞の組の語順について以下の 9 つの傾向を抽出している [4]、

- (1) 位格(ニ, デ, カラ, ヲ)は他の格に先行する
- (2) トキの位格はトコロの位格に先行する
- (3) ガは位格を除く他の格に先行する
- (4) 与格のニは対格のヲに先行する
- (5) カラは着格のニ, へに先行する
- (6) 長い補語は短い補語に先行する
- (7) 文脈指示を含む補語は先行する
- (8) 補語同士がかかり受けを構成する場合, かかりの補語が先行する
- (9) 慣用句では, 特定の補語は動詞の直前に位置する

佐伯はこれらの傾向を成分条件に基づく語順傾向と構文条件に基づく語順傾向の大きく 2 つに分類している。

成分条件に基づく語順傾向とは、補語自身の意味、機能に備わった支配条件に基づいて生じる語順傾向のことで、兎玉のいう文法機能に関する要因にほぼ相当する。(1) から (9) のうち、(1) から (5) がこれにあたる。これらの傾向をまとめると、おおよそ次のようになる。

位格(トキ>トコロ)>主格>与格>対格

ただし、 $X > Y$  は  $X$  が  $Y$  に先行することを意味する。

一方、構文条件に基づく語順傾向とは、補語の構文的な意味、機能に備わった支配条件に基づいて生じる語順傾向のことで、兎玉のいう依存関係、および形態に関する要因に相当する。(6) から (9) がこれにあたる。

佐伯は (1) から (5) の成分条件に基づく語順傾向については、かかりの深さと広さという観点から、(6) から (9) の構文条件に基づく語順傾向については、かかり先のあいまい性という観点から語順傾向の必然性を説明している。次章では語順傾向に影響を与える佐伯のかかりの広さと深さについて説明し、かかりの広さに基づく語順の推定モデルを提案する。

## 2. 語順の推定モデル

### 2.1. かかりの深さと広さ

日本語の単文は単純化すると次の規則で表現できる [1]。ここでいう後置詞句は1章で述べた佐伯の補語に対応する。

文 → 後置詞句\*, 述語, 時制.

後置詞句 → 名詞句, 助詞.

述語 → 動詞 | 形容詞 | 形容動詞語幹, 「だ」 | 名詞句, 「だ」.

ただし, “\*” は直前の要素の0回以上の繰り返しを, “|” は選言を表す。本論文では助詞としてガ, ヲ, ニ, カラ, ヘ, ト, ヨリ, デの8つの格助詞, 述語としては動詞のみを考える。また, 時制, アスペクトは扱わない。すなわち, 我々が扱うのは格助詞によって有標化された名詞句の並びの後に動詞が来るような単文である。

我々の関心は, このように単純化した日本語の単文について, かかり要素(後置詞句)がどのような順序に並ぶかという問題である。佐伯はこの問題をかかりの深さと広さという概念を用いて説明している [4]。

かかりの深さとは, 直観的にはかかり要素と述部の距離を表し, より文末の受けにかかると方がかかりの深さが深いという。かかりの深い要素のほうが先行しやすいというのが佐伯の観察である。これは非交差の原則とも関係がある。また, 1章でも述べたように日本語の述部の語順にはかなり厳格な規則があるので, 述部の要素を語順によって階層化し, かかり要素が述部のどの階層にかかるかによって, かかりの深さの絶対的な指標を求めることができる。かかりの深さという概念は述部の階層構造を前提としているので, 述部に動詞のみを仮定する我々の考察対象では役に立たない。

一方, かかりの広さとは, かかり要素がどの程度述部を限定するかという概念であり, 述部を厳しく限定するほどかかりの広さが狭いと考える。たとえば, 「東京カラ」や「京都へ」のようなかかり要素は, 何らかの移動を表す動詞が文末に来ることを予想させるが, 「私ガ」のようなかかり要素は, 動詞の限定がよりゆるやかである。かかりの広さは述部に動詞しか含まない単文にも適用できるため, 以下ではかかりの広さを考察の対象とする。

本章の最初でも述べたように, 日本語では文の主辞である動詞が文末に位置するという特徴がある。一般に動詞は文全体の意味の中心的な役割を担うので, 言語による情報伝達を考えると, これは一見不合理のように思える。しかし, 人間は実際に何の不自由もなく

日本語を通じて情報を伝達できるし、相手の発話中に動詞が出現する前に返答することさえできる。これは、人間が他人の発話を理解する際に、文の後半を予測しながら理解していることを示唆している。情報伝達の効率化という点からすると、聞き手が発話の後半を自然に予測できるような発話が望ましい。佐伯は動詞が文末に位置するという日本語の特徴をふまえ、かかりの深さが深く、かかりの広さが広いかかり要素から順に提示する方が聞き手にとって望ましいとしている。いいかえれば、より制約の少ないかかり要素を先に提示することになる。佐伯はこの主張の根拠として言い誤りの訂正に要するコストをあげている。以下は佐伯の例である[4]。

(1) しかし、→ 涙を(いや) 涙が → あふれてきた。

(2) 涙を → しかし、(いや) 涙が → しかし → あふれてきた。

(1)は佐伯のいう基本語順であり、(2)はその逆である。各要素を順次発話するとき、述部の発話の直前で述部とかかり要素のかかり受けの矛盾に気がつくとする、(1)の場合、述部を強く限定するかかり要素(涙を)が述部に近い位置にあるので、訂正は1要素ですむが、(2)では、述部と述部を限定するかかり要素(涙を)の間に、述部を必ずしも限定しない他の要素(しかし)があるために2要素の訂正が必要となる。このような語順は、文の表すことからの構成要素を周辺的なものから順次列挙して行って述語の意味を補完し、最後に述語で全体の意味を締めくくるという日本語の求心性[7]とも一致する。我々の目的は佐伯のかかりの広さを定量化し、それをを用いて語順の推定モデルを構築することである。次節では、動詞の結合価情報に基づくモデルについて述べる。

## 2.2. 結合価行列と優先度行列

日本語の動詞は、動詞によって必須要素としてとる格助詞の種類、および格助詞と結び付く名詞句の意味的な性質が異なっている。このような側面から文を分析しようとするのが結合価文法である[6]。たとえば、「歩く」という動詞はガ格の名詞句として動物性名詞句をとり、ヲ格として場所を表す名詞句をとる、といった分析ができる。このように、動詞がどのような意味的性質をもつ名詞句をどの格にとるかを規定するものを、その動詞の結合価パターンと呼ぶ。また、名詞句の意味的性質を表現するために、あらかじめ用意したいくつかの意味素性を用いて表す。上述の例では、「歩く」という動詞は[ANI:ガ, LOC:ヲ]という結合価パターンをもつと考えられる。ただし、ANI, LOCはそれぞれ動物性、場所性を表す意味素性である。

我々のモデルの基本的な考え方は、意味素性と格の各組合せについて、その組合せを結

合価パターンに含む動詞の数でかかりの広さを定量化するというものである。意味素性と格の組合せは後置詞句の特徴を規定すると考えることができるので、このモデルによって後置詞句のもつかかりの広さの定量的な値を計算することができる。したがって、どの後置詞句が先行しやすいかという指標を与えることができる。以下では、このモデルを定式化する。

ある動詞がある格をとらないときは、その格が仮想的な意味素性NONをとると考えると、結合価パターンは $N_p$ 項行ベクタで表現できる。これを結合価ベクタと呼ぼう。ただし、 $N_p$ は格の数で、ここでは $N_p = 8$ である。結合価ベクタにおける要素の位置は考察の対象としている8つの格助詞に対応し、ベクタの要素は意味素性となる。たとえば、「歩く」は次のような結合価ベクタを持つ。

$$\begin{array}{cccccccc} \text{ガ} & \text{ヲ} & \text{ニ} & \text{カラ} & \text{ヘ} & \text{ト} & \text{ヨリ} & \text{デ} \\ [ \text{ANI} & \text{LOC} & \text{NON} & \text{NON} & \text{NON} & \text{NON} & \text{NON} & \text{NON} ] \end{array}$$

一般に1つの動詞は統語的、意味的な違いによって複数の結合価ベクタをとることがある。以下、「動詞」は結合価パターンの違いを考慮したものであるとする。したがって、各動詞は唯一の結合価ベクタを持つことになる。

ここで、 $N_v$ 個の動詞について、各動詞の結合価ベクタを縦方向に接続した $N_v$ 行 $N_p$ 列の行列 $VM$ を考える。これを結合価行列と呼ぶ。

$$VM = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1N_p} \\ s_{21} & s_{22} & \cdots & s_{2N_p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N_v 1} & s_{N_v 2} & \cdots & s_{N_v N_p} \end{bmatrix}$$

次に、各意味素性について、その意味素性 $N_v$ 個を要素とする $N_v$ 項行ベクタを考え、これらの行ベクタを意味素性の数 $N_s$ だけ縦方向に接続した次のような $N_s$ 行 $N_v$ 列の行列 $SM$ を考える。これを意味素性行列と呼ぶ。

$$SM = \begin{bmatrix} s_1 & s_1 & \cdots & s_1 \\ s_2 & s_2 & \cdots & s_2 \\ \vdots & \vdots & \ddots & \vdots \\ s_{N_s} & s_{N_s} & \cdots & s_{N_s} \end{bmatrix}$$

$N_v$

意味素性行列を結合価行列に左からかけて得られる $N_s$ 行 $N_p$ 列の行列を優先度行列と呼ぶ。ただし、意味素性同士の積は以下のように定義する。

$$s_i \times s_j \begin{cases} 1 & \text{if } s_i = s_j \\ 0 & \text{otherwise} \end{cases}$$

すなわち、優先度行列  $PM$  は、

$$PM = SM \times VM = \begin{matrix} & p_1 & p_2 & \cdots & p_{N_p} \\ \begin{matrix} s_1 \\ s_2 \\ \vdots \\ s_{N_s} \end{matrix} & \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1N_p} \\ n_{21} & n_{22} & \cdots & n_{2N_p} \\ \vdots & \vdots & \ddots & \vdots \\ n_{N_s 1} & n_{N_s 2} & \cdots & n_{N_s N_p} \end{bmatrix} \end{matrix}$$

優先度行列の各要素  $n_{ij}$  は  $N_v$  個の動詞の中で意味素性  $s_i$  と格  $p_j$  の組合せをとることのできる動詞の数を示している。したがって、優先度行列の要素の値が小さい場合には、その要素に対応する意味素性と格の組を結合価ベクタに含む動詞が少ないということなので、それだけ動詞を限定する力が強いことになる。逆に優先度行列の要素の値が大きい場合には、その意味素性と格の組は動詞を限定しにくいということになる。優先度行列が佐伯のいかかりの広さに関する定量的な尺度を与えるというのが我々の主張である。したがって、後置詞句の語順を考えると、優先度行列の要素の値が大きい格と意味素性を持つ後置詞句ほど先行しやすいことが予想できる。ここで注意して欲しいのは我々の議論では、単にどの格が先行しやすいかだけでなく、格とそれに結び付く意味素性までも考慮している点である。これにより、より詳細な語順に関する議論が可能になる。これについては3章で述べる。

ここで、確率理論の観点からかかりの広さについて考察する。優先度行列の各要素の値を優先度行列のすべての要素の値の総和で割ると、動詞の出現確率が一様分布にしたがうと仮定したときの意味素性と格の組合せの出現確率が得られる。すなわち、かかりの広さは動詞の出現確率が一様分布にしたがうと仮定したときの後置詞句（意味素性と格の組）の出現確率と考えることができる。この立場に立つと、かかりの広さが広い、つまり、出現確率の高い後置詞句が先行しやすいという傾向は自然な帰結である。

### 2.3. 結合価行列の縮退

優先度行列は、辞書中の、あるいは、聞き手の意識にのぼっているすべての動詞に関して、どのような後置詞句が入力されやすいかという優先度を与える。これに対して、ある後置詞句が入力された後にどのような後置詞句が入力されやすいかを考えよう。たとえば、名詞句の意味素性が  $s_i$ 、格助詞が  $p_j$  である後置詞句が入力された時点での優先度行列は

どのようになるかを調べよう。この場合、格  $p_j$  に意味素性  $s_i$  をもたない動詞が文末に現れることはないので、結合価行列において  $p_j$  列の要素に値  $s_i$  を持たない行は削除することができる。次に  $p_j$  の意味素性が決定されたので、結合価行列の  $p_j$  に相当する列も削除することができる。この操作を結合価行列の縮退と呼ぶ。

縮退した結合価行列に意味素性行列を左からかけると、意味素性  $s_i$  を持つ名詞句と格助詞  $p_j$  からなる後置詞句が入力された時点での優先度行列が得られる。この優先度行列は後置詞句  $(s_i, p_j)$  が入力されたという条件つきのかかりの広さを表している。以下、同様に新しい後置詞句が入力されるたびに、結合価行列の縮退と優先度行列の計算をおこない、次に入力される後置詞句の優先度を計算することができる。

後置詞句が次々に入力されると結合価行列を縮退することができるが、この縮退によって文末に現れる可能性のある動詞の候補も徐々に絞られる。何も入力されない状態ではすべての動詞が候補となる。各動詞の結合価ベクトルは固有の非NON要素を持っており、後置詞句が入力されるたびに結合価行列の縮退によって非NON要素が削除される。初期状態の非NON要素の数から現在の非NON要素の数を引き、これを初期状態の非NON要素の数で割ったものをその結合価ベクトルの充足度と呼ぶことにする。すると、各動詞の持つ結合価ベクトルの充足度によって次入力動詞であると仮定した場合の動詞の優先度を求めることができる。

たとえば、ある段階で結合価ベクトルの充足度が1の動詞はその直後に動詞が入力されるとしたらもっとも入力されやすい動詞であるといえる。もし、次の入力が動詞ではなく後置詞句であれば、結合価行列の縮退によって充足度1の結合価ベクトルを持つ動詞に相当する行は削除される。なお、我々のモデルは後置詞句、動詞の各々については優先度を与えるが、次入力動詞か後置詞句か動詞かに関する優先度は与えない。

### 3. 実験

本章では、IPAL動詞辞書の結合価パタンの情報を用いて優先度行列を計算し、佐伯の提案した語順傾向と比較することによって我々の推定モデルの有効性を検討する。

#### 3.1. IPAL動詞辞書

IPAL動詞辞書[10]は計算機による日本語処理のために作成された動詞辞書で、基本的な和語動詞861語に関する統語的、意味的情報を含んでいる。ここでは、実験に必要な情報についてのみ述べる。

IPAL 動詞辞書の見出し語はひらがな表記で、同音異義語は同音異義語番号を見出し語にふることによって区別している。861語という数は同音異義語を区別したときの数である。一般にひとつの動詞が複数の結合価パターンをとることができ、結合価パターンによって意味が異なる場合がある。IPAL 動詞辞書ではサブエントリ番号によって結合価パターンおよび意味の違いを区別している。IPAL 動詞辞書は3370のサブエントリを持っている。まとめると、IPAL 動詞辞書の各見出し語は次のような階層を持つ。

見出し(ひらがな表記) > 同音異義語番号 > サブエントリ番号

IPAL 動詞辞書で扱う格は、ガ、ヲ、ニ、カラ、へ、ト、ヨリ、デ、φである。ここで、φは格助詞を必要としないはだか格を表す。本実験では、はだか格以外の8つの格助詞について考慮する。それぞれの格には、その格がとりうる名詞句の意味素性が割り当てられている。IPALでは20種の意味素性を設定している。

格、意味素性は交替可能な場合があり、その場合は交替可能な要素を“/”で区切っている。たとえば、「あおぐ」の同音異義語番号001、サブエントリ番号002の結合価パターンは、

[HUM/ORG:ガ, HUM/ORG:ニ/カラ\*, ABS:ヲ]

であるが、交替可能な要素をすべて展開すると次の8通りの結合価パターンになる。

[HUM:ガ, HUM:ニ\*, ABS:ヲ]      [HUM:ガ, ORG:ニ\*, ABS:ヲ]

[HUM:ガ, HUM:カラ\*, ABS:ヲ]      [HUM:ガ, ORG:カラ\*, ABS:ヲ]

[ORG:ガ, HUM:ニ\*, ABS:ヲ]      [ORG:ガ, ORG:ニ\*, ABS:ヲ]

[ORG:ガ, HUM:カラ\*, ABS:ヲ]      [ORG:ガ, ORG:カラ\*, ABS:ヲ]

また、“\*”は任意要素を表し、結局、このエントリは10種類の結合価パターンを持つことになる。ただし、この実験では任意要素と必須要素の区別はおこなわない。すなわち、“\*”は単に無視する。

### 3.2. 結合価行列と優先度行列

まず、IPAL 動詞辞書から見出し語、同音異義語番号、サブエントリ番号、結合価パターン情報を抽出する。前項で述べたように、結合価パターンの交替要素について結合価パターンを展開し、展開した結合価パターンに0から順に結合価パターン番号をふる。したがって、各動詞は統語的、意味的情報の違いも含めて、見出し語、同音異義語番号、サブエントリ番号、結合価パターン番号の4つ組で表現できる。ただし、ここでいう「動詞」は2章で述べたように統語的、意味的な違いも考慮したものである。こうして8829個の動詞が得られ



るが、ここで、ガーガ、ニーニのように同じ格を複数持つ結合価パターンを持つ動詞については結合価ベクトルが構成できないので削除する。これを除くと8062個の動詞が残る。

以上のようにして得られた8062の動詞から2章で述べたように、結合価ベクタ、結合価行列を構成する。IPAL動詞辞書の意味素性の数は20なので20行8062列の意味素性行列を作り、それを結合価行列に左からかけると20行8列の優先度行列を得る。この優先度行列を表1に示す。また、その要素の値について降順に整列したときの上位20の意味素性と格助詞の組を表2に示す。

### 3.3. 格の優先順序

かかりの広さが広い、すなわち優先度行列の要素の値が大きいかかり要素ほど先行するという佐伯の仮説にしたがえば、かかりの広さという観点から各動詞について最適な後置詞句配置が決定できる。各動詞について以下の手続きにより、最適な後置詞句の配列を計算する。

(1) 動詞を1つ取り出し、以下の処理を結合価パタンのすべての非NON要素についてお

表1 IPAL動詞辞書の優先度行列(初期状態)

	ガ	ヲ	ニ	カラ	ヘ	ト	ヨリ	デ
CON	350	756	146	55	14	12	4	86
ANI	436	96	56	2	0	6	0	4
HUM	3898	566	600	150	34	78	5	4
ORG	1515	207	397	176	55	48	5	5
PLA	39	19	12	0	0	0	0	3
PAR	38	237	101	29	5	1	0	107
NAT	88	34	19	2	3	2	0	10
PRO	422	501	143	30	14	22	5	479
PHE	228	195	83	8	0	7	0	162
ABS	603	1654	660	138	40	60	4	310
ACT	85	345	168	10	1	2	0	204
MEN	25	121	32	2	0	4	0	16
LIN	67	120	50	13	4	8	0	48
CHA	23	89	48	8	0	4	2	14
REL	2	26	10	0	0	0	0	0
LOC	109	497	954	588	401	7	0	6
TIM	21	55	41	2	1	3	4	2
QUA	3	38	68	26	9	3	0	22
DIV	81	108	50	6	0	15	4	12
---	0	0	0	0	0	130	0	0

こなう。

- (2) その動詞の結合価ボタンに含まれる後置詞句（意味素性と格の組）の中から現在の優先度行列で最大の値を持つものを選択する。
  - (3) 選択した後置詞句に基づいて結合価行列を縮退し、新しい優先度行列を計算する。
  - (4) 新しい優先度行列と結合価ボタン中の残りの後置詞句について(2)から繰り返す。
- このようにして得られた各動詞の最適な後置詞句配列から2つの格の優先関係を計算する。

表2 初期状態の優先度行列の要素の一部

順位	意味素性：格	動詞数	順位	意味素性：格	動詞数
1	HUM:ガ	3898	11	PRO:ヲ	501
2	ABS:ヲ	1654	12	LOC:ヲ	497
3	ORG:ガ	1515	13	PRO:デ	479
4	LOC:ニ	954	14	ANI:ガ	436
5	CON:ヲ	756	15	PRO:ガ	422
6	ABS:ニ	660	16	LOC:へ	401
7	ABS:ガ	603	17	ORG:ニ	397
8	HUM:ニ	600	18	CON:ガ	350
9	LOC:カラ	588	19	ACT:ヲ	345
10	HUM:ヲ	566	20	ABS:デ	310

表3 格の優先関係

X	Y	X > Y	X < Y	X	Y	X > Y	X < Y
ガ	ヲ	4384	1266	ニ	カラ	395	45
ガ	ニ	2824	804	ニ	へ	0	9
ガ	カラ	986	259	ニ	ト	56	5
ガ	へ	475	106	ニ	デ	300	77
ガ	ト	409	0	カラ	へ	287	13
ガ	ヨリ	32	1	カラ	ト	0	2
ガ	デ	1386	105	カラ	デ	102	98
ヲ	ニ	1408	790	へ	デ	35	46
ヲ	カラ	601	284	ト	デ	18	41
ヲ	へ	305	113	ヨリ	デ	0	17
ヲ	ト	204	31				
ヲ	ヨリ	11	0				
ヲ	デ	836	281				

たとえば、結合価パタンの最適配列は次のような形式で得られる。

443 28 HUM:ガ HUM:ニ ACT:ヲ

ここで、最初の数字は結合価パタンの識別番号、次の数字はその結合価パターンを持つ動詞の数、そして残りが後置詞句の配列である。この項目からは、ガ>ニ、ガ>ヲ、ニ>ヲという優先関係にそれぞれ28点ずつを与える。このようにして得られた結合を表3に示す。表3から次のような格の優先関係が読みとれる。

ガ>ヲ>ニ>カラ>デ

しかしながら、1つの格は一般に複数の役割を持つので各々の格がどのような意味素性の名詞句をとるときにどのような優先関係を持つのかという点を含めて議論しないと意味がない。そこで、1章で紹介した佐伯の語順に関する観察に沿って実験結果を検討する。

### 3.4. 検討

#### (傾向1) 位格は他の格に先行する

佐伯は位格の格助詞としてニ、デ、カラ、ヲを考察の対象としている。後述する(傾向3)とも関係するが、ガに先行できるのは一般に位格だけである。したがって、ここでは、ガと位格の関係について調べる。

まず、ニについて、ニがガに先行するのは動詞が所動詞であり、主格がモノである傾向が強く、逆にガが先行するのは動詞が能動詞であり、主格がヒトである傾向が強いという佐伯の観察がある。我々の手法では、格のとり意味素性を手がかりにこの分析をおこなうことができる。実際、ニ>ガ、ガ>ニとなるそれぞれの場合について意味素性の分布を調べると、ニ>ガの場合、ガの意味素性はABS(抽象物)、PRO(生産物)、CON(具体物)、PHE(現象)、ANI(動物)がいずれも10%以上を占め、これら5つの合計で72%となる。ヒトも含まれる可能性のあるANIは単独では約13%を占めるに過ぎない。また、この時のニはLOC(場所)とABSで77%以上を占める。一方、ガ>ニの場合、ガの意味素性はHUM(人間)とORG(組織)で86%以上を占めている。格がとる名詞句の性質という点では佐伯の観察と同様の結果が得られた。

次にデについて考えよう。デはニに比べるとガに先行しにくいという佐伯の主張は、表3のガとニ、ガとデの優先関係の比率に現れている。ガ>ニとなるのは、2824:804(3.5:1)であるが、ガ>デとなるのは、1386:105(13.2:1)という比率である。また、意味素性についてニと同様に調べてみると、ガ>デの場合のガはHUMとORGで82%を占める。また、この時のデはPROとABSで50%以上となり、ガ>デの場合は

デが具格を表す場合が多いと推定できる。逆に、デ > ガの場合にガがHUMをとる例はなく、ANIが24%をとるだけである。

カラについて、佐伯はガ > カラとなる場合は、着格(ニ、へ)も同時にとりやすいが、カラ > ガの場合は着格をとりにくいという観察を述べているが、我々の実験結果からはこのような観察は得られなかった。

最後に、ヲはガに先行しにくいという傾向は表3から読みとれる。特に、意味素性にLOCをとる場所の位格については、ガに先行するものはわずか15%しかなかった。

(傾向2) トキの位格はトコロの位格に先行する

位格内の優先順位については、トキとトコロの位格を両方含む結合価パターンがないために評価できなかった。これは、結合価パターンが主に必須格を対象としているのに対して、位格は必ずしも動詞の必須格にならないためであると考えられる。

(傾向3) ガは位格を除く他の格に先行する

ガの優先性は表3においていずれもガが他の格に先行していることから読みとれる。

(傾向4) 与格のニは対格のヲに先行する

ニ > ヲ, ヲ > ニそれぞれの場合について、ニのとり意味素性について調べたところ、ニ > ヲの場合はニ格に与格補語となりやすいと思われる意味素性HUMをとる場合が全体の25%であるのに対し、ヲ > ニの場合は16%であった。この差は9%であり、この実験からは(傾向4)について明確な傾向は観察できなかった。

(傾向5) 発格(カラ)は着格(ニ、へ)に先行する

意味素性をLOCに限定して、カラとニ、への組合せの優先関係を調べたところ、カラとニについてはLOC:ニ > LOC:カラが98%を占め、カラとへについてはLOC:カラ > LOC:へが100%を占めていた。この結果からは、発格が着格に先行するとはいえない。発格、着格はかかりの広さという観点からはどちらも同等であり、その順序関係は別の制約によって規定されることが考えられる。

まとめ

IPALの動詞辞書について我々のモデルが推定する語順傾向を佐伯の観察と比較した結果、(傾向1)と(傾向3)については強い相関が、(傾向4)については弱い相関が観察できる。特に(傾向1)に関して、我々のモデルではニ格にどのような意味素性を持つ名詞句をとるかによって細かい分析が可能である点に注目して欲しい。宮島は川端康成の新文章読本から抽出した文について語順傾向を調査している[3]。その結果、(傾向4)

と(傾向5)を除き, 佐伯とほぼ同様な傾向を観察している。宮島の結果と我々の実験結果から(傾向4)は必ずしも強い傾向ではないことが予想できる。

#### 4. おわりに

本論文では佐伯の提案したかかりの広さという概念を用い, 後置詞句の並びと動詞という構造を持つ日本語の単文の語順を推定するモデルを提案した。このモデルは, 意味素性と格の各組合せについて, その組合せをとる動詞の数を動詞の結合価パターンから計算し, もっとも多くの動詞がとる組合せがかかりの広さが広いと考え, 文の前方に配置されるという考え方に基いている。また, 実際にIPAL基本動詞辞書から結合価パターンを抽出し, IPAL中の各動詞についてかかりの広い順に後置詞句を配列し, その語順傾向が佐伯の分析に一致することを示した。このモデルの自然言語処理への応用としては, 音声認識における認識結果の候補の絞り込み, 名詞句の意味のあいまい性の解消, 文生成における基本語順の決定などが考えられる[8]。本論文では, 文脈の影響, 省略語, 係り助詞などについては考慮していないが, これらの現象を我々のモデルでどのように扱うかについては, 文献[8]を参照して欲しい。

#### 謝辞

貴重な資料を提供していただきました国立国語研究所の宮島達夫先生に感謝いたします。

#### 参考文献

- [1] 井上和子(編). 日本語の基本構造. 三省堂, 1983.
- [2] 久野暉. 談話の文法. 大修館書店, 1978.
- [3] 宮島達夫. カカリの位置. 計量国語学, 23:3-11, 1962.
- [4] 佐伯哲夫. 現代日本語の語順. 笠間書院, 1975.
- [5] 児玉徳美. 語順の普遍性. 山口書店, 1987.
- [6] 石綿敏雄, 荻野孝野. 結合価からみた日本文法. 文法と意味I, 朝倉書店, 1983.
- [7] 中島文雄. 日本語の構造 — 英語との対比 — . 岩波新書(黄版)第373巻, 岩波書店, 1987.
- [8] 徳永健伸, 岩山真, 乾健太郎, 田中穂積. 日本語語順の推定モデルとその応用. 情報処理学会 自然言語処理研究会, NL81-2, 1991.
- [9] 北原保雄. 日本語助動詞の研究. 大修館書店, 1981.
- [10] 計算機用日本語基本動詞辞書. 情報処理振興事業協会, 1986.

(1991年5月31日受付)