自 然 言 語 処 理 81－2
（1991. 1. 17）

日本語語順の推定モデルとその応用

徳永健伸, 岩山真, 乾健太郎, 田中穂積
東京工業大学 工学部
take@cs.titech.ac.jp

日本語のかかり要素の語順を規定すると考えられる要因のひとつである「かかりの広さ」という概念を用い, 日本語単文の次入力を推定するモデルを提案する. 「かかりの広さ」は直観的には, かかり要素がかかる動詞に対する限定力と考えることができ, 日本語では, かかりの広い要素が先行するとされている. モデルの基本的な考え方は, 動詞の結合価パタンから, 特定の意味素性と格の組合せをとりうる動詞の数を計算し, この動詞の数でかかりの広さを定量化しようというものである. また, このモデルが音声認識, 名詞句の意味の推定, 文生成のための語順の基礎的情報として利用できることも述べる.

A Estimation Model of Japanese Word Order and Its Applications for Natural Language Processing

TOKUNAGA Takenobu, IWAYAMA Makoto, INUI Kentaro and TANAKA Hozumi
Department of Computer Science, Tokyo Institute of Technology
(2-12-1 Ôokayama Meguro-Ku Tokyo 152 Japan)

This paper proposes a estimation model of the word order for Japanese simple sentences, which consist of several postpositional phrases followed by a verb. In Japanese word order, the postpositional phrase which has semantically loose relation with the verb, tends to precede in the sentence. As each postpositional phrase is obtained, our model estimates the semantic type and the postposition of the following postpositional phrase by using this tendency. The model also gives a estimation about the verb, which locates in the end of the sentences in Japanese. Our model can be applied to the word sense disambiguation of noun phrases, the speech recognition and the sentence generation.

1 はじめに

日本語は語順が比較的自由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 [8]. 一方, 述部にかかる要素間の語順については, 述部の語順ほど厳格な規則は見い出されていない.

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, 中でも依存関係を最重要視している. しかし, 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いってはいない.

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もとづく語順傾向と構文条件にもとづく語順傾向の大きく 2 つに分類している.

成分条件にもとづく語順傾向とは, 補語そのものの意味, 機能にそなわった支配条件にもとづいて生じる語順傾向のことで, これは児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 に優先することを意味する.

一方, 構文条件にもとづく語順傾向とは, 補語の構文的な意味, 機能にそなわった支配条件にもとづいて生じる語順傾向のことで, 児玉のいう依存関係, 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, 成分条件にもと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 (1) から (5) の語順傾向を説明している. また, 構文条件にもとづく語順傾向について, かかり先のあいまい性という観点から (6) から (9) の語順傾向の必然性を説明している.

本稿では, 佐伯のかかりの広さを定量化し, かかりの広さにもとづいた日本語語順の推定モデルを提案する. このモデルは日本語の文要素 (後置詞句または, 動詞に限定) が次々に入力されるとき, 各入力の段階での次入力に対する推定を優先度という形で与えるものである. このモデルの妥当性を検証するために情報処理振興事業協会技術センターが作成した計算機用日本語基本動詞辞書 (以下 IPAL 動詞辞書) を用いた実験をおこなった. 以下, 2 章では, 佐伯の語順に関する知見について説明し, その定量化と我々のモデルについて述べる. 3 章では, 実験に用いた IPAL の内容と実験について述べる. 4 章では, 我々のモデルの自然言語処理への応用について述べ, 最後に 5 章では, モデルの拡張について述べる.

2 語順の推定モデル

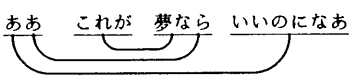
2.1 かかりの深さと広さ

日本語の単文は, 単純化すると次のような規則で表現できる [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」.

ここで, "+" は直前の要素の 1 回以上の繰り返しを, "|" は選言を表す. 本稿では, 助詞としてガ, ヲ, ニ, カラ, へ, 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である. たとえば,「ああ, これが夢ならいいのになあ」という文は, 次のようなかかり受け関係を持っている.



すなわち,「ああ」は「夢なら」と「いいのになあ」にかかり,「これが」は「夢なら」にかかっている.「ああ」と「これが」を比べるとかかりの深さが深い, つまり, より文末の受けにかかる「ああ」の方が先行するということである. これは, 非交差の原則とも関係する. また, 1 章でも触れたように, 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. 佐伯はこの指標を実験によって求め, おおむね次のような結果を得ている.

感動語 > 接続語 > 題目語 > 評釈語 > 時間的情態語
> 主格補語 > 情態語 > 着格補語 > 対格補語

このように, 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使えない.

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, 述部を厳しく限定するほどかかりの広さは狭くなる. 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我々の考察対象となりうる.

2 章の最初でも触れたように, 英語などの言語と異なり, 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大きな特徴がある. 一般に動詞は文全体の意味の中心的な役割を担うので, 言語によるコミュニケーションを考えると, これは一見不合理なように思える. しかし, 人間は実際に何の不自由もなく日本語を通じてコミュニケーションすることができるし, 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, 文の後半を推定しながら理解していることを示唆している. コミュニケーションの効率化という観点から考えると, 聞き手が発話の後半を推定しやすいような発話が望ましいということになる. 佐伯はこのような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ことが望ましいとしている. 言い替えれば, 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直しの訂正に要するコストをあげている. たとえば,

(1) あら → あなたは → ×ロンドンへ, (いや) ○大阪で → 留守番ね.

(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に, 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いたとすると, (1) の場合, 述部を強く限定するかかり要素 (ロンドンへ) が述部に近い位置にあるので, 訂正は少なくてすむが (この場合 1 要素), (2) では, 述部を限定するかかり要素 (涙が) の間に他の要素があるために 2 要素の訂正が必要となる. このような語順は述語の表していることがらの構成要素を周辺的なものから順次列挙してい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もと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, および各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の名詞句をとり, 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] という格パタンを持つことになる. 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.

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, ここで N_p は格の数で, ここでは $N_p = 8$ である. 各ベクタにおいて要素の位置はここで考察の対象としている 8 つの格助詞に対応し, 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.

ガ	ヲ	ニ	カラ	ヘ	ト	ヨリ	デ
[ANI	LOC	NON	NON	NON	NON	NON	NON]

一般に 1 つの動詞は統語的, 意味的な違いにより, 複数の結合価ベクタをとることがある. 以下,「動詞」といえば, このような統語的, 意味的な下位分類まで考慮したものを指すものとする. したがって, このように考えると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, 各結合価ベクタを縦方向に連接した N_v 行 N_p 列の行列を考えよう. これを結合価行列と呼ぶ. すなわち, 動詞, 格助詞をそれぞれ, v, p とし, 動詞 v の結合価ベクタの格助詞 p の位置の意味素性を $S(v,p)$ で表すと, 結合価行列 VM は次のように表現できる.

$$VM = \begin{bmatrix} S(v_1,p_1) & S(v_1,p_2) & \cdots & S(v_1,p_{N_p}) \\ S(v_2,p_1) & S(v_2,p_2) & \cdots & S(v_2,p_{N_p}) \\ \vdots & \vdots & \ddots & \vdots \\ S(v_{N_v},p_1) & S(v_{N_v},p_2) & \cdots & S(v_{N_v},p_{N_p}) \end{bmatrix}$$

次に, 同一の意味素性を要素として持つ N_v 項行ベクタを意味素性の数 N_s だけ縦方向に連接した次のような N_s 行 N_v 列

の行列SMを考えよう．これを意味素性行列と呼ぶ．

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

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されるものとする．

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

すなわち，優先度行列PMは，

$$PM = SM \times VM = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1N_p} \\ n_{21} & n_{22} & \cdots & n_{2N_p} \\ \vdots & \vdots & \ddots & \vdots \\ n_{N_s1} & n_{N_s2} & \cdots & n_{N_sN_p} \end{bmatrix}$$

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．この優先度行列が佐伯のいう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ということが予想できる．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というのではなく，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．

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，何の情報も与えられない時に，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直後に次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．まず，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の意味素性は決定されたので，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ぼう．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．この優先度行列はこの後置詞句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，その次にく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．もちろん，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タは固有の数の非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タ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タの充足度によって，次入力が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タの充足度が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だったら，結合価行列の縮退によって，充足度1の結合価ベクタを持つ動詞は棄却されることに注意しよう．また，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，次入力が後置詞句か動詞かに関する優先度は与えないことにも注意して欲しい．以上をまとめると，我々のモデルでは，優先度行列と結合価ベクタの充足度を用いることにより，動詞，後置詞句別の次入力の優先度を計算することができる．

ここで，結合価行列の縮退と優先度行列の計算に関する計算コストについて調べよう．いま，2つの意味素性の比較にかかるコストをC_{cmp}，2つの自然数の加算のコストをC_{add}とすると，結合価行列の縮退は，入力された後置詞句の意味素性と各動詞の該当する格の意味素性を動詞の数N_vだけ比較すればよいので，$C_{cmp} \times N_v$のコストで計算できる．また，優先度行列の計算は，行列式の計算であるから，$N_s \times N_p \times (C_{cmp} + C_{add}) \times N_v$のコストで計算できる．ここで，$N_s$，$N_p$はそれぞれ意味素性の数，格の数である．意味素性同士の積は前節で定義したように意味素性の比較なので，そのコストはC_{cmp}に等しい．したがって，新たに後置詞句が入力された時に，結合価行列を縮退し，優先度行列を計算するには，$C_{cmp} \times N_v + N_s \times (N_p - 1) \times (C_{cmp} + C_{add}) \times (N_v - N_v')$のコストがかかる．ここで，$N_v'$は結合価行列の縮退によって削除された動詞の数である．これらの定数の中で，一般的にもっとも大きくなるのは動詞の数N_vであるが，結合価行列の縮退と優先度行列の計算にはN_vに対して線形の計算コストしかかからない．また，後置詞句の入力にともなって結合価行列は小さくなるので計算コストも減少する．

3 実験

本章では，IPAL動詞辞書の結合価パタンの情報をもちい，実際に優先度行列を計算し，我々の推定モデルの有効性について考察する．

3.1 IPAL 動詞辞書

IPAL 動詞辞書 [9] 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, 同音意義語は同音意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 語という数は同音意義語を区別したときの数である. 見出し語数では 802 語となる. 一般にひとつの動詞が複数の結合価パタンをとることができ, 結合価パタ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 > 同音意義語番号 > サブエントリ番号

IPAL 動詞辞書で扱う格は, ガ, ヲ, ニ, カラ, ヘ, 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ふられている. 意味素性は表 1 に示す 20 個からなる. このうち "---" は文要素を表す. ---は「〜ト思う」などのように, ト格にしか現れない. 正確にはこれらの意味素性には 2 レベルの階層があり, CON は ANI から PRO の上位素性, ABS は ACT から QUA までの上位素性となっているが, 簡単のため実験では, すべての素性を同一に扱った.

表 1 IPAL の意味素性

略号	意味素性名	例
CON	具体物	
ANI	動物	鼠, 牛, 虎
HUM	人間	父, 友達, 先生
ORG	組織・機関	政府, 企業, 大学
PLA	植物	せり, なずな
PAR	生物の部分	手, 足, 鼻
NAT	自然物	海, 山, 川
PRO	生産物・道具	はさみ, 自動車
PHE	現象	雨, 風, 匂い
ABS	抽象物	
ACT	動作・作用	散歩, 勉強
MEN	精神	心, 喜び, 苦悩
LIN	言語作品	小説, レポート
CHA	性質	優しさ, 美しさ
REL	関係	原因, 結果, 根拠
LOC	空間・方角	西, 左
TIM	時間	月曜, 朝
QUA	数量	3 人, 10kg, 10m
DIV	制限緩やか	
---	文要素	「太郎が死んだ」

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 "/" で区切ってある. たとえば, 「あおぐ」の同音意義語番号 001, サブエントリ番号 002 の結合価パタンは,

[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, 同音意義語番号, サブエントリ番号, 結合価パタン情報を抽出する. 前節で述べたように, 結合価パタンの交替要素について結合価パタンを展開し, 展開した結合価パタンに 0 から順に結合価パタン番号をふる. 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, 同音意義語番号, サブエントリ番号, 結合価パタン番号の 4 つ組で表現できる. ただし, ここでいう「動詞」は 2.2 節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−ガ, ニ−ニなどのように同じ格を複数持つ結合価パタ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.

これらの動詞のいくつかは同一の結合価パタンを共有している. 結合価パタンによって動詞を分類すると 2225 の結合価パタンが得られる. 表 2 は各結合価パタンを持つ動詞の数の分布を表している. 表 2 からほとんどの結合価パタンのグループが 20 以下の動詞しか含まないということがわかる. これは, 結合価パタンが決まれば候補の動詞がほとんどの場合 20 以下に抑えられるということを示している.

表 2 動詞数別の結合価パタンの頻度

動詞の数			頻度
100	〜		3
80	〜	100	4
60	〜	80	5
40	〜	60	16
20	〜	40	35
	〜	20	2162

以上のようにして得られた 8062 の動詞から 2.2 節で述べたように, 結合価ベクタ, そして結合価行列を構成する. IPAL の辞書では, 意味素性の数が 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列の優先度行列が得られる．この優先度行列の各要素についてソートした上位20を表3に示す．

表3　初期状態の予測行列の要素の一部

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

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，各動詞についてかかりの広さという観点から最適な後置詞句の配置が優先度行列を用いて決定できる．すなわち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要素についておこなう．

(2) その動詞の結合価パタン中で，初期状態の優先度行列で最大の値を持つ後置詞句(意味素性と格の組)を選択する．

(3) 選択した後置詞句にもと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パ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．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果を表4に示す．

表4からおおよそ，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．

表4　格の優先関係

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

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．(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，格のとる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てみると，ニ＞ガの場合，ガの意味素性はABS, PRO, CON, PHE, ANIがいずれも10%以上を占め，これら4つの合

計で 72%となる．ヒトも含まれる可能性のある ANI (動物) 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 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 と ORG で 86%以上を占めている．格がとる名詞句の性質という観点からは，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 4 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，4384 : 1266 (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，この時のガは 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，ヘ) 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 4 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ンがないために評価できなかった．これは，結合価パタンが主に必須格を対象としているのに対して，位格は必ずしも必須格として動詞がと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 4 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，ニのとる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．10%程度の差ではあるが (傾向 4) が観察できる．

(傾向 5) 発格 (カラ) は着格 (ニ，ヘ) 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と考えられる．

4 応用

我々の推定モデルは日本語の後置詞句の並びと動詞からなる単文について，文要素 (後置詞句または動詞) が次々に入力されると仮定して，次要素の優先度を計算するものである．この点においては，入力の意味解釈を漸次的におこなうアプローチ [10, 2] と同様であるが，我々のモデルは各要素が入力された時点での次の要素を意味素性と格の組合せに対する優先度という形で推定できる点が特徴である．以下，この推定モデルの自然言語処理への応用について考察する．

4.1 名詞句の語義のあいまい性の解消

日本語の場合，文の統語的，意味的な主辞は文末の動詞であり，その動詞にかかる名詞句の意味の選択制限によって動詞の意味，ひいては文全体の意味が決まってくる．我々の推定モデルでは，動詞の意味のあいまい性に関しては，2.3 節で述べたように，動詞の結合価ベクタの充足度という尺度で優先度を与えることができる．しかし，一般には名詞句の意味にもあいまい性が存在するので，理想的には，名詞句と動詞が互いに相手の意味のあいまい性を解消するようなモデルが望ましい．我々の推定モデルは後置詞句が入力されたら，その時点でその格をとる名詞句の意味素性の優先度を与えることができる．したがって，名詞句の意味素性という限られた範囲ではあるが，名詞句の意味のあいまい性も部分的には解消していることになる．

4.2 音声認識における候補の絞り込み

音声認識では認識結果の候補をさまざまな情報を利用して絞り込むことが重要となる．我々のモデルが与える優先度情報は，候補を絞り込むための情報のひとつとして使用できる．特に，日本語の音声認識においては名詞句などの自立語に比べ，セグメント数の小さい助詞などの附属語の認識率が低いという傾向がある [3]．現在の日本語の言語処理，特に意味解析が助詞の情報に非常に依存していることを考えると音声から意味構造を抽出する音声理解のためには，助詞の認識率を向上させる必要がある．我々の推定モデルは名詞句の意味素性と格の組合せに対して優先度を与えるが，名詞句の認識結果を手がかりに助詞を推定することもできる．

4.3 文生成における語順の決定

文を生成する立場からすると，出力すべき内容はすべて与えられているので，文を解析する場合と異なり，名詞句や動詞の意味のあいまい性は存在しない．生成ではどのような順序で後置詞句を並べるかという語順が重要な問題となる．我々の推定モデルは佐伯のかかりの広さという概念に基づき，日本語ではかかりの広さの広い要素が先行するという仮説の上に構築したものである．したがって，3.3 節で格の優先順位を算出するときにおこなったように，出力すべき動詞の結合価パタンから最適な後置詞句配列を計算することができる．もちろん最終的な語順を決定するには，文脈の焦点や聞き手のモデルなど，さまざまな情報も考慮しなければな

らないが, 我々のモデルが与える情報は, 文を生成するときの語順に関するもっとも基本的な情報として利用することができる.

5 拡張

これまでの議論は, 文脈を考慮しない単文内における後続入力の推定に限られていた. 文脈情報を我々のモデルにどのように取り込むかは重要な問題である. また, 実際の文では, 文要素の省略も頻繁に生じる. 文要素の省略をどう扱うかも重要な問題である. 以下, これらの問題に対する我々のモデルの拡張について, その基本的な考え方を述べる.

5.1 文脈情報の扱い

2.2 節で導入した結合価行列では, 各動詞がすべて等確率で出現することを前提としており, この前提にもとづいて優先度行列を計算している. しかしながら, 実際には, 文脈的な情報や対象領域の性質から動詞の出現確率には偏りがあると考えるのがより現実的である. 我々の推定モデルにこのような文脈あるいは対象領域による動詞の偏りを導入するひとつの考え方は, 各動詞, すなわち, 結合価行列の列に重みを付加し, 優先度行列を計算する時にその重みを考慮した計算をおこなうことが考えられる. 具体的には, 意味素性同士の積を以下のように再定義する. ここで, w_vは, その意味素性を持つ動詞の重みである.

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

さらに, 文脈によっては, 焦点や提題化によって特定の要素が文の前方に位置しやすいことも考えられる. このような名詞句の前置されやすさに関する偏りも動詞の出現頻度の偏りと同様に, 意味素性行列の各行に重みを付加することによってモデルに導入することができる. ただし, 我々のモデルでは, あくまでも意味素性のレベルでしか名詞句の意味を扱えないので精度上の限界はある. 具体的には, 意味素性同士の積を以下のように再定義する. ここで, w_sは当該の意味素性の重みである.

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

5.2 省略語の推定

実際の文では, 文要素の省略が頻繁に起こる. 省略が起こると推定モデルで高い優先度で推定されている要素が入力されないで, より低い優先度の要素が入力されることになる. ある段階で高い優先度で推定された要素はその後の段階でも優先度の上位に位置する傾向があるので, 入力と各時点での優先度の履歴を調べれば, 過去に何度も高い優先度で推定されているにも関わらず, 実際には入力されていないような要素は省略された可能性があると判断できる. 具体的にどのような判断基準で省略要素を推定するかは, 実際のデータをもとにさらに詳しく調べる必要がある.

6 おわりに

本稿では, 佐伯の提案したかかりの広さという概念を用い, 日本語の単文の次入力を推定するモデルを提案した. 基本的な考え方は, 動詞の結合価パタンから, 意味素性と格の組合せをとりうる動詞の数を計算し, 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仮定に基づいている. また, 実際に IPAL 基本動詞辞書から結合価パタンを抽出し, 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よく一致することを確かめた. 我々のモデルは音声認識, 名詞句の意味の推定, 文生成のための語順の基礎的情報として利用できることも述べた. このモデルの拡張としては文脈情報の取り込みや省略要素の推定などが考えられるが, その基本的な考え方についても説明した. ただし, これらの拡張に際しては具体的なデータを用いてさらに精度の検証をおこなう必要がある. また, 本稿ではもっとも単純な構造である単文しか対象としなかったが, 埋め込み文を含む場合や, 受動化や使役化によって格交替を起こす場合なども含めてモデルの拡張を検討する必要がある.

謝辞

本稿のドラフトに対し, 有益なコメントを頂いた東工大の奥村学, 伊藤克亘両氏に感謝いたします.

参考文献

[1] 井上和子 (編), 日本語の基本構造, 三省堂, 1983.

[2] 奥村学, 田中穂積, 自然言語解析における意味的曖昧性を増進的に解消する計算モデル, 人工知能学会誌, 4(6), 1989.

[3] 岡田美智男, 伊藤彰則, 牧野正三, 城戸健一, 構文駆動型連続 DP 法による連続音声中からの活用語のスポッティング, 電子情報通信学会論文誌, J70-D(12), 1987.

[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 -, 岩波新書第 373 巻, 岩波書店, 1987.

[8] 北原保雄, 日本語動詞の研究, 大修館書店, 1981.

[9] 計算機用日本語基本動詞辞書, 情報処理振興事業協会, 1986.

[10] C. S. Mellish. *Computer Interpretation of Natural Language Descriptions*. Ellis Horwood, 1985.