

語順規則による言語の分類と 2 パラメータモデル

統計数理研究所 上 田 澄 江
統計数理研究所・総合研究大学院大学 伊 藤 栄 明

(1994 年 11 月 受付)

1. 序

語順規則によって世界の言語の分類を試みた場合、側置詞が分類の指標として重要であることは従来より指摘されてきた (Greenberg (1966), Hawkins (1983), Bybee (1990)). ここで側置詞とは前置詞・後置詞の総称であり、後置詞とは、学校へ、東京に、の助詞“へ”、“に”を指し、前置詞とは to school の“to”を指す。通常、側置詞による分類では前置詞・後置詞を共にもたない無側置詞言語は側置詞をもつ言語とは分類上関係のないものとして考慮の外とされてきた。しかし Tsunoda et al. (1995) では、階層クラスタ分析によって、その無関係とされた無側置詞言語は日本語を典型とする後置詞をもつ言語と同じカテゴリーに属するということが示された。

本稿ではそれに関する統計的解析について述べる。すなわち、側置詞をもたない「無側置詞言語」を「後置詞をもつ言語」の中に統合して、世界の言語を

「前置詞をもたない言語 (後置詞をもつ言語と無側置詞言語)」と「前置詞をもつ言語」

に 2 分割することの妥当性について統計的な議論を行う。また、言語の語順規則による分類は多数の項目に関係するようにみえるが、実は、本質的にごく少数の項目、例えば 2 項目 (2 パラメータ) で説明し得るだろうことについて考察する。

角田の語順表で扱われた 130 言語 (表 1~表 3, 角田 (1991)) は、言語の一般的な系統分類上の 37 語族 124 言語と語族が不明である 6 言語を含み世界中に分布していると思われる (図 1, 亀井 他 (1988, 1989, 1992), 角田 (1991)). 例えば、世界の言語の語順の分類においていわれる、

“SOV 言語は 45%, SVO 言語は 35%, VSO 言語は 18%位の割合で起こる。” (柴谷 他 (1982))

にもほぼ見合ったものとなっている。すなわち角田の語順表から得られるこれらの割合はそれぞれ、43, 36, 15%である。ここで SOV 言語とは日本語に代表される言語で、文の並びが、主語、目的語、動詞の順である言語を指し、SVO 言語とは英語に代表される言語で、主語、動詞、目的語の語順をもつ言語を指す。VSO (動詞、主語、目的語の順)をもつ言語には、アイルランド語や、フィリピンのイロカノ語などがある。

以降、2 節で語順表の数値化の方法およびその数値を用いた階層クラスタ分析の結果から、語順表の中の項目 2「名詞と側置詞」の順序によって世界の言語がほぼ 2 分されることを示す。3 節で主成分分析結果について、4 節で便宜的に言語を分類する。5 節では分類を主に規定して

表1. 語順の表 (角田 (1991) 大付録から抜粋). 語順の表の項目2~10, 12, 18および19は, 日本語と同じ語順なら+, 日本語と逆の語順なら-で示す. それ以外の項目は記号, 言葉, その他で示す. 2つ以上の語順がある場合は, 普通とみなす語順をより普通なものから順に示す.

	日本語	英語	タイ語	テイグリーニヤ語
1 S, OとV	SOVなど	SVO	SVO	SOVなど
2 名詞と側置詞	+	-	-	-
3 所有格と名詞	+	+, -	-	+
4 指示詞と名詞	+	+	-	+
5 数詞と名詞	+	+	-	+
6 形容詞と名詞	+	+	-	+
7 関係節と名詞	+	-	-	+
8 固有名詞と普通名詞	+	-, +	-	-
9 比較の表現	+	-	-	その他
10 本動詞と助動詞	+	-	-, +	+; - (?)
11 副詞と動詞	Vより前	様々	様々	Vより前
12 副詞と形容詞	+	+, -	-	+
13 疑問の印	文末	無し	文末; 焦点の後	様々
14 一般疑問文でのS, V倒置	無し	有る	無し	無し
15 疑問詞	平叙文式	文頭	平叙文式	平叙文式
16 特殊疑問文でのS, V倒置	無し	有る	無し	無し
17 否定の印	動詞語尾	Vの直後	焦点の直前	不明
18 条件節と主節	+	+, -	+, -	+
19 目的節と主節	+	-	-	-

	フィンランド語	イロカノ語	ベンガル語	中国語
1	SVOなど	VSO	SOV	SVOなど
2	+, -	-	+	-
3	+など	-	+	+
4	+	+	+	+
5	+	不明	+	+
6	+	+, 名詞を挟む(?)	+	+
7	-, +	-	その他	+
8	+, -	-	+	+
9	-, +	不明	不明	その他
10	-	-	+	-
11	様々	様々(?)	Vの直前; など	様々
12	+	-	+	+
13	2番目	無し	質問の焦点の直後(?)	文末
14	無し	不明	無し	無し
15	文頭(?)	文頭	不明	平叙文式
16	無し	不明	無し(?)	無し
17	否定助動詞が 本動詞の直前(?)	文頭	Vの後	SとVの間
18	-, +	+, -	+	+
19	+	- (?)	不明	不明

表2. 語順の19項目と例文(角田(1991)を参考にした).

1	S, OとV	私は本を買った。(主語, 目的語, 動詞の順)
2	名詞と側置詞	東京へ
3	所有格と名詞	私の家
4	指示詞と名詞	この家
5	数詞と名詞	3人の子供
6	形容詞と名詞	大きい家
7	関係節と名詞	昨日読んだ本
8	固有名詞と普通名詞	名古屋大学
9	比較の表現	(この本は) <u>あの本より</u> 難しい
10	本動詞と助動詞	食べている
11	副詞と動詞	いつも速く歩く(副詞は動詞より前)
12	副詞と形容詞	とてもきれい
13	疑問の印	読みましたか(疑問の印は文末)
14	一般疑問文でのS, V倒置	あれは富士山ですか?(倒置無し) Is that Mt. Fuji? (英語は倒置有り)
15	疑問詞	あなたは昨日どこへ行きましたか?(平叙文式) Where did you go yesterday? (英語は文頭)
16	特殊疑問文でのS, V倒置	あれは何ですか?(倒置無し) What is that? (英語は倒置有り)
17	否定の印	読まない(否定の印は動詞語尾)
18	条件節と主節	もし明子があれば, 花子は行きます.
19	目的節と主節	花子が出かけられるように, 明子は家にいた.

いると考えられる項目の選択方法を取り上げる. これらを踏まえた上で分類に関わっている項目に対する考察を6, 7節でおこなう.

言語類型論ではここで用いた項目のことをパラメータと言い, タイトル「…2パラメータモデル」のパラメータはこの意味で用いる. 階層クラスタ分析および主成分分析には Becker et al. (1988) によるソフトウェア S 言語の関数を使用した.

2. 角田の語順表の数値化の方法と階層クラスタ分析

角田の130言語19項目にわたる語順表の数値化は語順表で使用されている表現の違いを反映するように配慮した. その結果, 各言語の特徴を87次のベクトルで表現した. 原則的には次の規則に従った.

- (1) データの語順表現で情報の不確定性を示す“?”, あるいは別の語順の可能性を示す“など”の表現は数値化に際しては考慮しない.“不明”はS言語表現で“NA (Not Available)”とする.
- (2) 語順表にあらわれた表現のちがいに応じて, 各項目のベクトルの次数の長さを決定し [0, 1] の値を割り当てる. 項目内の要素の和は1とする.
- (3) 同項目内に複数の語順をもつものは, 使用頻度の高い順に表示される. 従って先にかかれた語順ほど大きい数値を割り当てる. 比率は予め定める.

角田の語順表は, 項目2~10, 12, 18および19は日本語を基準とし, 日本語と同じ語順をもつ

表3. 130言語リスト (ただし, 117番はチョクトー語, チカソー語の2言語を含む).

1 日本語 (Japanese)	65 トング語 (Tongan)
2 韓国語・朝鮮語 (Korean)	66 サモア語 (Samoan)
3 蒙古語 (Mongolian)	67 ニウエ語 (Niuean) (ニウエ島)
4 エベンキ語 (Evenki) (シベリア)	68 マオリ語 (Maori) (ニュージーランド)
5 トルコ語 (Turkish)	69 ワルング語 (Warrungu)
6 マリ語 (Mari) (ボルガ川流域)	70 カルカトゥング (Kaikatuungu)
7 ハンガリー語 (Hungarian)	71 ディヤリ語 (Diyari)
8 フィンランド語 (Finnish)	72 アリヤワラ語 (Alyawarra)
9 アブハズ語 (Abkhaz)	73 ワルピリ語 (Warlpiri)
10 アジゲ語 (Adyge)	74 ジャル語 (Djaru)
11 カバルディアン語 (Kabardian)	75 クニヤンティ語 (Kuniyanti)
12 アバル語 (Avar)	76 アムエシャ語 (Amuesha) (ペルー)
13 グルジア語 (Georgian)	77 ハカル語 (Jaqaru)
14 ロシア語 (Russian)	78 アイマラ語 (Aymara)
15 ポーランド語 (Polish)	79 グアラニ語 (Guarani) (パラグワイなど)
16 チェコ語 (Czech)	80 ウルブ・カーボル (Urubu-Kaapor) (ブラジル)
17 ブルガリア語 (Bulgarian)	81 カネラ語 (Canela) (ブラジル)
18 セルボ・クロアチア語 (Serbo-Croatian) (ユーゴスラビア)	82 ビラハ語 (Piraha) (ブラジル)
19 スウェーデン語 (Swedish)	83 ヒシカリャナ語 (Hixkaryana)
20 ノルウェー語 (Norwegian)	84 アパライ語 (Apalai)
21 デンマーク語 (Danish)	85 ケチュア語 (Quechua) (ペルーなど)
22 ドイツ語 (German)	86 トゥユカ語 (Tuyucan) (コロンビアなど)
23 オランダ語 (Dutch)	87 トル語 (Tol) (ホンジュラス)
24 英語 (English)	88 ハイランド・チョンタル語 (Highland Chontal) (メキシコ)
25 アイルランド語 (Irish)	89 ワラバイ語 (Walapai) (アメリカ)
26 ウェールズ語 (Welsh) (イギリス)	90 サウスイースタン・ポモ (Southeastern Pomo) (アメリカ)
27 ブレトン語 (Breton) (フランス)	91 イースタン・ポモ (Eastern Pomo) (アメリカ)
28 フランス語 (French)	92 マム語 (Mam)
29 ポルトガル語 (Portuguese)	93 イシル語 (Ixil)
30 スペイン語 (Spanish)	94 キチュ語 (Quiche)
31 イタリア語 (Italian)	95 ポコムチ語 (Pokonchi)
32 ルーマニア語 (Rumanian)	96 ラビナル・アチ語 (Rabinal Achi)
33 現代ギリシャ語 (Modern Greek)	97 カクチケル語 (Cakchiquel)
34 ペルシア語 (Persian)	98 ケクチ語 (K'ekchi')
35 パンジャビ語 (Panjabi) (インド)	99 ハカルテク語 (Jacaltec)
36 ヒンディー語 (Hindi) (インド)	100 トホラバル語 (Tojolabal)
37 ベンガリ語 (Bengali) (バングラデシュ)	101 チョントル・マヤ語 (Chontal Mayan)
38 バスク語 (Basque) (フランス、スペイン)	102 チョルティ語 (Chorti)
39 エジプトのアラビア語 (Egyptian Arabic)	103 コバラ・トリケ語 (Copala Trique)
40 現代ヘブライ語 (Modern Hebrew) (イスラエル)	104 イスマス・ザポテク (Isthmus Zapotec)
41 ティグリニャ語 (Tigrinya) (エチオピア)	105 ビピル語 (Pipil)
42 ハウサ語 (Hausa) (ナイジェリア)	106 ナワトル語 (Nahuatl)
43 ヨルバ語 (Yoruba) (ナイジェリア)	107 ヤキ語 (Yaqui)
44 スワヒリ語 (Swahili) (タンザニア、ケニア)	108 パパゴ語 (Papago)
45 ハヤ語 (Haya) (ケニア、ウガンダ)	109 ホピ語 (Hopi)
46 タミル語 (Tamil)	110 チェメフエヒ語 (Chemehuevi)
47 カンナダ語 (Kannada)	111 コマンチ語 (Comanche)
48 ブルシャスキ語 (Burushaski) (パキスタン)	112 ルイセニョ語 (Luiseno)
49 チベット語 (Tibetan)	113 カイオワ語 (Kiowa)
50 ミゾ語 (Mizo) (インド)	114 ナバホ語 (Navajo) (アメリカ)
51 ビルマ語 (Burmese)	115 サルシー語 (Sarcee) (カナダ)
52 北京の中国語 (Mandarin Chinese)	116 スレイビー語 (Slavey) (カナダ)
53 タイ語 (Thai)	117 チョクトー語 (Choctaw) / チカソー語 (Chickasaw)
54 ラオ語 (Lao) (ラオス)	118 ユチ語 (Yuchi) (アメリカ)
55 カンボジア語 (Cambodian)	119 オマハ・ポンカ語 (Omaha-Ponca)
56 ベトナム語 (Vietnamese)	120 グコタ語 (Dakota)
57 マレイ語 (Malay) (マレーシアなど)	121 ブラックフット語 (Blackfoot)
58 インドネシア語 (Indonesian)	122 アティカメク語 (Atikamekw)
59 タガログ語 (Tagalog) (フィリピン)	123 サハプティン語 (Sahaptin)
60 イロカノ語 (Ilokano) (フィリピン)	124 ネズ・パース語 (Nez Perce)
61 カパンパンガン語 (Kapampangan) (フィリピン)	125 コースト・ツィムシアン語 (Coast Tsimshian)
62 ビコル語 (Bikol) (フィリピン)	126 ギトックサン語 (Gitksan)
63 パラウ語 (Palauan) (パラオ島)	127 エスキモー語 (Eskimo) (カナダ、アラスカなど)
64 チャモロ語 (Chamorro) (グアム、サイパン)	128 チュクチ語 (Chukchi) (シベリア)
	129 ニブフ語 (Nivkh) (サハリン、アムール川付近)

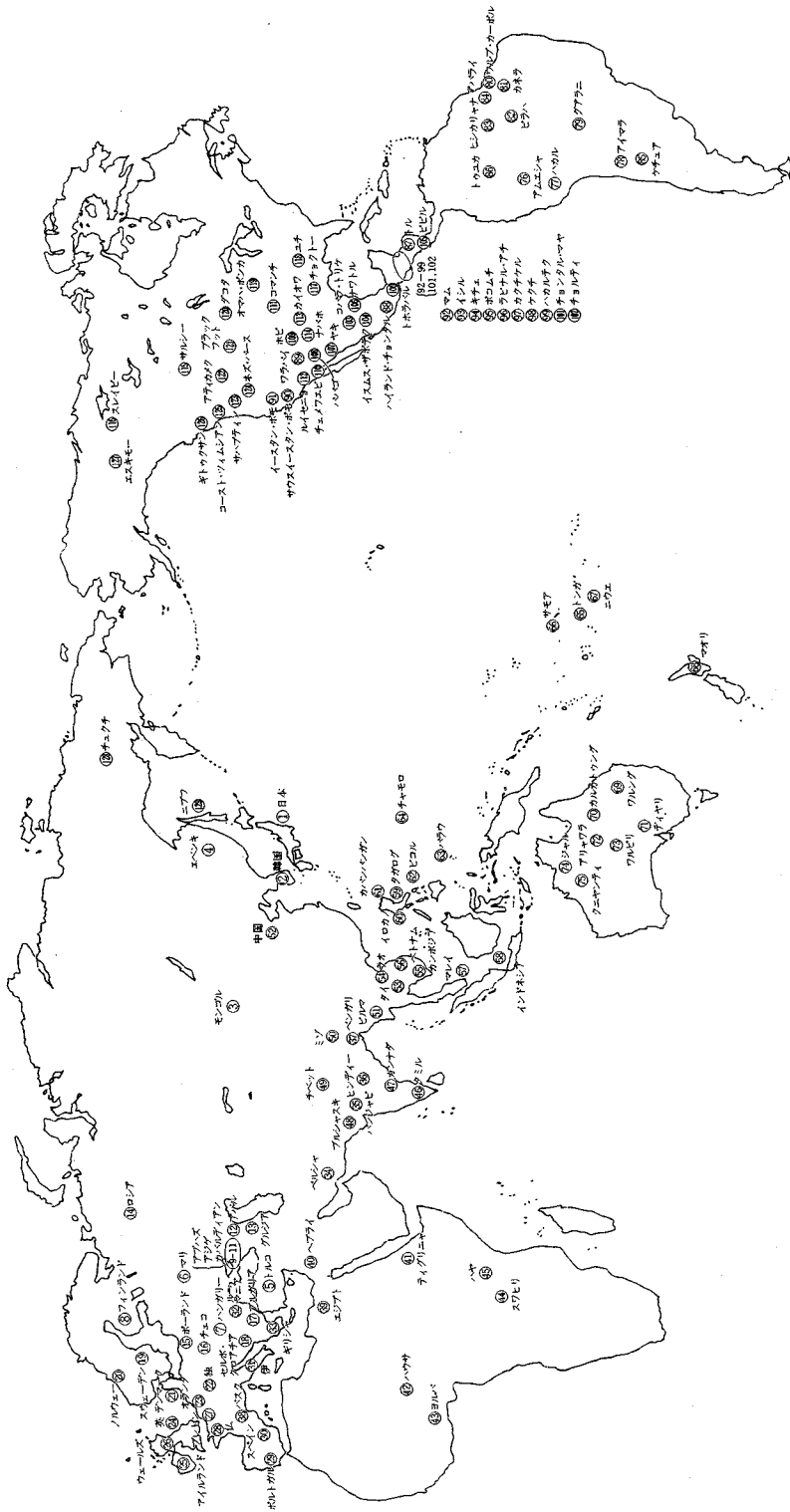


図1. 130 言語の分布の概略図 (亀井 他 (1988-1992), 角田 (1991) を参考に作成した).
番号は語順表に掲載された言語の順序である (表3).

ものに対して“+”を，日本語と反対の語順をもつものを“−”で表示し，そのどちらにも属さないものは言葉で表示している．また，その他の項目については言葉または“その他”で表示される．上記の規則に従い，例えば，項目1の「S, OとV」の語順においては，考えられる6通りの組合せ

(SOV, SVO, OSV, VSO, OVS, VOS)

から6次のベクトルで表現し，

日本語の場合：“SOV” …………… (1, 0, 0, 0, 0, 0)

英語の場合：“SVO” …………… (0, 1, 0, 0, 0, 0)

のように数値を充てた．項目2「名詞と側置詞」では，

(+, −, 側置詞無し, その他)

の4次のベクトルで表現した．表1から日本語，英語，フィンランド語についてみると，項目2「名詞と側置詞」では

日本語の場合：“+” …………… (1, 0, 0, 0)

英語の場合：“−” …………… (0, 1, 0, 0)

フィンランド語の場合：“+，−” … (0.6, 0.4, 0, 0)

のように数値を充てた．このようにして19項目に使用したベクトルの総次数は87次となり，各項目の次数は6次，4次，3次，…のように異なるが，項目内の要素の和は1である．階層クラスタ分析では，2つの言語*i*と*j*の距離はベクトルの各要素の差の絶対値の総和で測った．すなわち，

$$dist(i, j) = \sum_{k=1}^{19} \sum_{m=1}^{n(k)} |i(k, m) - j(k, m)|$$

$n(k)$ は項目 k のベクトルの次数

$i(k, m)$ は i 言語の k 項目第 m 番目の要素の値

とし，0に近い程2つの言語が類似しているとみなす．クラスタ分析には，130言語に対して19項目全部を平等に扱い，群平均法と最長距離法を用いた(図2, Tsunoda et al. (1995))．このとき複数の語順に対して優先順位をつけずに同比率，すなわち

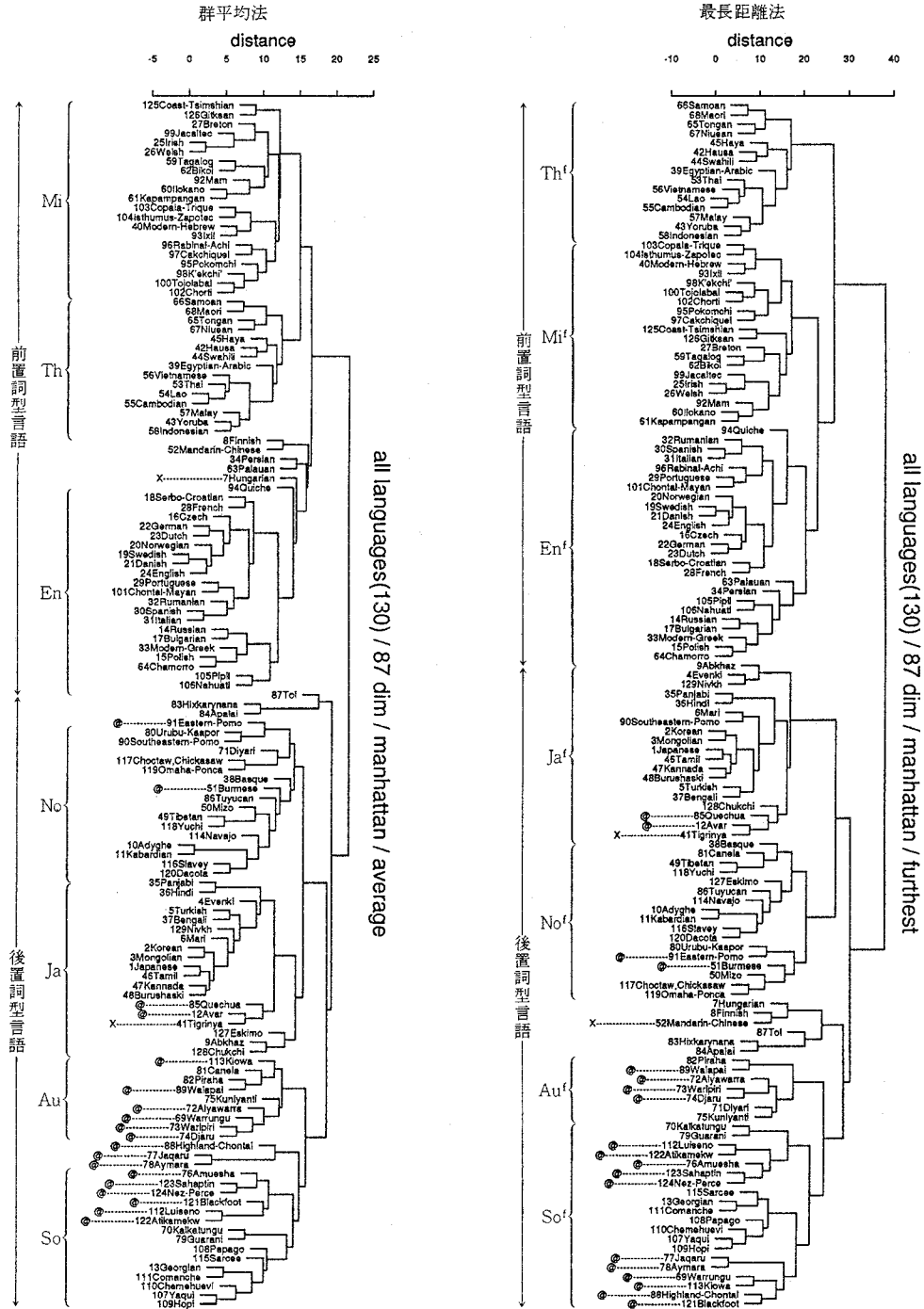
(0.6, 0.4, 0, 0)

の代わりに

(0.5, 0.5, 0, 0)

のような数値を与えた場合にも階層クラスタ分析では同様の結果を得る．つまり同比率で扱う場合も無側置詞言語を後置詞をもつ言語と同じカテゴリーに属するとすると(図2, @印が無側置詞言語)，項目2の「名詞と側置詞」の順序によってデンドログラムがほぼ2分されるのを見ることができる．これは他の18項目ではみられない現象である．デンドログラム上の2分割を仮定し，それぞれ

A：日本語と同じグループに分類され，当該項目が日本語と同じ順序である．



注) この図は *Linguistics*, 33-4 (Tsunoda et al. (1995)) に掲載されたものであり, Mouton de Gruyter Publishers, Berlin のご厚意によりここに掲載する。

図2. 階層クラスタ分析と7グループ分割。@印は無側置詞言語、×印は項目2による2分割を仮定した場合の例外を指す。

B：日本語と異なるグループに分類され、当該項目が日本語とは対称的な順序である。

のように分けるとき、A、Bの例外とみなされる言語（例外言語を図2、×印で示す。）を列挙すれば、

項目2：群平均法	{	1言語(ティグリニャ語) 1言語(ハンガリー語)	}	最長距離法	{	2言語(ティグリニャ語, 中国語) なし	}
----------	---	-----------------------------	---	-------	---	-------------------------	---

である。これに対して、項目2の次によく分類に関わっているように見える項目1「S、OとV」では、後述するOV言語かVO言語かをA、Bに対応させると

項目1：群平均法	{	8言語 1言語(ペルシャ語)	}	最長距離法	{	11言語(ハンガリー語, 中国語を含む.) 1言語(ペルシャ語)	}
----------	---	-------------------	---	-------	---	-------------------------------------	---

となり例外の多いことがわかる。以後、項目2によって2分される言語を

- a：後置詞型言語（前置詞をもたない言語）
- b：前置詞型言語（前置詞をもつ言語）

と呼ぶことにする。

数値化の仕方の詳細については、上田・伊藤（1992）で述べている。

3. 主成分分析

前節の階層クラスタ分析の結果分類された2つのタイプの言語は、同一項目内で対称的な関係にある。——データが日本語と同等の語順か否かで表現されているのでそれらが想定されていたとも考えられるのであるが。——実際、19項目のうちのいくつかは相互に強い関係があり、それらは同一項目内では互いに拮抗している（図3参照）。例えば、項目1と2、項目2と3の分割表（表4）からは、

“SOVの語順をもつ言語の多くは後置詞をもち、後置詞をもつならば所有格は名詞の前にくる。”

“SVOの語順をもつ言語の多くは前置詞をもち、前置詞をもつ言語の所有格は名詞の後に位置することが多い。”

などということが出来る。前者は日本語に近い語順をもつ言語の特徴であり、後者は日本語と逆の語順を示すものである。韓国語、ベンガル語が前者に属し、英語、タイ語は後者にあたる。

柴谷 他（1982）は、「S、OとV」の語順のSを除外してOV言語、VO言語と分類することによって、他の項目との顕著な対称関係が観察されるとして、次のような語順を掲げている。

OV言語	VO言語
2 名詞+後置詞 ……………	前置詞+名詞
10 本動詞+助動詞 ……………	助動詞+本動詞

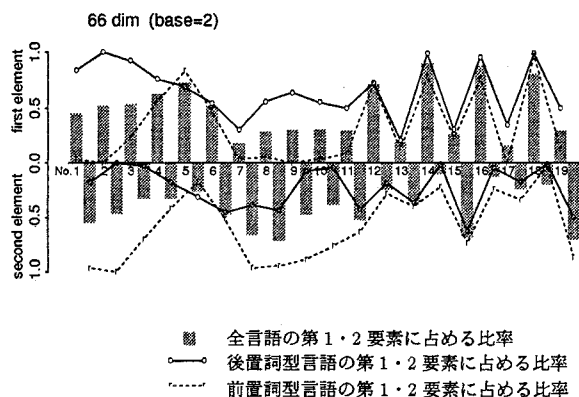


図3. 第1・2要素に占める比率.

表4. 項目2と項目1, 項目2と項目3の分割表 (87次元).

		項目1「S, OとV」						項目3「所有格と名詞」				
		SOV	SV0	OSV	VSO	OVS	VOS	+	-	名詞を挟む	その他	不明
項目2	後置詞	42	4	0	0	2	0	45	1	0	3	0
	前置詞	2	38	0	15	0	5	15	42	1	2	0
	無し	12	5	0	2	0	0	16	1	0	1	1
	その他	0	0	0	2	0	0	1	1	0	0	0

ここで項目1の"OSV"の語順をもつ言語は皆無ではない。アメリカのナバホ語は"SOV"と"OSV"の語順をもつが、優先順位の高い"SOV"をとってSOV型言語として取扱った。パパゴ語は項目1「S, OとV」に関して様々の語順をもち項目1から除いた。

- 11 副詞+動詞 動詞+副詞
- 7 関係節+名詞 名詞+関係節
- 6 形容詞+名詞 名詞+形容詞
- 4 指示詞+名詞 名詞+指示詞
- 5 数詞+名詞 名詞+数詞
- 8 固有名詞+普通名詞 ... 普通名詞+固有名詞
- 3 所有格+名詞 名詞+所有格

左端に付加した番号は角田の語順表の項目の番号に相当するものである。これらの関係の妥当性についての議論は別にして、ここで文章構造の骨格をなす項目1はその特性からSの位置に関わらず、OとVの順序で分類されることが一般的である。つまり、OV, VO言語の2次のベクトルで表現される。項目2は階層クラスタ分析の結果から

(後置詞型言語, 前置詞型言語, その他)

の3次のベクトルで表現できた。これにならい他の項目でもベクトルの次数を2ないし3にまとめて表現の単純化を図り, 可能な限り次のようにした。

- 第1要素: 日本語と同じ語順
- 第2要素: 日本語と対称的な関係にある語順
- 第3要素: その他

こうすることは日本語の語順との対比から互いに拮抗した語順の2グループに近い言語か, それとは別のグループに分類される言語かに焦点を絞るのに有効であると思われる。

項目1のベクトルは

- 第1要素: OV 言語 (SOV, OSV, OVS 言語)
- 第2要素: VO 言語 (SVO, VSO, VOS 言語)

であり, 項目2のベクトルは,

- 第1要素: 後置詞型言語 (後置詞をもつ言語と側置詞をもたない言語)
- 第2要素: 前置詞型言語 (前置詞をもつ言語)
- 第3要素: その他

と表現する。項目3も“名詞を挟む”と“その他”を統合して

- 第1要素: “所有格+名詞”の順序をもつ
- 第2要素: “名詞+所有格”の順序をもつ
- 第3要素: “名詞を挟む”, または“その他”

のように3次のベクトルで表した。

表5. 項目2と項目1, 項目3, 項目9の分割表 (66次元).

	項目1 「S, OとV」		項目3 「所有格と名詞」				項目9 「比較の表現」			
	OV	VO	+	-	その他	不明	+	-	その他	不明
項目 後置詞・無	56	11	61	2	4	1	34	4	16	14
項目 前置詞	2	58	14	42	4	0	0	48	6	6
項目 2 その他	0	2	1	1	0	0	0	0	0	2

同じ項目で2つ以上の語順がある場合には優先順位の最も高い語順を1として取扱った。

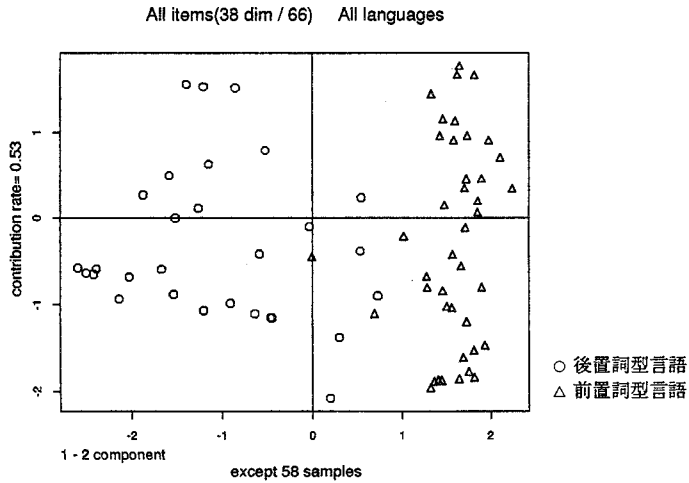


図4-1. 主成分分析 (38次元).

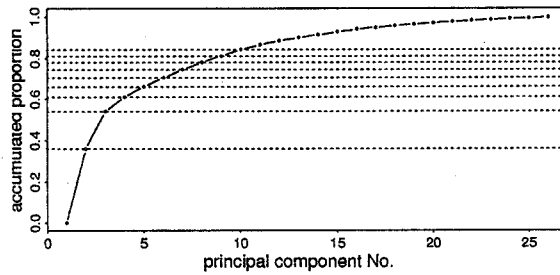


図4-2. 累積寄与率.

項目 11, 13, 17 については表現が多様で3次以下にまとめることはできず、それぞれ6, 5, 13 次の長さのベクトルになった。こうした数値化の見直しの結果、各言語の特徴を表した 19 項目は 66 次元に減少した (上田 他 (1995), p.5)。項目 1 と 2, 2 と 3, 2 と 9 の分割表を表 5 に示す。各項目の第 1・2 要素の占める比率は図 3 に示す。全 130 言語が各項目の第 1・2 要素に占める比率をヒストグラムで、後置詞型言語 (無側置詞言語を含む) の第 1・2 要素に占める比率を実線で、前置詞型言語の占める比率を点線で図示する。実線や点線がヒストグラムから離れているほど後置詞・前置詞型言語の特徴が顕著となる。

19 項目の第 1・2 要素だけをデータとしたとき、すなわち 66 次元のうちの 38 次元を用いたときの主成分分析の結果を図 4 に示す。図 4-1 の各言語の第 1, 2 主成分では、後置詞型・前置詞型言語を○, △で表示した。図 4-2 は累積寄与率, 図 4-3 は因子負荷量のペアを図示したものである。主成分分析では語順が“不明”の項目を1つでももつ言語は除外されている。

図 4-3 で、各項目の第 1 要素を大文字で、第 2 要素を小文字で表す。(例: A-OV 言語, a-VO 言語, B-後置詞型言語, b-前置詞型言語, C-“所有格+名詞”の順序, c-“名詞+所有格”の順序, 等々である。) 第 1, 2 主成分に対する因子負荷量は原点を中心に放射状に並び、各第 1・2 要素はほぼ原点对称の位置にある。主成分の寄与率は第 3 主成分以降急速に減少する。

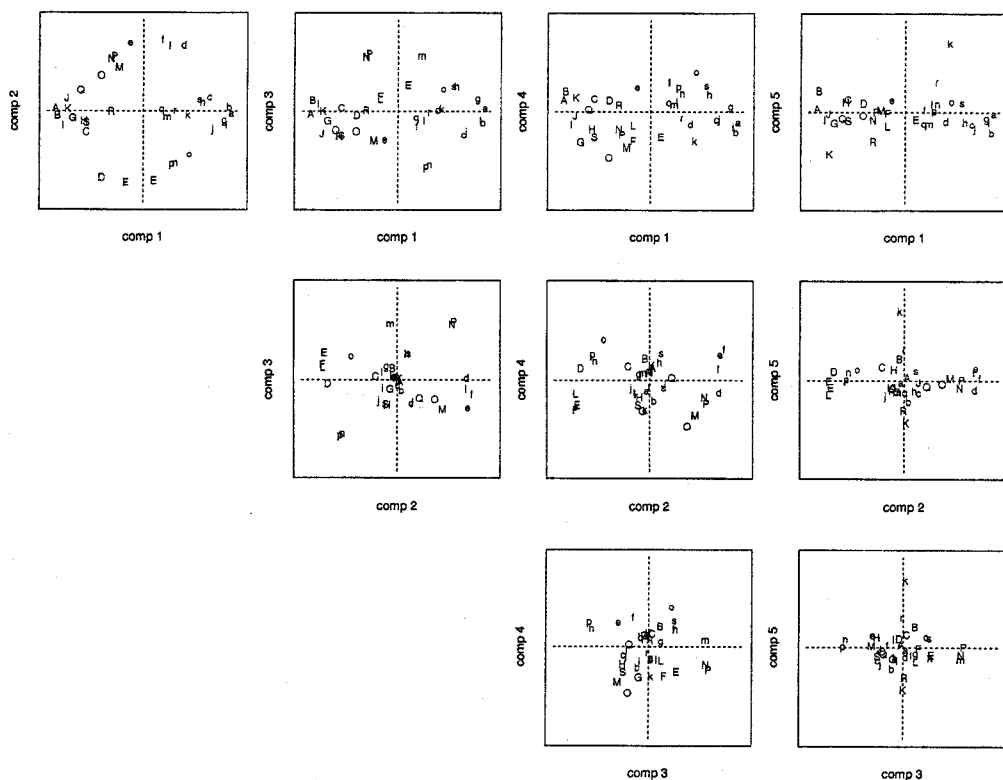


図4-3. 因子負荷量.

第1主成分を形成しているのが「S, O と V」や「名詞と側置詞」に代表される項目群であり、言語をほぼ2分する成分である。

19項目の因子負荷量の絶対値が0.5以上の項目は次のようになっている。()は $0.5 \leq$ 因子負荷量 < 0.6 の値をとる項目である。

- 第1主成分 … 1, 2, 3, 7, 8, 9, 10, 11, 17, (19)
- 第2主成分 … 4, 5, 6, 12, (14), (16)
- 第3主成分 … (13), (14), 16
- 第4主成分 … (15)
- 第5主成分 … 11

これらはおよそ次のことを示唆している。

第1主成分を構成しているのは、英語とタイ語に共通で日本語と対称的な関係にある項目

第2主成分を構成しているのは、日本語と英語に共通でタイ語と対称的な関係にある項目

第3主成分を構成しているのは、日本語とタイ語に共通で英語と対称的な関係にある項目

そして日本語、英語、タイ語の3言語は世界の言語の語順の3極端に位置する。なぜなら、3言語は第1, 2主成分をみる限りほぼ絶対値最大のところに位置し、主成分分析をアジア・ヨー

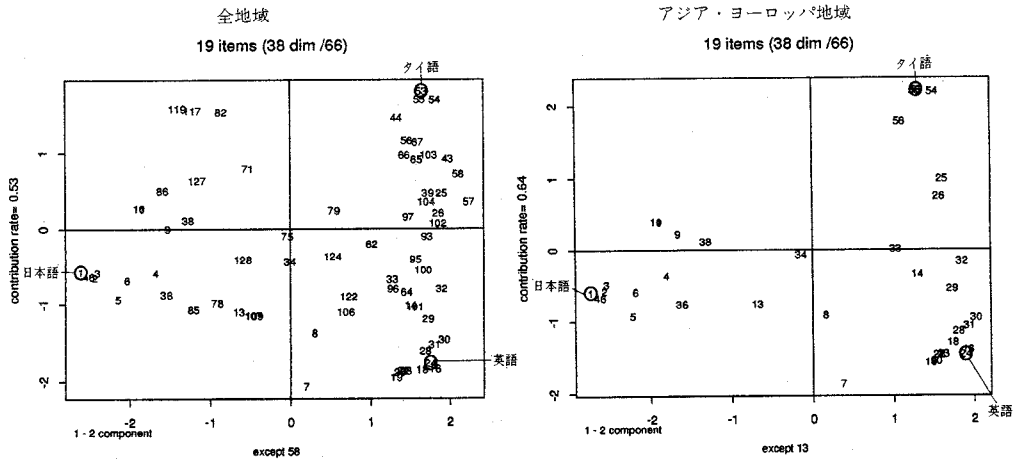


図5. 主成分分析 (38次元). 66次元データ, 87次元データの場合も同様の結果となる. ここでは66次元データの第1・2要素を抜き出した38次元の場合を取り上げる. ただし, アジア・ヨーロッパ地域の第2主成分の値の符号を変えた.

ロップに限定しても, 全体を対象にしても, 個々の言語のスコアにはほとんど変化がなく, 世界の言語はこの3言語の間隙を埋めるかのように分布するからである (図5).

主成分分析の記述には, 河口 (1973) を参考にした.

4. 7グループ分割

語順規則による分類において各グループを区別する主な項目群を取り出すために, まず便宜的なグループを生成する. クラスタ間の距離とその時点で生成されるグループ数との関係を見ると, 距離 14.53 で14グループが生成され各グループに属する言語数は,

21, 15, 2, 1, 1, 1, 22, 1, 2, 18, 19, 9, 3, 15

である (図2, 左図). グループ内に属する言語に共通した特徴をとらえるにはごく少数の言語で構成されるグループを除いた方がよいと考えられる. 3以下の言語で構成されるグループを除くと14グループは7グループとなり言語数は

21, 15, 22, 18, 19, 9, 15

である. この各々のグループを次のように名づける. 命名はグループに所属する言語または地域からとったが, 必ずしも隣接した地域の言語ばかりでなく, 例えば, Jaのグループにはエチオピアのティグリニャ語やペルーのケチュア語が含まれる.

後置詞型言語	1. No	18 言語	(北米地域の言語を含むグループ)
	2. Ja	19 言語	(日本語を含むグループ)
	3. Au	9 言語	(豪州地域の言語を含むグループ)
	4. So	15 言語	(南米地域の言語を含むグループ)
前置詞型言語	5. Mi	21 言語	(中米地域の言語を含むグループ)
	6. Th	15 言語	(タイ語を含むグループ)
	7. En	22 言語	(英語を含むグループ)

除外した言語：11 言語

ハンガリー語、フィンランド語、中国語、ペルシャ語、パラウ語、トル語、ヒシカリャナ語、アパライ語、ハイランド・チョンタル語、ハカル語、アイマラ語

また“No”に属するチョクトー語、チカソー語は19項目において同語順を示し分析では1言語として取り扱っている。ここで生成した7グループは、群平均法によるデンドログラムによって生成されたグループの中から次節で述べる手法を用いてAICの値を比較して選び出したものである(表6)。この7グループ分割にもとづいて以降の議論を進めていきたい。

5. 分割表データにおける項目の選択

言語の分類に必要な最適な項目を選択すること、また逆に選択された項目を用いて最適なグループ化を図る。これらには分割表データにおける変数選択の方法(Sakamoto and Akaike (1978), 坂元 (1985))を用いた。以下で述べる解析方法の表記はそれに従う。

7グループに分割する場合の説明変数としてベクトルの次数が4以上の11, 13, 17を除外し

表6. 群平均法によるデンドログラムの距離、グループ数、AIC.

距離	グループ数-言語数 (AIC)	k以下の言語で構成されるグループを除いた場合	
		k=2の場合 グループ数-言語数(AIC)	k=3の場合 グループ数-言語数(AIC)
14.29	15-129 (-124.02)	9-121 (-143.05)	7-115 (-151.88)
14.53	14-129 (-125.67)	8-121 (-150.76)	* 7-118 (-153.92)
14.58	13-129 (-127.67)	〃	〃
14.61	12-129 (-124.47)	8-122 (-147.77)	7-119 (-151.04)
14.98	11-129 (-126.04)	7-122 (-138.46)	6-119 (-141.73)
15.05	10-129 (-128.04)	6-122 (-148.16)	— (左に同じ)
15.54	9-129 (-128.73)	5-122 (-132.16)	—
15.90	8-129 (-130.35)	5-124 (-135.74)	—
15.94	7-129 (-131.46)	4-124 (-133.15)	—
16.13	6-129 (-132.81)	4-126 (-132.88)	—
16.53	5-129 (-131.88)	3-126 (-131.96)	—

ここで、言語数とは解析に用いた全言語の数である。

た16個の項目について考え、その中からグループ分割に有効な項目の組合せを選択する。今、目的変数 i_0 のカテゴリ数 c_0 は7, 説明変数 i_j ($j = 1, \dots, 16$) のカテゴリ数 c_j ($j = 1, \dots, 16$) は2ないし3とする。この17次元の各セルに対応する確率を $p(i_0, i_1, i_2, \dots, i_{16})$, 観測度数を $n(i_0, i_1, i_2, \dots, i_{16})$ と表すことにする。

$$\sum_{i_0, \dots, i_{16}} p(i_0, i_1, \dots, i_{16}) = 1$$

$$\sum_{i_0, \dots, i_{16}} n(i_0, i_1, \dots, i_{16}) = n$$

である。 n は全データ数, $\sum_{i_0, \dots, i_{16}} = \sum_{i_0=1}^{c_0} \sum_{i_1=1}^{c_1} \dots \sum_{i_{16}=1}^{c_{16}}$ と記す。このとき、観測度数の集合 $\{n(i_0, i_1, \dots, i_{16})\}$ を得る確率は、多項分布

$$P(\{n(i_0, i_1, \dots, i_{16})\} | \{p(i_0, i_1, \dots, i_{16})\}) = \frac{n!}{\prod_{i_0, \dots, i_{16}} n(i_0, i_1, \dots, i_{16})!} \times \prod_{i_0, \dots, i_{16}} p(i_0, i_1, \dots, i_{16})^{n(i_0, i_1, \dots, i_{16})},$$

$$\prod_{i_0, \dots, i_{16}} = \prod_{i_0=1}^{c_0} \prod_{i_1=1}^{c_1} \dots \prod_{i_{16}=1}^{c_{16}}, \quad c_0 = 7, c_1 = \dots = c_{16} = 2 \text{ ないし } 3$$

となる。ここで、 $p(i_0, i_1, \dots, i_{16})$ をパラメータとみなして、 $p(i_0, i_1, \dots, i_{16})$ に無関係な定数項を無視すると、対数尤度は

$$(*) \quad l(\{p(i_0, i_1, \dots, i_{16})\}) = \sum_{i_0, \dots, i_{16}} n(i_0, i_1, \dots, i_{16}) \log p(i_0, i_1, \dots, i_{16})$$

与えられる。

$$p(i_0, i_1, \dots, i_{16}) = p(i_0 | i_1, \dots, i_{16}) p(i_1, \dots, i_{16})$$

を(*)に代入して、

$$\begin{aligned} l(\{p(i_0, i_1, \dots, i_{16})\}) &= \sum_{i_0, \dots, i_{16}} n(i_0, i_1, \dots, i_{16}) \log \{p(i_0 | i_1, \dots, i_{16}) p(i_1, \dots, i_{16})\} \\ &= \sum_{i_0, \dots, i_{16}} n(i_0, i_1, \dots, i_{16}) \log p(i_0 | i_1, \dots, i_{16}) \\ &\quad + \sum_{i_0, \dots, i_{16}} n(i_0, i_1, \dots, i_{16}) \log p(i_1, \dots, i_{16}) \\ &= \sum_{i_0, \dots, i_{16}} n(i_0, i_1, \dots, i_{16}) \log p(i_0 | i_1, \dots, i_{16}) \\ &\quad + \sum_{i_1, \dots, i_{16}} \left\{ \sum_{i_0=1}^{c_0} n(i_0, i_1, \dots, i_{16}) \right\} \log p(i_1, \dots, i_{16}) \\ &= \sum_{i_0, \dots, i_{16}} n(i_0, i_1, \dots, i_{16}) \log p(i_0 | i_1, \dots, i_{16}) \\ &\quad + \sum_{i_1, \dots, i_{16}} n(i_1, \dots, i_{16}) \log p(i_1, \dots, i_{16}) \end{aligned}$$

となる。 i_0 に無関係な項を無視して条件付き対数尤度を考えれば

$$l(\{p(i_0 | i_1, \dots, i_{16})\}) = \sum_{i_0, \dots, i_{16}} n(i_0, i_1, \dots, i_{16}) \log p(i_0 | i_1, \dots, i_{16})$$

となり、これから与えられたモデルに対応した最大対数尤度を求めれば、AIC は次のように定義される。

$$\text{AIC} = -2 \times (\text{モデルの最大対数尤度}) + 2 \times (\text{モデルの自由パラメータ数}).$$

この AIC を最小とするモデルが最適なモデルと考えられる。

7 グループ分割のためには 16 項目の説明変数の中のどの項目を用いれば有効であるかを判定する。すなわち最適なモデルを選択する。主成分分析から第 5 主成分以降 16 項目との因子負荷量が 0.5 以上の項目はなく、第 4 主成分までを代表する 4 項目について考えることにした。モデルは次のものについて考える。

$$(5.1) \quad \text{MODEL}(I_0; I_1, I_2, I_3, I_4) : p(i_0 | i_1, \dots, i_{16}) = a(i_0 | i_1, i_2, i_3, i_4)$$

$$(5.2) \quad \text{MODEL}(I_0; I_1, I_2, I_3) : p(i_0 | i_1, \dots, i_{16}) = a(i_0 | i_1, i_2, i_3)$$

$$(5.3) \quad \text{MODEL}(I_0; I_1, I_2) : p(i_0 | i_1, \dots, i_{16}) = a(i_0 | i_1, i_2)$$

$$(5.4) \quad \text{MODEL}(I_0; I_1) : p(i_0 | i_1, \dots, i_{16}) = a(i_0 | i_1).$$

ここで i_1, i_2, i_3, i_4 は 16 項目の説明変数 i_1, \dots, i_{16} の重複しないいずれかの変数を指す。(5.1) のモデルは確率 $p(i_0 | i_1, \dots, i_{16})$ が 4 変数の値に依存して他の 12 変数の値の如何にかかわらず定まると考える。同様に (5.2) のモデルは 3 変数、(5.3) は 2 変数、(5.4) は 1 変数のみに依存して定まると考えるモデルである。そのときの AIC はそれぞれ次のように計算される。ここで、各式に共通に $(-2) \sum n(i_0) \log(n/n(i_0)) - 2(c_0 - 1)$ を加えた。

$$\begin{aligned} \text{AIC}(i_0; i_1, i_2, i_3, i_4) &= (-2) \sum_{i_0, i_1, i_2, i_3, i_4} n(i_0, i_1, i_2, i_3, i_4) \log \frac{n \cdot n(i_0, i_1, i_2, i_3, i_4)}{n(i_0)n(i_1, i_2, i_3, i_4)} \\ &\quad + 2(c_0 - 1)(c_1 c_2 c_3 c_4 - 1), \end{aligned}$$

$$\begin{aligned} \text{AIC}(i_0; i_1, i_2, i_3) &= (-2) \sum_{i_0, i_1, i_2, i_3} n(i_0, i_1, i_2, i_3) \log \frac{n \cdot n(i_0, i_1, i_2, i_3)}{n(i_0)n(i_1, i_2, i_3)} \\ &\quad + 2(c_0 - 1)(c_1 c_2 c_3 - 1), \end{aligned}$$

$$\begin{aligned} \text{AIC}(i_0; i_1, i_2) &= (-2) \sum_{i_0, i_1, i_2} n(i_0, i_1, i_2) \log \frac{n \cdot n(i_0, i_1, i_2)}{n(i_0)n(i_1, i_2)} \\ &\quad + 2(c_0 - 1)(c_1 c_2 - 1), \end{aligned}$$

$$\text{AIC}(i_0; i_1) = (-2) \sum_{i_0, i_1} n(i_0, i_1) \log \frac{n \cdot n(i_0, i_1)}{n(i_0)n(i_1)} + 2(c_0 - 1)(c_1 - 1),$$

$$c_0 = 7, c_1 = c_2 = c_3 = c_4 = 2 \quad \text{ないし} \quad 3.$$

次にグループ化について考える。前述の説明変数と目的変数を逆転させて、グループをいくつに分割すれば項目 2 の各要素の分布がより有効に説明できるかというように考えることによって、グループ化の問題はカテゴリー-区分法の選択の問題とみなすことができる。例えば、4 グループ分割の場合と 2 つを統合して 2 グループ分割にした場合を考える (表 7)。4 グループの場合 i_0 のカテゴリー数は 4、項目 2 の i_2 のカテゴリー数は 3 で、そのときのモデルは

$$\text{MODEL}(I_2; I_0) : p(i_2 | i_0) = a(i_2 | i_0), \quad i_2 = 1, 2, 3, \quad i_0 = 1, 2, 3, 4$$

で表される。2 グループの場合のモデルは i_0' のカテゴリー数は 2 となって

$$\begin{aligned} \text{MODEL}(I_2; I_0') : p(i_2 | i_0 = 1) &= p(i_2 | i_0 = 2) = a(i_2 | 1), \\ &: p(i_2 | i_0 = 3) = p(i_2 | i_0 = 4) = a(i_2 | 2), \quad i_2 = 1, 2, 3 \end{aligned}$$

と表現される。AIC は次のように求められる。

表7. グループの統合.

4 グループの場合の分割表				2 グループの場合の分割表			
グループ \ 項目-要素 (i ₀)	2-1	2-2	2-3	グループ \ 項目-要素 (i _{0'})	2-1	2-2	2-3
(i ₂)				(i ₂)			
グループ 1 : No. Au	26	0	0	グループ 1 : No. Au. Ja. So	59	1	0
グループ 2 : Ja. So	33	1	0	グループ 2 : Th. Mi. En	2	55	1
グループ 3 : Th	0	15	0				
グループ 4 : Mi. En	2	40	1				

$$AIC(I_2; I_0) = (-2) \sum_{i_0, i_2} n(i_2, i_0) \log \frac{n(i_2, i_0)}{n(\cdot, i_0)} + 2c_0(c_2 - 1),$$

$$AIC(I_2; I_{0'}) = (-2) \sum_{i_2=1}^3 \left\{ \sum_{k=1}^2 n(i_2, k) \log \frac{\sum_{k=1}^2 n(i_2, k)}{\sum_{k=1}^2 n(\cdot, k)} + \sum_{k=3}^4 n(i_2, k) \log \frac{\sum_{k=3}^4 n(i_2, k)}{\sum_{k=3}^4 n(\cdot, k)} \right\} + 2(c_0 - 2)(c_2 - 1),$$

$$i_2 = 1, 2, 3, \quad i_0 = 1, 2, 3, 4, \quad c_2 = 3, \quad c_0 = 4$$

ここで、 $n(i_2, i_0)$, $n(\cdot, i_0)$ は4グループの場合の観測度数である。後者の式は2グループの場合の分割表をつくれれば前者の形に表記できる。すなわち

$$AIC(I_2; I_{0'}) = (-2) \sum_{i_0', i_2} n(i_2, i_0') \log \frac{n(i_2, i_0')}{n(\cdot, i_0')} + 2c_{0'}(c_2 - 1),$$

$$i_2 = 1, 2, 3, \quad i_0' = 1, 2, \quad c_2 = 3, \quad c_{0'} = 2$$

となる。

6. 分類に関与する2パラメータ

言語を語順規則によって7グループに分割したとき、どの項目が分類に最も関わっているかについて考察する。主成分分析の第4主成分までを考慮に入れ、因子負荷量0.7以上を1つの目安とする。同主成分内で因子負荷量が0.7以上の項目は同時に選択されることはない。すなわち次の各組から1項目を選択して4項目の組合せを考えることにする。

- (1) 第1主成分との因子負荷量が0.5以上の項目 (1, 2, 3, 7, 8, 9, 10, 19)
- (2) 第2主成分との因子負荷量が0.5以上の項目 (4, 5, 6, 12, 14, 16)
- (3) 第3主成分との因子負荷量が0.5以上の項目と、上記 (1), (2) 以外の項目および (1), (2) の項目のうち因子負荷量が0.5以上0.7未満の項目 (3, 8, 14, 15, 16, 18, 19)
- (4) (3) と同じ (3, 8, 14, 15, 16, 18, 19)

これらの各組より1項目ずつ選び出して、重複しない4つの項目の組合せからAIC最小のものをみつける。

ここでは個々の言語が各項目に共通に次の4つのいずれに分類されるかをみて、3節で述べたように1, 2に注目し3と4を同じカテゴリーに入れてカテゴリー数を3に固定する。

1. 日本語と同語順

2. 日本語と対称的な関係にある語順
3. 上記1, 2以外の語順
4. 語順が“不明”

ただし、項目1は3, 4に該当する言語はなくカテゴリー数は2とする。項目2では4の“不明”の言語はなく、3に分類される“その他”の語順をもつのは2言語（コースト・ツィムシアン語、ギトックサン語）だけでいずれも

項目2「名詞と側置詞」：“その他、—” …………… (0, 0.4, 0.6)

のように数値が充てられるので、変数選択で選ぶまでもなく“その他”は2のカテゴリーに組み入れカテゴリー数は2とする（表5参照）。すなわち各項目のカテゴリー数は $c_1 = c_2 = 2, c_3 = \dots = c_{16} = 3$ に固定した。結果は

項目2, 5, 15, 16

の組合せが最小となった。ここで、

項目2「名詞と側置詞」の順序,
項目5「数詞と名詞」の順序,
項目15「疑問詞」の位置,
項目16「疑問文でのS, V倒置」の有無,

である。これは分類に関わる項目数を4に限定した場合であるが、実際には3個の説明変数（項目2, 5, 15）、または4個の説明変数（項目2, 5, 15, 16）より2個（項目2, 5）のモデルのAICが小さい（表8）。つまり7グループ分割でカテゴリー数を固定した場合“「名詞と側置詞」の順序”と“「数詞と名詞」の順序”の2項目（2パラメータ）が選択されるという結果になった。

項目の選択（4項目～1項目）とグループ化（7グループ分割～2グループ分割）の両面に対しておこなった場合のAICの小さい順に表示したのが表9-1である。2グループ分割では1個の説明変数をとるモデルが選択されるが、3以上の分割ではいずれも2個の説明変数をとるモデルが選択された。AIC最小となった4グループ分割2項目選択

“(No. Au)-(Ja. So)-Th-(Mi. En)”なる4グループ構成

“「名詞と側置詞」の順序と「数詞と名詞」の順序”の2項目（2パラメータ）

の分割表は表9-2に示す。

唯一の項目でAICの値を比較したのが表9-3である。2グループ分割から7グループ分割まで項目2「名詞と側置詞」が選択された。項目1「S, OとV」、項目9「比較の表現」がこれに続く。

上述したように表9においては各項目のカテゴリー数を固定した。これは日本語に近い言語に分類されるのか、あるいは日本語と対称的な語順をもつ言語として分類されるのかに焦点を当てたためである。参考までに表10-1（個々の項目のカテゴリー数を変数選択を用いて決めた場合）、表10-2（“その他”、“不明”に属する言語を取り去った場合）を掲げた。これらの結果に“その他”、“不明”に固有の片寄りがある程度入っているかはわからないが、項目2, 5, 15の

表8. 7グループ分割に関与する最適な変数 (1項目~4項目を選択した場合).

説明変数	A I C
項目 2	-127.15
項目 2, 5	-153.92
項目 2, 5, 1 5	- 72.28
項目 2, 5, 1 5, 1 6	329.92

表9-1. AICの小さい順の群平均法によるグループ構成 ($c_2 = 2, c_5 = c_{15} = c_{16} = 3$).

グループ数：構成 (118言語)	AIC(選ばれた項目)	AIC (3項目選択)
4 : (No. Au) - (Ja. So) - Th - (Mi. En)	-175.71(2・5)	-132.48(2・5・15)
4 : (No. Ja) - (Au. So) - Th - (Mi. En)	-171.49(2・15)	-122.77(2・5・15)
5 : (No. Au) - (Ja. So) - Th - Mi - En	-170.53(2・5)	-106.16(2・5・16)
5 : No - Au - (Ja. So) - Th - (Mi. En)	-167.76(2・5)	-105.31(2・5・15)
5 : (No. Au) - Ja - So - Th - (Mi. En)	-167.05(2・5)	-128.13(2・5・15)
5 : (No. Au) - Ja - So - Th - (Mi. En)	-166.74(2・15)	-128.13(2・5・15)

表9-2. AIC最小の分割表.

項目 2	項目 5	4 グループ (118言語)			
		No. Au	Ja. So	Th	Mi. En
後置詞・無し	+	4.56	30.6	0	1.6
	-	19.04	2.4	0	0.0
	名詞を挟む・不明	2.00	0.0	0	0.0
前置詞	+	0.24	1.0	5	38.4
	-	0.16	0.0	10	0.8
	名詞を挟む・不明	0.00	0.0	0	2.2

表9-3. 1項目のみでグループ化した場合のAICの値 (群平均法, $c_1 = c_2 = 2, c_9 = 3$).

グループ数：構成 (118言語)	AIC(項目 2)	AIC(項目 1)	AIC(項目 9)
2 : (No. Ja. Au. So) - (Th. Mi. En)	-133.60	- 99.92	- 92.33
3 : (No. Au) - (Ja. So) - (Th. Mi. En)	-131.73	-103.58	- 93.36
4 : (No. Au) - (Ja. So) - Th - (Mi. En)	-130.71	-102.50	- 92.47
5 : (No. Au) - Ja - So - Th - (Mi. En)	-129.89	-111.29	-102.10
6 : No - Ja - Au - So - Th - (Mi. En)	-128.75	-110.59	-100.31
7 : No - Ja - Au - So - Th - Mi - En	-127.15	-110.65	- 99.76

表10-1. AICの小さい順の群平均法によるグループ構成 (変数選択によってカテゴリー区分を決めた場合, $c_2 = c_5 = c_{15} = 2$).

グループ数：構成 (118言語)	AIC(選ばれた項目)	カテゴリー区分	
		項目5	項目15
5 : (No. Au) - Ja - So - Th - (Mi. En)	-203.15(2・5・15)	ⓑ	ⓐ
6 : No - Ja - Au - So - Th - (Mi. En)	-195.00(2・5・15)	ⓑ	ⓐ
6 : (No. Au) - Ja - So - Th - Mi - En	-192.66(2・5・15)	ⓑ	ⓐ
4 : (No. Au) - (Ja. So) - Th - (Mi. En)	-187.64(2・5・15)	ⓑ	ⓐ
6 : No - Ja - (Au. So) - Th - Mi - En	-184.72(2・5・15)	ⓑ	ⓐ
7 : No - Ja - Au - So - Th - Mi - En	-184.51(2・5・15)	ⓑ	ⓐ

- ⓐ { カテゴリー1. 日本語と同語順+その他+不明
 カテゴリー2. 日本語と対称的な語順
 ⓑ { カテゴリー1. 日本語と同語順+その他
 カテゴリー2. 日本語と対称的な語順+不明

表10-2. AICの小さい順の群平均法によるグループ構成 (その他, 不明の言語を除いた場合, $c_2 = c_5 = c_{15} = 2$).

グループ数：構成 (95言語)	AIC(選ばれた項目)
5 : (No. Au) - Ja - So - Th - (Mi. En)	-149.39(2・5・15)
4 : (No. Au) - (Ja. So) - Th - (Mi. En)	-141.14(2・5)
5 : (No. Au) - Ja - So - Th - (Mi. En)	-139.52(2・15)
6 : No - Ja - Au - So - Th - (Mi. En)	-139.43(2・5・15)
4 : (No. Ja) - (Au. So) - Th - (Mi. En)	-138.92(2・15)
4 : (No. Au) - (Ja. So) - Th - (Mi. En)	-137.64(2・5・15)

- カテゴリー1. 日本語と同語順
 カテゴリー2. 日本語と対称的な語順

中から選択されていることは共通している。

同様の方法を最長距離法のデンドログラムから生成した7グループに適用した。このときの7グループを群平均法と対応させて

$$Mi-Mi^f, Th-Th^f, En-En^f, No-No^f, Ja-Ja^f, Au-Au^f, So-So^f$$

と名づける。それぞれに属する言語数は

$$20, 15, 25, 18, 19, 7, 20$$

であり、同じ条件で除いた6言語 (ハンガリー語, フィンランド語, 中国語, トル語, ヒンカリャナ語, アパライ語) は群平均法と共通である (図2, 右図)。対応したグループを構成している言語は Th^f を除いて群平均法と同一ではない。一例を上げれば最長距離法の En^f は次の言語を含む。

$$En^f(25 \text{ 言語}) = \begin{cases} \text{“En” に属する全 22 言語} \\ \text{“Mi” に属する 1 言語} \\ \text{群平均法では孤立している 2 言語(ペルシャ語, パラウ語).} \end{cases}$$

群平均法による解析結果を示した表 9 には最長距離法では表 11 が対応し、表 10 には表 12 が対応する。

表11-1. AIC の小さい順の最長距離法によるグループ構成 ($c_2 = 2, c_5 = c_{15} = c_{16} = 3$).

グループ数：構成 (123言語)	AIC(選ばれた項目)	AIC (3 項目選択)
4 : (No ^f . Au ^f) - (Ja ^f . So ^f) - Th ^f - (Mi ^f . En ^f)	-185.24 (2 · 5)	-135.12 (2 · 5 · 15)
5 : (No ^f . Au ^f) - (Ja ^f . So ^f) - Th ^f - Mi ^f - En ^f	-182.08 (2 · 5)	-110.22 (2 · 5 · 15)
5 : (No ^f . Au ^f) - Ja ^f - So ^f - Th ^f - (Mi ^f . En ^f)	-178.25 (2 · 5)	-137.37 (2 · 5 · 16)
6 : (No ^f . Au ^f) - Ja ^f - So ^f - Th ^f - Mi ^f - En ^f	-175.09 (2 · 5)	-112.47 (2 · 5 · 15)
4 : (No ^f . Ja ^f) - (Au ^f . So ^f) - Th ^f - (Mi ^f . En ^f)	-174.99 (2 · 15)	-132.86 (2 · 5 · 15)
4 : (No ^f . Au ^f) - (Ja ^f . So ^f) - En ^f - (Mi ^f . Th ^f)	-173.56 (2 · 5)	-110.69 (2 · 5 · 15)

表11-2. AIC 最小の分割表.

項目 2	項目 5	4 グループ (123言語)			
		No ^f . Au ^f	Ja ^f . So ^f	Th ^f	Mi ^f . En ^f
後置詞・無し	+	3.56	34.6	0	1.6
	-	19.04	2.4	0	0.0
	名詞を挟む・不明	1.00	1.0	0	0.0
前置詞	+	0.24	1.0	5	40.4
	-	0.16	0.0	10	0.8
	名詞を挟む・不明	0.00	0.0	0	2.2

表11-3. 1 項目のみでグループ化した場合の AIC の値 (最長距離法, $c_1 = c_2 = 2, c_9 = 3$).

グループ数：構成 (123言語)	AIC(項目 2)	AIC(項目 1)	AIC(項目 9)
2 : (No ^f . Ja ^f . Au ^f . So ^f) - (Th ^f . Mi ^f . En ^f)	-140.24	- 95.63	- 95.54
3 : (No ^f . Au ^f) - (Ja ^f . So ^f) - (Th ^f . Mi ^f . En ^f)	-138.30	- 98.96	- 93.68
4 : (No ^f . Au ^f) - (Ja ^f . So ^f) - Th ^f - (Mi ^f . En ^f)	-137.23	- 98.43	- 92.60
5 : (No ^f . Au ^f) - Ja ^f - So ^f - Th ^f - (Mi ^f . En ^f)	-136.70	-106.09	-105.61
6 : No ^f - Ja ^f - Au ^f - So ^f - Th ^f - (Mi ^f . En ^f)	-136.62	-107.15	-104.70
7 : No ^f - Ja ^f - Au ^f - So ^f - Th ^f - Mi ^f - En ^f	-135.63	-106.67	-105.45

表12-1. AICの小さい順の最長距離法によるグループ構成(変数選択によってカテゴリー区分を決めた場合, $c_2 = c_5 = c_{15} = 2$).

グループ数: 構成 (123言語)	AIC(選ばれた項目)	カテゴリー区分	
		項目5	項目15
5 : (No ^f . Au ^f) - Ja ^f - So ^f - Th ^f - (Mi ^f . En ^f)	-208.46 (2・5・15)	Ⓐ	Ⓐ
6 : (No ^f . Au ^f) - Ja ^f - So ^f - Th ^f - Mi ^f - En ^f	-201.28 (2・5・15)	Ⓒ	Ⓐ
6 : No ^f - Ja ^f - Au ^f - So ^f - Th ^f - (Mi ^f . En ^f)	-199.90 (2・5・15)	Ⓐ	Ⓐ
4 : (No ^f . Au ^f) - (Ja ^f . So ^f) - Th ^f - (Mi ^f . En ^f)	-194.48 (2・5)	Ⓐ	—
7 : No ^f - Ja ^f - Au ^f - So ^f - Th ^f - Mi ^f - En ^f	-192.31 (2・5・15)	Ⓒ	Ⓐ
5 : (No ^f . Au ^f) - (Ja ^f . So ^f) - Th ^f - Mi ^f - En ^f	-191.67 (2・5)	Ⓐ	—

- Ⓐ { カテゴリー1. 日本語と同語順+その他+不明
 カテゴリー2. 日本語と対称的な語順
 Ⓒ { カテゴリー1. 日本語と同語順
 カテゴリー2. 日本語と対称的な語順+その他+不明

表12-2. AICの小さい順の最長距離法によるグループ構成(その他, 不明の言語を除いた場合, $c_2 = c_5 = c_{15} = 2$).

グループ数: 構成 (99言語)	AIC(選ばれた項目)
5 : (No ^f . Au ^f) - Ja ^f - So ^f - Th ^f - (Mi ^f . En ^f)	-160.13 (2・5・15)
4 : (No ^f . Au ^f) - (Ja ^f . So ^f) - Th ^f - (Mi ^f . En ^f)	-153.88 (2・5)
6 : (No ^f . Au ^f) - Ja ^f - So ^f - Th ^f - Mi ^f - En ^f	-151.05 (2・5・15)
5 : (No ^f . Au ^f) - (Ja ^f . So ^f) - Th ^f - Mi ^f - En ^f	-150.95 (2・5)
5 : (No ^f . Au ^f) - Ja ^f - So ^f - Th ^f - (Mi ^f . En ^f)	-149.67 (2・5)
5 : No ^f - Au ^f - (Ja ^f . So ^f) - Th ^f - (Mi ^f . En ^f)	-148.79 (2・5)

- カテゴリー1. 日本語と同語順
 カテゴリー2. 日本語と対称的な語順

表9~表12では分類の仕方, グループ数, カテゴリー数を問わず, 項目2「名詞と側置詞」の順序が共通に選択される. 次にこの項目2に続く分類に関わる重要な項目を選択するとき, 表9-1, 11-1は項目5もしくは項目15の2つの可能性を与える. すなわち語順規則から言語を眺めたとき

“名詞と側置詞”と「数詞と名詞」の順序”の2項目(2パラメータ)

もしくは

“名詞と側置詞”の順序と「疑問詞」の位置”の2項目(2パラメータ)

によって世界の言語はよく分類されるといってよいと思われる. 一見文章構造に関わりをもたない項目が分類にもっとも関与している.

7. 差の分布と2分割

表9～表12は、項目2「名詞と側置詞」が他の項目に突出して重要な指標であることを示すものであった。無側置詞言語を後置詞をもつ言語と同じカテゴリーに属するとすることで、世界の言語がほぼ2分されることが分割表データにおける項目の選択からも明らかにされた。

次に個々の項目にこだわらずに、グループ内の言語の第1要素と第2要素の差の分布からグループの特徴を調べてみる。言語*i*の項目*k*の第1要素を*i(k, 1)*、第2要素を*i(k, 2)*、対象とする項目の集合を*S_j*(*j* = 1, 2, 3, 4)と記述するとき

$$\frac{1}{\# \{S_j\}} \sum_{k \in S_j} \{i(k, 1) - i(k, 2)\}$$

の値を言語毎に求め、グループ内の言語の分布を図示した(図6)。ここで

上段：全130言語、
中段：後置詞型言語、
下段：前置詞型言語

で、それぞれ左端より、項目の集合*S_j*(*j* = 1, 2, 3, 4)が次の場合に対する分布である。

- S*₁ = {16項目, *k* = 1, ..., 19, *k* ≠ 11, 13, 17}
- S*₂ = {第1主成分を形成する項目群, *k* = 1, 2, 3, 7, 8, 9, 10, 19}
- S*₃ = {*S*₁ - *S*₂, *k* = 4, 5, 6, 12, 14, 15, 16, 18}
- S*₄ = {項目2・5, *k* = 2, 5}.

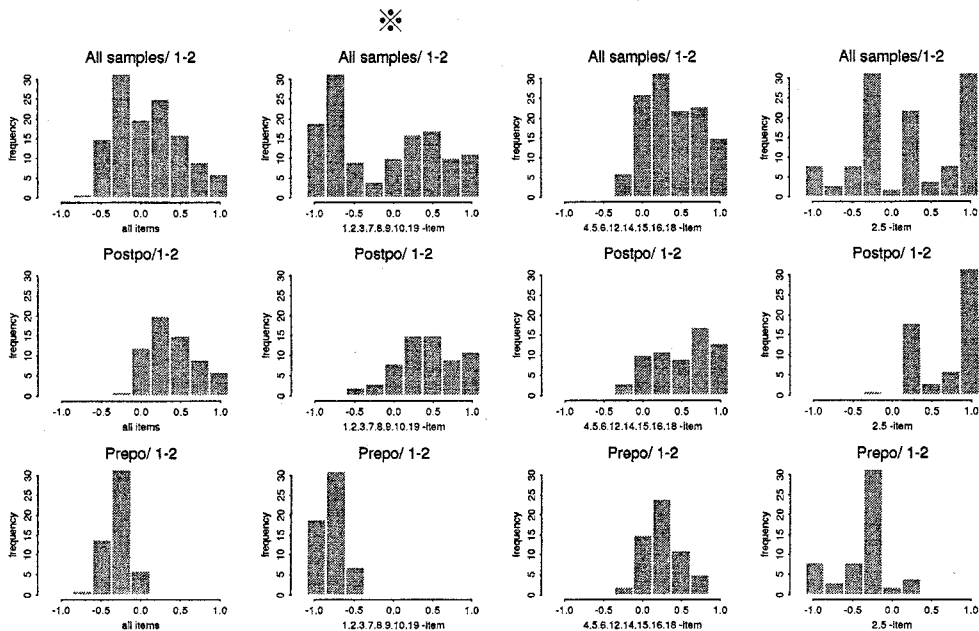


図6. 第1要素と第2要素の差の分布.

値が1に近いほど日本語と同一の語順規則をもつ項目の多い言語であることを示し、-1に近くなるほど日本語から遠ざかった逆の特徴をもつ。

※印の第1主成分と相関の高い項目の図では、前置詞型言語と後置詞型言語の2つの山形がみられ、両者の共通部分はわずかである。つまり第1主成分が言語を2分していることが読み取れる。これは、側置詞をもたない言語を後置詞をもつ言語と同じカテゴリーに属するとみなすことによって世界の言語が2分されるとしたことと重なる。項目2は差の分布からも第1主成分を十分に説明している。

8. 語順規則の統計的推移

Hawkins (1983), Bybee (1990) は前置詞, 後置詞が言語の分類に強くかかわっていることを指摘し, Greenberg (1966) は2つないしは3つの項目間の比較によって語順の規則について列挙した。角田の語順表はこれらの規則を統計的に裏付けるものであり, その規則の多くの部分が互いに独立と思われる語順「名詞と側置詞」と、「数詞と名詞」の順序もしくは「疑問詞」の位置でかなりよく説明されることがわかった。

今後, これらの2つの項目を基本とするランダム生成モデルを提案し, 語順規則の推移をとらえてみたい。語順規則があるパターンに収束していくのか, 永久に変化を続けていくものなのか, マルコフ連鎖として問題を定式化することにより解析したいと考えている。

謝 辞

査読者, 編集委員の方々に適切なご助言と貴重なご意見をいただきました。ここに感謝いたします。

参 考 文 献

- Becker, R.A., Chambers, J.M. and Wilks, A.R. (1988). *The New S Language*, Wadsworth, Pacific Grove, California (渋谷政昭, 柴田里程 訳 (1991). 『S言語 I・II』, 共立出版, 東京).
- Bybee, J. (1990). Asymmetries in the affixation of grammatical material, *Studies in Typology Diachrony*, Benjamins, Amsterdam.
- Greenberg, J.H. (1966). Some universals of grammar with particular reference to the order of meaningful elements, *Universals of Language* (ed. J.H. Greenberg), 73-113, MIT Press, Cambridge, Massachusetts.
- Hawkins, J.A. (1983). *Word Order Universals*, Academic Press, New York.
- 亀井孝, 河野六郎, 千野栄一 (1988, 1989, 1992). 言語学大辞典, 第1, 2, 3, 4巻, 三省堂, 東京.
- 河口至商 (1973). 『多変量解析入門』, 森北出版, 東京.
- 坂元慶行 (1985). 『カテゴリカルデータのモデル分析』, 共立出版, 東京.
- Sakamoto, Y. and Akaike, H. (1978). Analysis of cross classified data by AIC, *Ann. Inst. Statist. Math.*, **30**, 185-197.
- 柴谷方良, 影山太郎, 田守育啓 (1982). 『言語の構造——理論と分析——』, くろしお出版, 東京.
- 角田太作 (1991). 『世界の言語と日本語——言語類型論からみた日本語——』, くろしお出版, 東京.
- Tsunoda, T., Ueda, S. and Itoh, Y. (1995). Adpositions in word order typology, *Linguistics*, **33**(4), 741-761.
- 上田澄江, 伊藤栄明 (1992). 語順規則による世界の言語の階層クラスタ分析, *応用統計学*, **20**(3), 155-166.
- 上田澄江, 角田太作, 伊藤栄明 (1995). 言語の文法構造のデータの統計解析, 統計数理研究所共同研究リポート, No. 70.

The Classification of Languages
by the Two Parameter Model for Word Ordering Rule

Sumie Ueda

(The Institute of Statistical Mathematics)

Yoshiaki Itoh

(The Institute of Statistical Mathematics and
The Graduate University for Advanced Studies)

We analyze the data of word ordering rule for 130 languages collected by Tsunoda (1991, *Sekai no Gengo to Nihongo*, Kuroshio, Tokyo). The adposition (preposition and postposition) is a good indicator in linguistic typology to classify the languages into two groups. We compare the possible ways of classifications applying the statistical method introduced by Sakamoto and Akaike (1978, *Ann. Inst. Statist. Math.*, **30**, 185-197) on analysis of cross classified data and get a reasonable classification into 4 groups by using two parameters, i. e., adposition and numeral.