WALS データを用いた日本周辺言語の分析
Analysis of languages around Japan using The World Atlas of Language Structures' data

江原暉将
Terumasa EHARA

諏訪東京理科大学
Tokyo University of Science, Suwa
eharate @ rs. suwa. tus. ac. jp

1　はじめに

WALS online[1]では、The World Atlas of Language Structures[Haspelmath et. al, 2005]に基づき、言語学上の素性 142 個を取り上げ、世界の 2,560 言語に対して、その素性値をデータとして提供している。本研究では、WALSデータを用いて日本周辺の言語に対する素性値を分析し、日本語と素性値が類似する言語が何かを明らかにすることを目的としている。分析方法としては多次元尺度構成法を用いる。

言語学的分析に多次元尺度構成法を用いる研究には[江原, 1995]、[Croft and Poole, 2006]、[Albu, 2006]、[Ehara, 2006]などがある。[Ehara, 2006]では WALS データを用いているが語順に関する素性のみを利用している。今回は広く他の素性も用いて分析する。

2　使用データ

WALS データの中からまず日本周辺言語を抽出する。WALS には各言語にその使用中心地に当たる経緯度が記述されている。その値を用いて、東経 75 度から 180 度、北緯 0 度から 90 度の範囲の言語をまず抽出する。そこから手話(7 言語)とピジン・クレオール(4 言語)を除外した。その結果、429 言語が得られた。

WALS ではすべての言語に対して、すべての素性値が記述されているわけではなく、沢山の素性値が記述されている言語もあれば、ごく少数の素性値しか記述されていない言語もある。素性値の記述が少ないと、類似性の比較を行うときの信頼性が損なわれるので、上記 429 言語から素性値の記述が 40 以上の言語を選択した。その結果 78 言語が得られた。今回は、これらの言語を分析対象とする。

次に使用する素性を選択する。素性の中には他の素性から一意に決めることができるものが含まれている。例えば、主語、目的語、動詞の語順(素性番号 81)は主語と動詞の語順(素性番号 82)と目的語と動詞の語順(素性番号 83)から決定できる。このように他の素性から一意に決められる素性をまず除外した。そのような素性の番号は 3, 81, 95, 96, 97 である[2]。

WALS の素性は 11 種に分類されている。そのうち、語彙、手話、その他に分類された素性(129 から 142)は今回の分析では用いなかった。その結果 123 個の素性が得られた。

3　分析方法と結果

多次元尺度構成法を適用するために各言語間の非類似度を求める。ここでは素性値間の相対ハミング距離を非類似度とする。言語 l_1 と l_2 間の相対ハミング距離 $d(l_1,l_2)$ は以下のようにして求められる。F を素性の全体とし、$f(l)$ を言語 l に対する素性 $f \in F$ の素性値とする。素性値が定義されていない場合には、$f(l)=\phi$ とする。

$$d(l_1,l_2)=\frac{\sum_{f\in F} d_f(l_1,l_2)}{\sum_{f\in F} e_f(l_1,l_2)} \qquad (1)$$

ここで

$$d_f(l_1,l_2)=\begin{cases}0 & f(l_1)=\phi \vee f(l_2)=\phi \vee f(l_1)=f(l_2)\\ 1 & \text{else}\end{cases}$$

$$e_f(l_1,l_2)=\begin{cases}0 & f(l_1)=\phi \vee f(l_2)=\phi\\ 1 & \text{else}\end{cases}$$

つまり、l_1 と l_2 で共通して定義されている素性に

[1] http://wals.info/

[2] 素性番号および素性値番号の意味については WALS online を参照のこと。

対する両者の素性値の不一致の割合である。l_1とl_2の素性値がすべて一致すれば$d(l_1,l_2)=0$となり、すべて不一致であれば$d(l_1,l_2)=1$となる。l_1とl_2で共通して定義されている素性がない場合は、(1)の分母がゼロとなるが、今回の解析では、そのような例はなかった。相対ハミング距離で距離行列を求め、Torgerson の方法で内積行列を求めた。加算定数は Torgerson の単純法で計算した。内積行列に対して固有値を求めた。求まった固有値を大きい順に表 1 に示す。表 1 には累積寄与率も示されている。0.5 以上の正の固有値の総数は 25 個であった。これらの固有値を用いて空間配置を構成した。

表 1 上位の固有値と累積寄与率

順位	固有値	累積寄与率
1	10.63	0.19
2	4.09	0.26
3	2.92	0.31
4	2.48	0.35
5	1.97	0.38
6	1.85	0.42
7	1.69	0.45
8	1.48	0.47
9	1.42	0.50
10	1.39	0.52

空間配置において日本語と距離の近い言語とその距離を表 2 に示す。表 2 には、距離行列の構成で用いた相対ハミング距離で日本語と距離の近い言語およびその距離も示してある。両者での順位は類似しているが同一ではない。

表 2 日本語と距離の近い言語

	空間配置での距離		相対ハミング距離	
順位	言語名	距離	言語名	距離
1	Korean	0.6273	Korean	0.2973
2	Mangghuer	0.6926	Mangghuer	0.2973
3	Khalkha	0.6934	Newari (Kathmandu)	0.3235
4	Garo	0.7100	Nepali	0.3409
5	Burmese	0.7146	Khalkha	0.3585
6	Newari (Kathmandu)	0.7236	Jingpho	0.3714
7	Lepcha	0.7554	Malayalam	0.3721
8	Ladakhi	0.7642	Dagur	0.3721
9	Nepali	0.7646	Burmese	0.3905
10	Tamang	0.7697	Telugu	0.4000
11	Jingpho	0.7835	Yakut	0.4048
12	Meithei	0.7854	Tamang	0.4054
13	Dagur	0.8169	Garo	0.4096
14	Mandarin	0.8239	Buriat	0.4118
15	Telugu	0.8351	Ladakhi	0.4203
16	Ainu	0.8542	Meithei	0.4286
17	Yakut	0.8554	Tamil	0.4333
18	Mundari	0.8598	Tibetan (S S)	0.4333
19	Kannada	0.8648	Lahu	0.4426
20	Malayalam	0.8664	Mandarin	0.4495

各言語を第 2 固有値までの空間に配置したものを図 1 に示す。第 2 固有値までの累積寄与率は 26%と大きくないため、表 2 と図 1 では差が見られる。

図 1 では、目的語(O)と動詞(V)の語順を表す素性(83)の素性値が、OV である言語を菱形(青)で、VO である言語を四角(赤)で表している。OV 言語は右側に、VO 言語は左側に集中している。広範な素性を用いた今回の分析でも[Ehara, 2006]と同様に目的語と動詞の語順が空間配置の最も主要な要因であることが分かる。

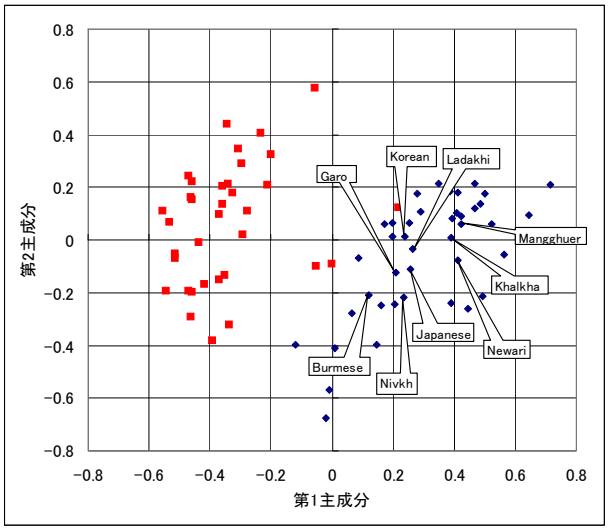

図 1　2 次元空間配置

25 次元の空間配置を元にクラスター分析を行って得られた樹状図を図 2 に示す[3]。樹状図は Tibetan(Standard Spoken)とそれ以外にまず分岐する。2 段目はNicobarese(Car)とJapaneseの間で分離している。Nicobareseより上の部分はAcehneseを除いて 83 番の素性値がすべてVOであり、Japaneseより下の部分はMandarin、Cantonese、Kashmiriを除いてすべてOVである。Acehnese は No dominant order であり、Mandarin、Cantonese、KashmiriはVOである。Tibetan(Standard Spoken)は 83 番の素性値が定義されていない。このように樹状図から見ても、少数の例外を除いて、VO型とOV型でまず分離されることが分かる。

[3]本分析には「Excel アドイン工房」のクラスター分析ソフトを利用した。
http://www.jomon.ne.jp/~hayakari/index.html

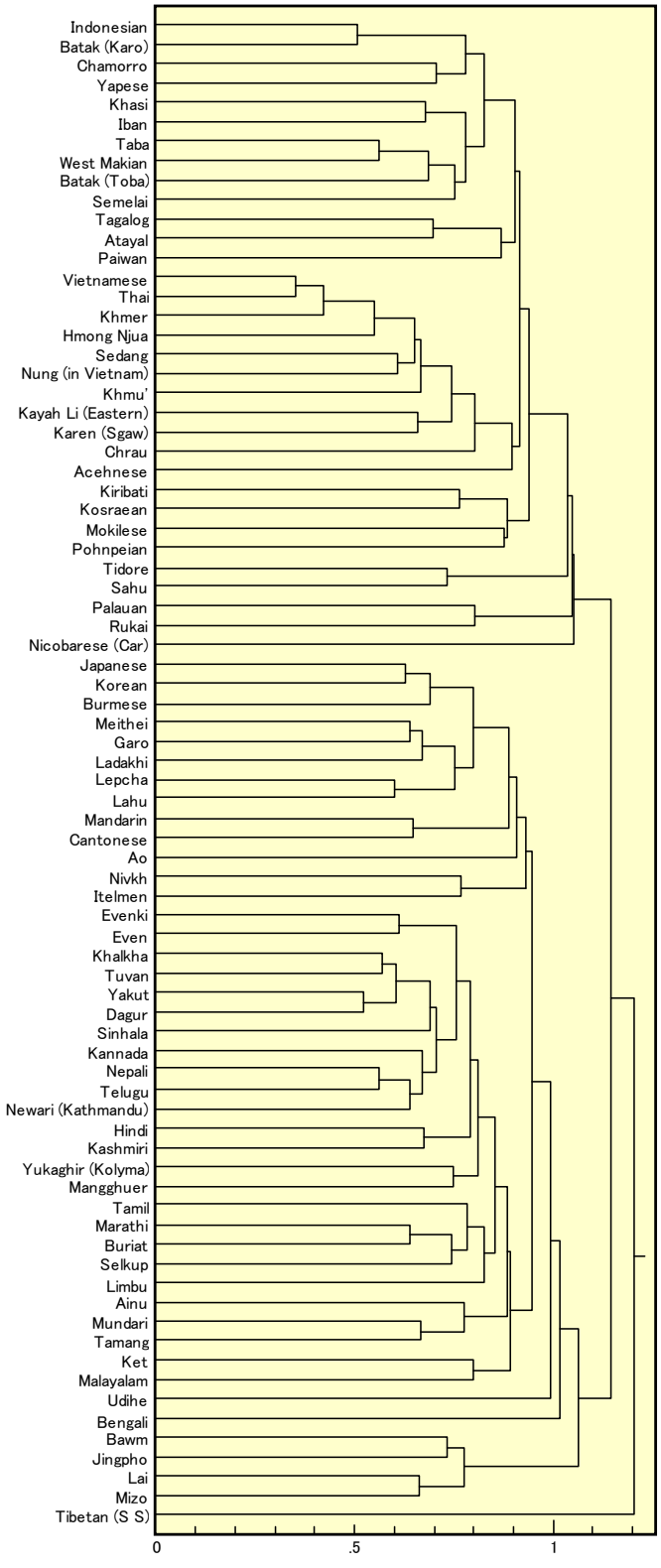

図 2 クラスター分析結果

4 おわりに

WALS データを用いて、日本周辺の言語に対する各種素性値を多次元尺度構成法によって分析した。その結果、Korean、Mangghuer、Khalkha、Burmese、Newari などの言語が日本語と素性値が近いことなどが分かった。これらの言語と日本語の素性値の比較を付表 1 に示す。WALS によると、Mangghuer と Khalkha は Mongolic に、Burmese と Newari は Tibeto-Burman に属している。Japanese と Korean は孤立している。

今回は、語彙に関する比較は行わなかった。WALS には語彙に関する素性値も記述されているが、WALS では言語類型論の立場からの素性のみが記述されており、語彙比較の立場からはデータが不足していると考え除外した。

表 2 を見ると日本語と最も素性値の類似している Korean でも相対ハミング距離が 0.29 である。つまり日本語と Korean とでは 29%の素性値が異なっていることが分かる。このような意味で、本分析は概括的なものであり、個々の言語の比較は詳細になされるべきことは言うまでもない。

5 参考文献

[Albu, 2006] Albu, Mihai: Quantitative Analyses of Typological Data, Von der Fakultät für Mathematik und Informatik der Universität Leipzig Dissertation Dorctor Rerum Naturalium, 2006.

[Croft and Poole, 2006] Croft, William and Poole, Keith T. : Inferring universals from grammatical variation: multidimensional scaling for typological analysis, Unpublished Manuscript, 2006.

[江原, 1995]江原暉将 : 多次元尺度構成法を用いた語順パラメータの間の関係付け, 言語処理学会第 1 回年次大会発表論文集, pp.173-176.

[Ehara, 2006] Ehara, Terumasa: Word order characteristics analyzed by multi dimensional scaling, 言語処理学会第 14 回年次大会発表論文集, A1-3.

[Haspelmath et. al, 2005] Haspelmath, Martin; Dryer, Matthew S.; Gil, David and Comrie, Bernard (Ed.): The World Atlas of Language Structures, Oxford University Press, 2005.

付表１ 日本語と距離の近い言語の素性値番号

素性番号	Japanese	Korean	Mangghuer	Khalkha	Burmese	Newari	素性番号	Japanese	Korean	Mangghuer	Khalkha	Burmese	Newari
1	2	3	3	3	4	3	64	2	2	2			
2	2	3	2	3	3	1	65	2	1		2	2	
4	4	1	4	4	4	2	66	1	1		1	4	
5	2	1	2	2	2	2	67	2	2		2	2	
6	3	1	2	1	1	1	68	4	4		3	2	
7	1	2	1	1	1	1	69	2	2	2	2	2	2
8	1	2	2	2	2	2	70	4	4		4	5	
9	3	2	2	2	1		71	2	2		2	2	
10	2	2		2	2		72	4	4		1	4	
11	1	1	1	1	1	1	73	2	2		1	2	2
12	2	2	2		2	2	74	1	3		2	3	
13	2	1	1	1	3	1	75	3	3		3	3	
14				1			76	2	2		3	2	
15				5			77	2	2		3	1	
16				2			78	2	2		6	1	
17				5			79	4	4		4	4	
18	1	1	1	1	1	1	80	1	1		1	1	
19	1	1	1	1	5	1	82	1	1	1	1	1	1
20	1	1		1	1		83	1	1	1	1	1	1
21	1	1		1	1		84	3		3		6	
22	3	4		2	2		85	1	1	1	1	1	1
23	2	2		2	2		86	1	1	1	1	1	1
24	2	2		2	2		87	1	1	1	1	2	1
25	2	2		2	2		88	1	1	1	1	1	1
26	2	2	2	2	2	2	89	1	1	1	1	2	2
27	2	1		1	2		90	2	2	2	2	2	2
28	1	1		4	1		91	1	1		1	1	1
29	1	1		1	1		92	2	6	6	2	6	2
30				1	1		93	2	2		2	2	2
31				1	1		94	2	2	2	2	4	5
32				1	1		98	2	2		2	2	
33	9	2	7	2	7	8	99	2	2		2	2	
34	2		4	4		3	100	1	1		1	1	
35	8	6		8	2		101	5	5	5	5	5	
36	1	1		1	1	1	102	1	1		1	1	
37	4	5	4	5	5		103	1	1		1	1	
38	1	5	1	5	5		104	1	1		1	1	
39	3	3		3	2		105	1	1	1	1	2	
40	1	1		1	1		106	2	2		2	1	
41	3	3		2	2		107	1	1		1	2	
42	3	2		3			108	3	3		3	3	
43	1	1		2	1		109	8	8		8	8	
44	2	3		6	3		110		2			2	
45	4	4		2	4		111	2	2		2	4	
46	1	1					112	1	2	2	1	6	1
47	1	1		1	1		113	2	3		3	2	
48	2	2		2	2		114	5	4		1	3	
49	7	6		7	7		115	1	1			1	
50	2	2		2	2		116	1	2	2	1	2	1
51	9	1	6	1	1	1	117	1	1		1	1	
52	2	2	2	2			118	3	3		2	1	
53	4	4		5	8		119	1	1		2	1	
54	4	4		4	4		120	1	1		2	2	
55	3	3		1	3	3	121	1	1		1	1	
56	3			3	3		122	4	4		4	4	
57			4	2		4	123	4	4		4	4	
58	2	2		2	2		124	4	1			1	
59	1	1		1	1		125	2	1		3		
60	2	6		6	3		126	2	1		3		
61	7	6		2	2		127	1	1				
62	4	1		2	5		128	1	1				
63	2	1		1	2								