Word order characteristics analyzed by multi dimensional scaling
多次元尺度構成法を用いた語順類型の分析

Terumasa EHARA
江原暉将

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

1 Introduction

There are many languages in the world. These languages have the similarity and the difference each other. Linguistic typology intends to clarify these characteristics. The word order universal is one of the main topics in the field [Greenburg, 1966], [Comrie, 1981], [Hawkins, 1983], [Tsunoda, 1991], [Matsumoto, 2006].

This paper presents word order characteristics analyzed by multi dimensional scaling (MDS). The results are language distribution in the one and two dimensional eigen space and interpretation of eigen vectors by word order characteristics. The research methodology is same as the previous work [Ehara, 1995]. The new thing is the database. In this paper, the data is obtained from [Dryer, 2005]. We also compare the results with the former results which uses the data mainly from [Tsunoda, 1991].

2 Word order characteristics and their parameterization

Dryer presents 13 word order characteristics in the book of "The World Atlas of Language Structures"(WALS) [Dryer, 2005]. We select 7 characteristics from them, because of to hold the richness of the quantity of languages. They are listed in Table 1. They are all binary characteristics. We use three values for the characteristics. For example, in the case of "Order of subject (S) and verb (V)"; "SV", "VS" and "No dominant order" are used. The languages which have another characteristic value are not used in our analysis. In other words, we only consider languages which have one of the three characteristic values for all 7 characteristics. We can obtain 576 languages from [Dryer, 2005] by this filtering.

When we parameterize the characteristics, the value same to Japanese is coded with +10. The opposite value is coded with -10. "No dominant order" is coded with 0. So, for the case of order of S and V, SV is coded with +10 and VS is coded with -10, because Japanese has SV order.

Table 1 Word order characteristics and their parameterization

No.	Word order characteristics	+10	-10
1	Subject(S) and Verb(V) in a declarative sentence	SV	VS
2	Object(O) and Verb(V) in a declarative sentence	OV	VO
3	Noun(N) and Adposition(Ap)	N-Ap	Ap-N
4	Genitive(G) and Noun(N)	GN	NG
5	Determiner(D) and Noun(N)	Dm-N	N-Dm
6	Adjective(A) and Noun(N)	AN	NA
7	Numeral(Nm) and Noun(N)	Nm-N	N-Nm

3 Analysis by the Multi dimensional scaling

Like [Ehara, 1995], Torgerson's multi dimensional scaling algorithm is applied to analyze the above data of 576 languages. The additive constant obtained by Torgeson's simple method is zero. We get the inner product matrix by Young-Householder transformation which uses the center of mass as the origin. Eigen values of the inner product matrix are shown in Figure 1. The cumulated contribution ratio up to the second eigen value is 75%. Figure 2 shows the histograms of 576 languages on the one dimensional eigen spaces which correspond to the first and second principal components. Figure 3 shows the mapping of 576 languages in the two dimensional eigen space, which we named “word order space”.

The histogram of languages on the first principal component is nearly concave. The histogram of languages on the second principal component is nearly convex.

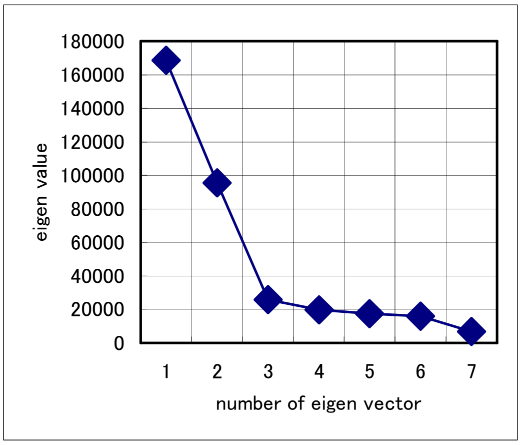
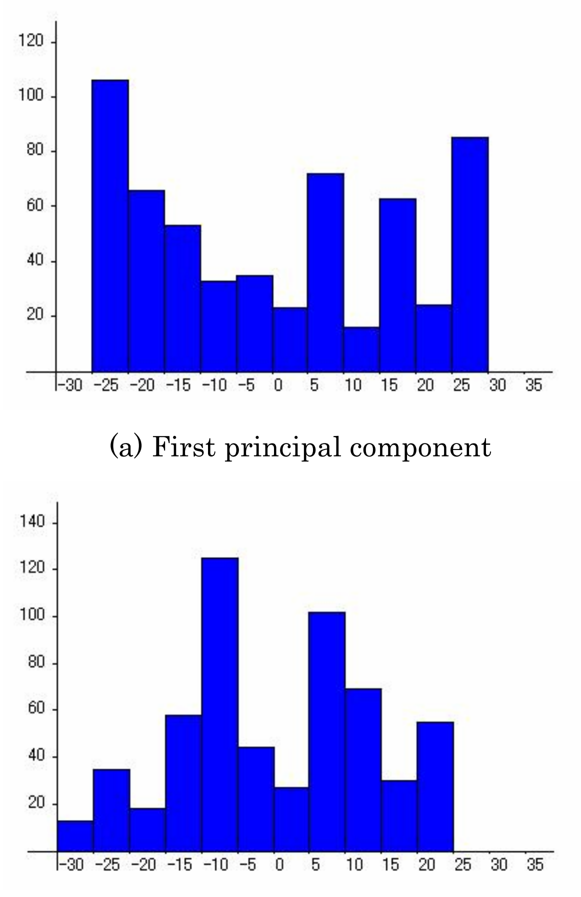


Fig. 1 Eigen values of inner product matrix

4 Relation between word order parameters

We can investigate relations between word order parameters from the mapping of languages in the word order space.

For the parameter Pk (k=1,...,7), we divide the language set $L=\{Li;\ i=1,...,576\}$ to three disjoint subsets. $Lk+ = \{Li;\ Pk(Li) = +10\}$, $Lk0 = \{Li;\ Pk(Li) = 0\}$, $Lk- = \{Li;\ Pk(Li) = -10\}$. When the average positions of $Lk+$ and $Lk-$ in the word order space are $qk+$ and $qk-$ respectively, we define the parameter vector qk (for Pk) as $qk = qk+ - qk-$. The angle ak formed qk and the first principal axis is listed in Table 2 with number of elements of $Lk+$, $Lk-$ and $Lk0$.



(a) First principal component

(b) Second principal component

Fig. 2 Histogram of 576 languages projected on the principal component

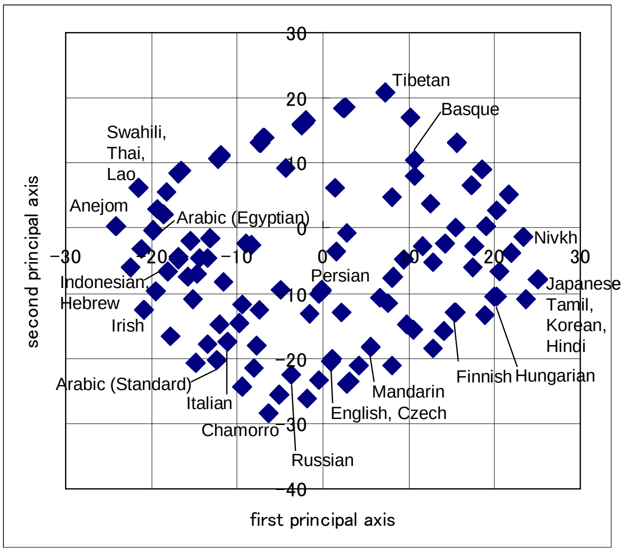


Fig. 3 Distribution of 576 languages in the word order space

From Table 2 we divide word order parameter set to three disjoint subsets. They are *P1* = {Pk; 0≤ak≤20}, *P2* = {Pk; 70≤ak≤90}, *P3* = {Pk; 20<ak<70}.

3 members of *P1* correlate to the first principal component. Members of *P1* are, also, mutually correlated.

From this analysis, first principal axis is interpreted as one of the three members of *P1.* They are the order of object and verb, the order of noun and adposition and the order of genitive and noun. Since the number of members of Lk0 is the fewest for k=2, the order of object and verb is most suitable for the interpretation of the first principal axis. No word order characteristic is well interpreted as second principal axis.

Table 2 Classification of word order parameters

Group No.	Parm No.	Angle(deg)	Lk+	Lk−	Lk0
3	1	48.0	459	84	33
1	2	16.8	241	307	28
1	3	16.8	271	275	30
1	4	15.0	311	227	38
3	5	37.3	169	377	30
3	6	27.7	271	305	0
3	7	65.3	268	281	27

5 Comparing with the former results

[Ehara, 1995] uses 20 word order characteristics which include all characteristics listed in Table 1. However, the parameter values are real valued between -10 and +10 in [Ehara, 1995]. Instead of the real valued, this paper uses the three leveled, +10, -10 and 0. [Ehara, 1995] uses 126 languages' data mainly from [Tsunoda, 1991]. The intersection of language set of [Ehara, 1995] and this paper includes 65 languages. They are listed in Figure 4. The mapping of these 65 languages in two dimensional word order space of [Ehara, 1995] and of this paper is shown in Figure 5. Correspondence lines between the results of [Ehara, 1995] and this paper are also shown in the Figure. Looking at Figure 5, we can recognize the similarity of the shapes of two mappings. Our result is roughly obtained from the former to rotate it counterclockwise.

Abkhaz, Alyawarra, Aymara, Basque, Bulgarian, Burmese, Burushaski, Chamorro, Choctaw, Chukchi, Coast Tsimshian, Comanche, Copala Trique, Czech, Danish, Dutch, Egyptian Arabic, English, Evenki, Finnish, French, Georgian, German, Guarani, Hindi, Hopi, Hungarian, Indonesian, Irish, Isthumus Zapotec, Italian, Japanese, Kannada, Korean, Lao, Luisenyo, Mam, Mandarin Chinese, Modern Greek, Modern Hebrew, Navajo, Nivkh, Norwegian, Palauan, Panjabi, Persian, Piraha, Polish, Quechua, Russian, Serbo-Croatian, Slavey, Spanish, Swahili, Swedish, Tamil, Thai, Tibetan, Tongan, Turkish, Urubu-Kaapor, Vietnamese, Welsh, Yaqui, Yoruba

Fig. 4 Languages analyzed in both [Ehara, 1995] and this paper

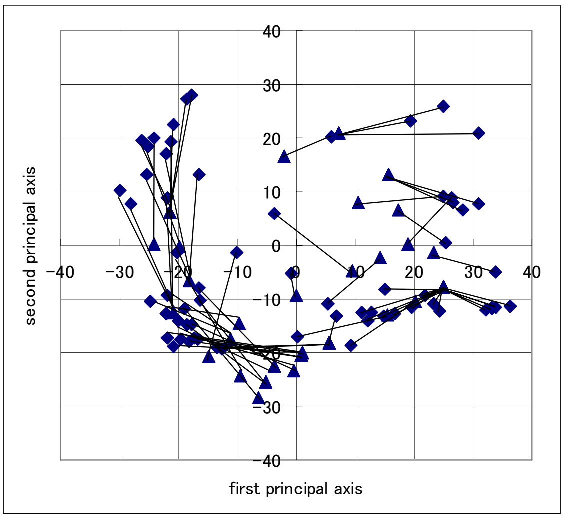


Fig. 5 Comparison of analyzed results.
◆:[Ehara, 1995], ▲:this paper

6 Related works

Some researchers apply MDS technique to linguistic typology study.

Croft uses MDS results to make classification of linguistic phenomena. They are structure of indefinite pronouns and structure of tense and aspect expressions [Croft and Poole, 2006].

Albu uses WALS data to get distance measures between languages [Albu, 2006]. However, Albu's interest is to investigate phylogenetic structure of languages. Our interests are on the word order similarities of phylogenetically different languages.

7 Conclusion

Quantitative study of word order typology using multi dimensional scaling is presented. Cumulated contribution ratio is 75% up to second principal component. First principal component can be interpreted by the order of object and verb and second principal component can not be interpreted by any word order characteristics. In [Ehara, 1995], first principal component was interpreted by the same, but second principal component was interpreted by the order of adjective and noun.

8 Bibliography

[Albu, 2006] Albu, Mihai: Quantitative Analyses of Typological Data, Von der Fakultät für Mathematik und Informatik der Universität Leipzig Dissertation Dorctor Rerum Naturalium, 2006.
http://lingweb.eva.mpg.de/phylogenetictools/QALD.pdf

[Comrie, 1981] Comrie, Bernard: Language Universals and Linguistic Typology, University of Chicago Press, 1981.

[Croft and Poole, 2006] Croft, William and Poole, Keith T. : Inferring universals from grammatical variation: multidimensional scaling for typological analysis, Unpublished Manuscript, 2006.
http://www.unm.edu/~wcroft/Papers/MDSpaper-3.pdf

[Dryer, 2005] Dryer, Matthew S.: Word Order, The World Atlas of Language Structures, Chapter F, pp.330-397, Oxford University Press, 2005.

[Ehara, 1995] Ehara, Terumasa: Relation among Word Order Parameters Analyzed by Multi-Dimensional Scaling, Proceedings of The First Annual Meeting of The Association for Natural Language Processing, pp.173-176, 1995 (Originally in Japanese).
http://www.rs.suwa.tus.ac.jp/eharate/ehara/ronbun.files/relation_among_word_order_parameters_ENG.pdf
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, 1995.
http://www.rs.suwa.tus.ac.jp/eharate/ehara/ronbun.files/relation_among_word_order_parameters.pdf

[Greenburg, 1966] Greenburg, Joseph: Language universals, with special reference to feature hierarchies, Janua Linguarum, Series Minor 59, Mouton, 1966.

[Hawkins, 1983] Hawkins, John A.: Word Order Universals, Academic Press, 1983.

[Matsumoto, 2006] Matsumoto, Katsumi: A Worldwide Perspective on Languages - Historical Linguistics and Linguistic Typology -, Sanseido, 2006 (in Japanese).
松本克己：世界言語への視座－歴史言語学と言語類型論－，三省堂, 2006.

[Tsunoda, 1991] Tsunoda, Tasaku: Languages of the World and Japanese, Kurosio Shuppan, 1991 (in Japanese).
角田太作：世界の言語と日本語，くろしお出版, 1991.