Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica D 198 (2004) 333–339
PHYSICA D
www.elsevier.com/locate/physd

The Ising model for changes in word ordering rules in natural languages

Yoshiaki Itoh, Sumie Ueda*

The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan

Received 3 February 2003; received in revised form 14 September 2004; accepted 21 September 2004

Communicated by Y. Kuramoto

Abstract

The order of 'noun and adposition' is an important parameter of word ordering rules in the world's languages. The seven parameters, 'adverb and verb' and others, depend strongly on the 'noun and adposition'. Japanese as well as Korean, Tamil and several other languages seem to have a stable structure of word ordering rules, while Thai and other languages, which have the opposite word ordering rules to Japanese, are also stable in structure. It seems therefore that each language in the world fluctuates between these two structures like the Ising model for finite lattice.

Keywords: Ising model; Word ordering rules; Languages

1. Majority rule for word ordering rule

Natural languages provide an attractive object of various interdisciplinary studies. The problem on how the universals of languages evolve has recently attracted considerable attention from physicists [1]. The importance of adpositions (prepositions and postpositions) as a parameter is recognized in word order typology [2,3]. In previous studies [4–6] based on the word order data of Tsunoda (Table 1) [7], in which Japanese is taken as a standard for comparison, 130 languages are classified by using 19 word order parameters. The result is that, except for one or two languages, the 130 languages are nicely divided into two groups: (a) prepositional languages, and (b) other languages (i.e. postpositional languages and adpositionless languages). Adpositionless languages behave like postpositional languages in terms of other word order parameters. The second important parameter of word order

* Corresponding author. Tel.: +81 354 218 770; fax: +81 354 218 750.
E-mail address: ueda@ism.ac.jp (S. Ueda).

0167-2780/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.physd.2004.09.006

Table 1
Word order data by Tsunoda [7]: from parameters 2 to 10, and also for parameters 12, 18 and 19, the plus sign '+' and minus sign '−' are used – wherever possible – with Japanese as the standard of comparison

No.	Word order parameters	Japanese	English	Thai	Punjabi
1.	S, O and V	SOV, etc.	SVO	SVO	SOV
2.	Noun and adposition	+	−	−	+
3.	Genitive and noun	+	+, −	−	+
4.	Demonstrative and noun	+	+	−	+
5.	Numeral and noun	+	+	−	+
6.	Adjective and noun	+	+	−	+
7.	Relative clause and noun	+	−	−	Other; +
8.	Proper noun and common noun	+	−, +	−	NA (not available)
9.	Comparison of superiority	+	−	−	+
10.	Main verb and auxiliary verb	+	−	−, +	+
11.	Adverb and verb	Before V	Various	Various	Immediately after S
12.	Adverb and adjective	+	+, −	−	+
13.	Question marker	Sentence-final	None	Sentence-final immediately after focus of question	None
14.	S, V inversion in general questions	None	Present	None	None
15.	Interrogative word	Declarative sentence type	Sentence-initial	Declarative sentence type	Immediately before verb
16.	S, V inversion in special questions	None	Present	None	None
17.	Negation marker	Verbal suffix	Immediately after verb	Immediately before focus of negation	Immediately before verb
18.	Conditional clause and main clause	+	+, −	−, +	+
19.	Purpose clause and main clause	+	−	−	−

This '+/−' method is used for some other parameters as well when applicable. Japanese is convenient as the standard of comparison. Thus, if a given language has 'noun + postposition' like Japanese, then it will be assigned '+' for parameter 2. If a given language has 'preposition + noun' in contrast with Japanese, it will be assigned '−' for this parameter. If a given language has some other order, then an explanation is given as much as possible. If such an explanation is not feasible, it is simply presented as 'other'. If a given language has two orders, such as 'adjective + noun (+)' and 'noun + adjective (−)', then the order that appears to be more common is presented first. Thus, if 'adjective + noun' is more common than 'noun + adjective' in a given language, we have '+/−' rather than '−/+' for the parameter 6 of this language. When a given language has alternative possibilities other than the word order listed, this is shown with 'etc.'. The expression 'NA (not available)' indicates that no information is available.

is 'numeral and noun'. By applying a statistical method for categorical data [8], the clustering of the 130 languages is well explained by the two word ordering parameters, 'noun and adposition' and 'numeral and noun' [5,6].

Greenberg [2] makes use of the position of adpositions in 7 of the 45 universals he proposes (pp. 110–112). For example, in languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions, the former almost always precedes the latter. Tsunoda's data show that out of the 19 word order parameters, the parameters 1 (i.e. 'S, O and V'), 3 (i.e. 'genitive and noun'), 7 (i.e. 'relative clause and noun'), 8 (i.e. 'proper noun and common noun'), 9 (i.e. 'comparison of superiority'), 10 (i.e. 'main verb and auxiliary verb') and 19 (i.e. 'purpose clause and main clause') depend strongly on the parameter 2 (i.e. 'noun and adposition') [5,6]. For example, our study of the data gives the following statistical laws on cooccurrence with three parameters.

(i) If a language has preposition, and if the language is the VO language, then the genitive tends to follow the noun.
If a language has postposition, and if it is the OV language, the genitive tends to precede the noun.
(ii) If the main verb of a language follows the auxiliary verb, and if the relative clause follows the noun, the purpose clause tends to follow the noun.
If the main verb of a language precedes the auxiliary verb, and if the relative clause precedes the noun, the purpose clause tends to precede the noun.

We study the cooccurrence of word ordering rules of the world's languages by using the Ising model. While the spontaneous magnetization can occur only in the thermodynamic limit, in a finite lattice, the system makes excursions from states with uniformly negative magnetization through an intermediate mixed-phase state to states with uniformly positive magnetizations [9]. Our ternary interaction model of a finite population makes similar excursions, if we introduce a mutation to the other type for each particle. Consider a system of particles of two types A and B. At each step, three particles are taken at random. Suppose that two of the three particles are of the type A (B) and one is of the type B (A); the three particles become three particles of the type A (B) with probability p and the three particles of the type B (A) with probability $1 - p$. We continue this step sequentially. As a result, if p is larger than 1/2, the type of the majority is advantageous to the type of the minority, while if p is less than 1/2, the type of the minority is advantageous. The model for p less than 1/2 becomes a simple caricature of the Ising model for the temperature higher than the critical point, although the model does not have space variables, since we neglect the position of the particles. The case of p larger than 1/2 corresponds to the Ising model for the temperature lower than the critical point, where the majorities are advantageous to the minorities. Here, we apply the ternary interaction model to study word ordering rules of the world's 130 languages [7].

A pair of very similar word orders may appear only by chance. For instance, the difference between Tamil and Japanese is as small as the difference between Italian and Spanish in Tsunoda's word order table [7]. Japanese as well as Korean, Tamil and several other languages seem to have a stable structure of word ordering rules. We put the value $+1$ for each of the eight parameters 'noun and adposition' and the above seven parameters for Japanese. We put -1 for each of the eight parameters for the languages which have the opposite word ordering rules like Thai. The changes in the values of the parameters of each language may depend on the values of other parameters as can be seen from the above (i) and (ii) for three parameters. It seems that each language in the world fluctuates between these two stable structures. Hence, the system of word ordering rules in Thai, or the opposite system like Japanese, Korean, Mongolian and others, seems to be reasonably described by the above ternary interaction model of the majority rule, and could be explained in terms of the economy of communications minimizing misunderstanding. Each language in the world has its own history and personality. Word ordering rules of each language may change at random like a mutation in a population of eight individuals, each of which corresponds to a parameter of word ordering rules. The biological mechanism for performing language may be ultimately functionally driven by the need for rapid and efficient communications in real time [1]. Hawkins [3] discusses the relationship between innateness and functional pressures in the explanation of language universals and argues that some innate processing mechanisms have responded to functional pressure that makes rapid and efficient communications possible. This observation seems to support our stochastic model for the change of word ordering rule. It seems that each language fluctuates between the above two stable structures like the Ising model for finite lattice. The fractal concept [10] is useful for understanding the structure of sentences [11]. The parse tree is convenient for representing the phrase structure rule in generative grammar [12]. English is a prepositional language and Japanese is a postpositional language. English sentences are usually represented by right branching parse trees, while Japanese sentences are represented by left branching parse trees. Evolution of word ordering rules or evolution of parse trees, which may be closely related with each other, are interesting problems and will be able to be discussed by using evolutionary game theory [13].

2. Interaction among parameters

We consider the numerical values of the above eight parameters. For the parameters of each language that can not clearly be classified into $+1$ or -1, we assign a numerical value between $+1$ and -1 by a digitization of the description in Table 1.

We take an arithmetic mean of the values of the eight parameters for each language. When some parameter values are unknown, an arithmetic mean is taken only for the parameter values available. We call the arithmetic mean as the measure of 'postposition-ness'. Fig. 1 shows the histogram of the measure of 'postposition-ness' considering the eight parameters for the 130 languages [5].

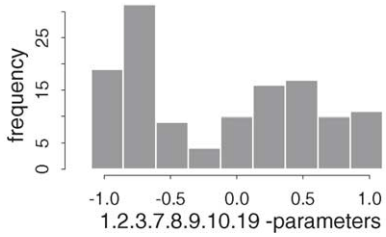

Fig. 1. The histogram for the measure of 'postposition-ness' considering the eight parameters on the 130 languages.

The changes of word ordering rules could be represented by the above ternary interaction model of eight particles $1, 2, \ldots, 8$, by assuming that $p = 1$ and the mutations for each particle changing its sign occur at a certain rate at the end of each step. The trajectory of the number n_+ of particles of the type $+1$ is given in Fig. 2a for the first 1000 steps, where it is assumed that a mutation to the opposite sign occurs in the system with probability 0.06 for a particle chosen at random at each step. The histogram for the numbers of visits has two modes as shown in Fig. 2b.

We represent word ordering rules of each of the 130 languages by a vector of 66-dimension. A Manhattan distance d between two languages is defined as Eq. (1). A random walk is made over the 130 languages to simulate changes of word ordering rules. Choose a language at random at first. Then, we choose the second language randomly from those languages staying within a given Manhattan distance $d = 0.6$ from the first language. The third language is chosen similarly, i.e. chosen from the languages staying within the distance $d = 0.6$ from the second language. Such random steps are repeated many times. Fig. 3a shows a resulting trajectory obtained from the first 1000 such steps. The histogram of the number of visits in Fig. 3b, which is calculated for the first 10,000 steps, is also characterized by two modes similar to those in Figs. 1 and 2b. The states in which the eight parameters have the same sign seem to stay relatively stable over the random walk.

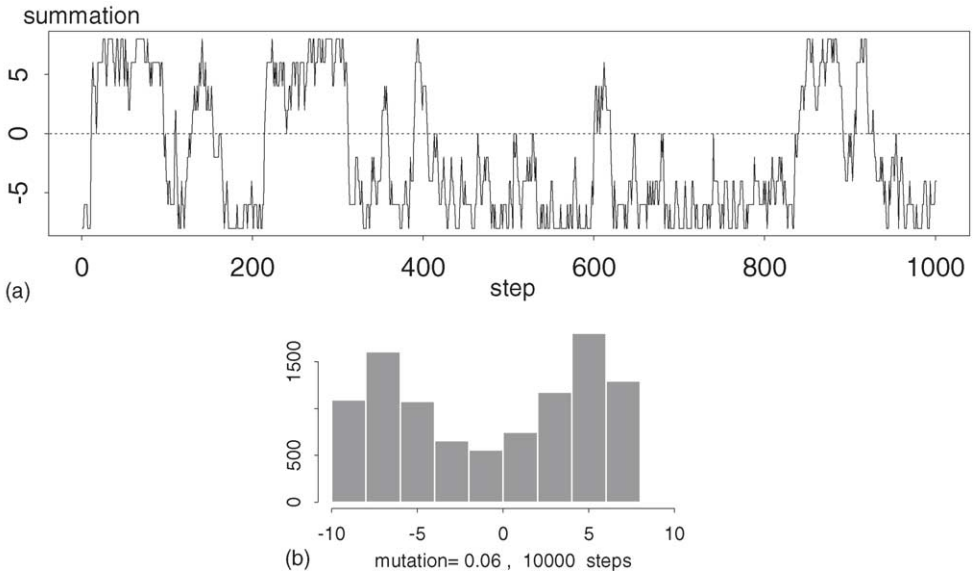

Fig. 2. The simulation of the majority rule model of 10,000 steps: (a) the trajectory of the first 1000 steps; (b) the histogram with two modes for numbers of visits by the trajectory.

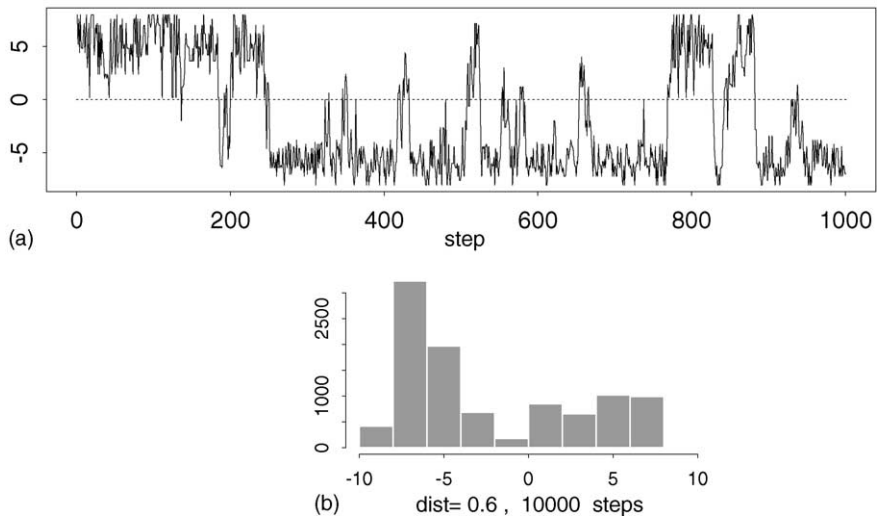

Fig. 3. The random walk on the 130 languages of 10,000 steps: (a) the trajectory of the first 1000 steps starting from Japanese; (b) the histogram of the numbers of visits by the trajectory.

3. Digitization of word ordering rule

Our numerical methods are briefly mentioned. We associate a 66-dimensional vector to each language in order to digitize the word order data. In the data by Tsunoda, 'NA' was used for the parameter when the corresponding value was currently unavailable [7]. Our present study as well as our previous study [4] is based on the data by Tsunoda [7]. Our method can be extended to define the distance between two languages including 'NA', applying the idea in the S language [14], although we do not go into its details in this article.

A numerical value between 0 and 1 was assigned to each of the 19 parameters according to the description of word orders concerned. The sum of the components corresponding to a given parameter should be equal to 1, i.e.

$$\sum_{m=1}^{n_k} x_{i,k,m} = 1,$$

for $i = 1, \ldots, 130$, $k = 1, \ldots, 19$, where n_k is the dimension of parameter k and $x_{i,k,m}$ is the value of the m-element of parameter k of language i.

We represent the state of the parameter 1 by a two-dimensional vector for the six possible combinations of the parameter S, O and V. The OV language contains SOV, OSV and OVS languages. The VO language contains SVO, VSO and VOS. In the case of Japanese which has 'SOV, etc.', $(1, 0)$ is assigned. In the case of English which has 'SVO', $(0, 1)$ is assigned. Similarly, in the case of Modern Greek, which has 'SVO, VSO, etc.', $(0, 1)$ is assigned. For a parameter with a three-dimensional vector, we represent it by $(1, 0, 0)$ if the order is the same to Japanese. If its order is reversed to Japanese, then we represent it by $(0, 1, 0)$. Otherwise, we represent it by $(0, 0, 1)$. For example, the parameter 3 (genitive and noun) for English is '$+, -$' which is represented by $(0.6, 0.4, 0)$. When a parameter of a language has more than one word order possibility, the ratio is given by the similarity of the competing possibilities. We use 6-, 5- and 13-dimensional vectors, respectively, for the parameters 11, 13 and 17. For each of the 130 languages, we construct a 66-dimensional vector from the 19 vectors by arranging the coordinates of the 19 vectors from the vector for parameter 1 to the vector for parameter 19. We give two examples for the digitization. The 19 slashes in the following vector classify the 66 coordinates into the 19 parameters.

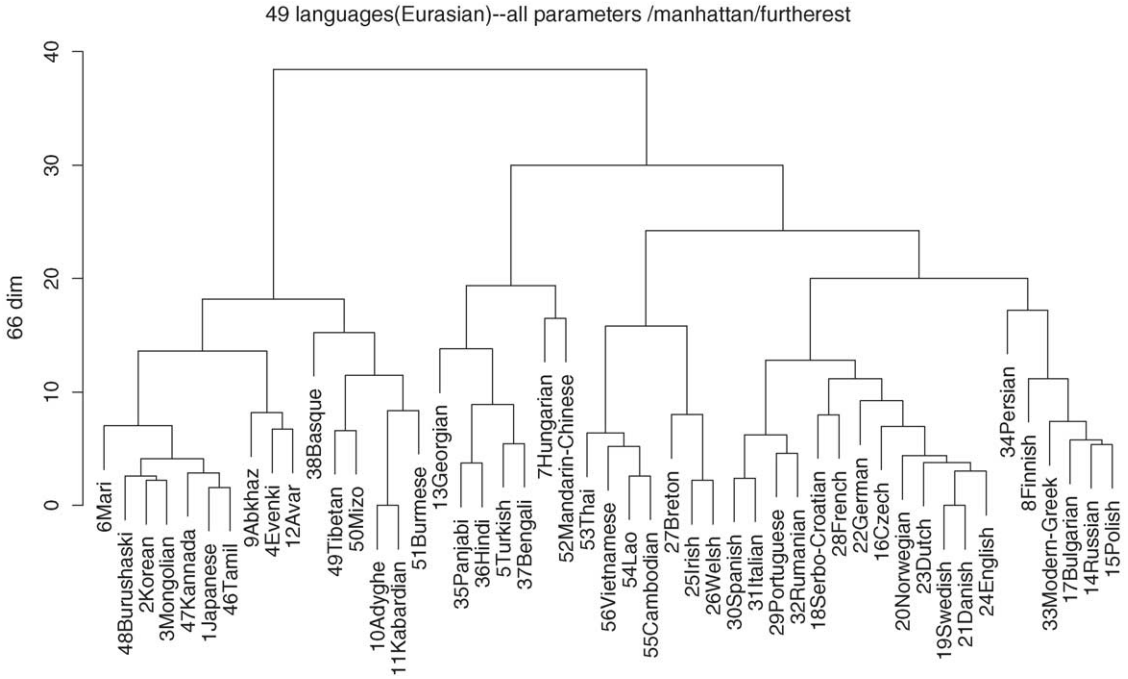

Fig. 4. The hierarchical clustering for the Eurasian languages applying the furthest method.

Japanese:

(1, 0/1, 0, 0/1, 0, 0/1, 0, 0/1, 0, 0/1, 0, 0/1, 0, 0/1, 0, 0/1, 0/1, 0, 0/1, 0, 0/1, 0, 0, 0, 0, 0/1, 0, 0/1, 0, 0,
0, 0/1, 0/1, 0, 0/1, 0/1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0/1, 0/1, 0/)

English:

(0, 1/0, 1, 0/0.6, 0.4, 0/1, 0, 0/1, 0, 0/1, 0, 0/0, 1, 0/0.4, 0.6/0, 1, 0/0, 1, 0/0, 1, 0, 0, 0,
0/0.6, 0.4, 0/0, 0, 1, 0, 0/0, 1/0, 1, 0/0, 1/0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0/0.6, 0.4/0, 1/)

We define the distance between two languages i and j as

$$d(i, j) = \frac{1}{19} \sum_{k=1}^{19} \sum_{m \in n_k} | x_{i,k,m} - x_{j,k,m} | \leq 2. \tag{1}$$

The distances between English and Japanese, between English and Thai, and between Japanese and Thai are 1.45, 1.05 and 1.52, respectively. By applying a standard software of statistical data analysis, the S language [14], to the above distances, we derived the hierarchical clustering for the Eurasian languages (Fig. 4), which are a part of the 130 languages for the random walk.

4. Evolution of languages

Our present study on Tsunoda's table seems to give a possible answer to an aspect of the problem how the language universals evolve. The word ordering rule can change at random under the influence of the functional

pressure, by which the communications become more rapid and efficient. The word ordering rule of each language in the world seems to fluctuate between the two stable structures, a typical postpositional language structure and a typical prepositional language structure, like the Ising model for finite lattice.

References

[1] J.A. Hawkins, M. Gell-Mann, Preface, in: J.A. Hawkins, M. Gell-Mann (Eds.), The Evolution of Human Languages, Addison-Wesley, Redwood City, 1992.
[2] J.H. Greenberg, Some universals of grammar with particular reference to the order of meaningful elements, in: J.H. Greenberg (Ed.), Universals of Language, MIT Press, Cambridge, MA, 1966, pp. 73–113.
[3] J.A. Hawkins, Word Order Universals, Academic Press, New York, 1983.
[4] T. Tsunoda, S. Ueda, Y. Itoh, Adpositions in word order typology, Linguistics 33 (4) (1995) 741–761.
[5] S. Ueda, Y. Itoh, The classification of languages by the two parameter model for word ordering rule, Proc. Inst. Stat. Math. 43 (2) (1995) 341–365 (in Japanese).
[6] S. Ueda, Y. Itoh, Classification of natural languages by word ordering rule, in: O. Opitz, M. Schwaiger (Eds.), Explanatory Data Analysis in Empirical Research, Springer-Verlag, 2002, pp. 180–187.
[7] T. Tsunoda, World's Languages and Japanese—Japanese observed from Linguistic Type Theory, Kuroshio Shuppan, Tokyo, 1991 (in Japanese).
[8] Y. Sakamoto, H. Akaike, Analysis of cross classified data by AIC, Ann. Inst. Stat. Math. 30 B (1978) 185–197.
[9] K. Binder, D.W. Heermann, Monte Carlo Simulation in Statistical Physics, Springer, Berlin, 1988.
[10] B.B. Mandelbrot, The Fractal Geometry of Nature, W.H. Freeman, San Francisco, CA, 1977.
[11] M. Tokieda, Japanese Grammar, Iwanami Shoten, Tokyo, 1950 (in Japanese).
[12] N. Chomsky, The Logical Structure of Linguistic Theory, Plenum Press, New York, 1975.
[13] M.A. Nowak, D.C. Krakauer, Evolution of languages, Proc. Nat. Acad. Sci. U.S.A. 96 (1999) 8028–8033.
[14] R.A. Becker, J.M. Chambers, A.R. Wilks, The New S Language, AT&T Bell Laboratories, Wadsworth & Brooks, California, 1988.