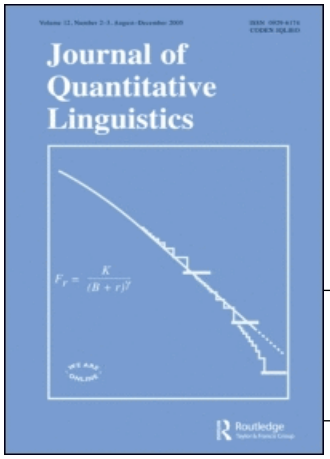

Journal of Quantitative Linguistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t716100702

On the dynamics of the compounding of Japanese kanji with common and proper nouns*

Katsuo Tamaoka; Peter Meyer; Shogo Makioka; Gabriel Altmann

Online Publication Date: 01 May 2008

To cite this Article: Tamaoka, Katsuo, Meyer, Peter, Makioka, Shogo and Altmann, Gabriel (2008) 'On the dynamics of the compounding of Japanese kanji with common and proper nouns*', Journal of Quantitative Linguistics, 15:2, 136 — 153

To link to this article: DOI: 10.1080/09296170801961801
URL: http://dx.doi.org/10.1080/09296170801961801

Journal of Quantitative Linguistics
2008, Volume 15, Number 2, pp. 136–153
DOI: 10.1080/09296170801961801

On the Dynamics of the Compounding of Japanese Kanji with Common and Proper Nouns*

Katsuo Tamaoka[1], Peter Meyer[2], Shogo Makioka[3] and Gabriel Altmann[4]
[1]Kashiwa, Japan; [2]Göttingen, Germany; [3]Osaka, Japan; [4]Lüdenscheid Germany

ABSTRACT

The present study examines the dynamics of the kanji combinations that form common (or general) and proper nouns in Japanese. The following three results were obtained. First, the degree of distribution results from two similar processes which are based on a steady-state of birth-and-death processes with different birth and death rates, resulting in a positive negative binomial distribution with the proper nouns and in a positive Waring distribution with common nouns. Second, all rank-frequency distributions follow the negative hypergeometric distribution used very frequently in ranking problems. Third, the building of kanji compounds follows a dissortative strategy. The higher the outdegree of a kanji, the more it prefers kanji with lower indegrees. A linear dependence can be observed with common nouns, whereas the relationship between compounded kanji is rather curvilinear with proper nouns. The actual analytical expression is not yet known.

1. INTRODUCTION

In previous articles (Tamaoka & Altmann, 2004; 2005) a database of 1945 basic kanji (Tamaoka et al., 2002; Tamaoka & Makioka, 2004) was used to show that Japanese kanji can combine with one another to different extents in order to build new compound words. For example the

*Address correspondence to: Katsuo Tamaoka, College of Foreign Studies and Graduate School of Language Education, Reitaku University, 1-1, 2-chome, Hikarigaoka, Kashiwa, Chiba, 277-8686, Japan. E-mail: ktamaoka@reitaku-u.ac.jp

kanji 学 /gaku/, meaning in English 'to learn' or 'learning', can be used as the left-hand component of compounds such as

学校 /gaQkoR/ (Q = a moraic geminate or voiceless obstruent and R = a moraic long vowel) 'school'
学生 /gakusei/ 'student'
学者 /gakusya/ 'scholar' or as the right-hand component, as in
留学 /ryuRgaku/ 'study aboard'
入学 /nyuRgaku/ 'school admission'
文学 /buNgaku/ (N = a moraic nasal) 'literature'

An interesting question arises as to the extent of this combinability. Since two kanji combined with one another can be considered as a graph with two vertices and one (directed) edge, the whole set of selected kanji (vertices) with all realized combinations (edges) establishes a (directed) graph G = {V, E} where V = set of vertices (kanji), E = set of edges (realized combinations), or a network. The edges are directed from the first part of the compound to the second. Graphs of this kind have been thoroughly examined not only in statistical physics but also in many other fields (for a general survey, see Albert & Barabási, 2002; for linguistics, Batagelj et al., 2002a, b; Ferrer i Cancho & Solé (2003); Ferrer i Cancho et al., 2004; Steyvers & Tenenbaum, 2005).

We considered such combinations separately for common (or general) nouns and for proper nouns, using 1934 commonly used kanji in the first case and 1430 kanji in the second case. The graph of combinations was represented as a matrix of 1934 × 1430 cells in a spreadsheet file.[1] Networks constructed in this way display a number of properties by which they can sometimes be classified as random graphs, small worlds, scale-free etc. The aim of the present study is to investigate some properties of these graphs as applied to Japanese two-kanji compound nouns.

2. OUTDEGREE AND INDEGREE

In a compound consisting of components, the edge is directed from the first component to the second. The number of edges going out from a

[1]Japanese has 1945 basic kanji, but some kanji with zero combinations were omitted.

vertex (here kanji) designates its outdegree, the number of edges ending at a vertex designates its indegree. As has been shown in the previous study (Tamaoka & Altmann, 2004), the outdegree and the indegree are not necessarily equal, i.e. there is not necessarily a quantitative symmetry. Moreover, an AB compound does not imply the existence of BA. For example, if the word 学者 /gakusya/ 'scholar' is considered as an AB compound, the BA compound of 者学 /syagaku/ does not exist in Japanese. As such, there is no guarantee of qualitative symmetry.

The study of the distribution of degrees is routine work in network analysis. The shape of the distributions tells us something about the type of the network and the process of attachment when new vertices are added. In literature we can find a number of distributions derived and used for these purposes. In linguistics it is reasonable to examine both the distribution of (in- and out-) degrees and their rank-frequency distribution, a problem introduced by Zipf (1935).

As already pointed out, the process of building a network of compounds is a stochastic birth-and-death process with simple birth and death rates leading to a negative binomial distribution (see Table 1a and Figure 1). In this process new compounds are built preferentially with those kanji which already have many compounds. It is to be noted that a random building process would lead to a Poisson distribution and the avoidance of those kanji would already build several compounds for new compounds which lead to a binomial distribution.

In analysing this process, the kanji that do not have a follower will be simply left out of outdegree counting and those which do not have a predecessor will be left out of indegree counting, because, from the point of view of network analysis, they are not yet a part of the graph. Thus, the frequency of $x = 0$ was not taken into account during fitting. This manipulation resulted in the 1934 $\times$ 1430 matrix. Since the total set of basic kanji consists of 1945 characters, we would expect a 1945 $\times$ 1945 matrix if all kanji were combinable with one another.

Considering now the same phenomenon in common nouns, it is easily shown that the negative binomial distribution does not yield a satisfactory fit. The difference between proper nouns and common nouns consists in their semantics: proper nouns need not have any meaning while common nouns must have a meaning. The fate of some proper nouns is also controlled by biological or other factors that influence linguistic creativity. Commonly-used family names (e.g. 山本 /yamamoto/, 鈴木 /suzuki/) and company names (e.g. ヤマハ /yamaha/,

Table 1a. Frequencies of outdegrees of Japanese two-kanji compound proper nouns (positive negative binomial distribution).

X	f_x	NP_x	X	f_x	NP_x	X	f_x	NP_x
1	216	210.57	35	7	6.84	69	3	2.69
2	139	116.13	36	3	6.61	70	2	2.63
3	81	81.04	37	5	6.39	71	5	2.58
4	72	62.47	38	4	6.18	72	3	2.52
5	44	50.90	39	4	5.98	73	2	2.47
6	52	42.96	40	4	5.79	74	6	2.41
7	42	37.16	41	0	5.61	75	4	2.36
8	36	32.73	42	6	5.44	76	1	2.31
9	36	29.22	43	6	5.27	77	4	2.26
10	27	26.37	44	5	5.12	78	1	2.21
11	28	24.01	45	7	4.97	79	1	2.17
12	17	22.02	46	4	4.83	80	5	2.12
13	15	20.32	47	6	4.69	81	1	2.08
14	14	18.85	48	3	4.56	82	1	2.04
15	14	17.57	49	8	4.43	83	4	2.00
16	17	16.43	50	4	4.31	84	3	1.96
17	18	15.43	51	3	4.20	85	1	1.92
18	12	14.53	52	2	4.08	86	0	1.88
19	12	13.71	53	4	3.98	87	1	1.84
20	14	12.98	54	7	3.87	88	0	1.81
21	12	12.31	55	4	3.77	89	0	1.77
22	13	11.70	56	5	3.68	90	1	1.74
23	7	11.14	57	3	3.59	91	0	1.70
24	4	10.63	58	4	3.50	92	1	1.67
25	10	10.15	59	5	3.41	93	1	1.64
26	8	9.71	60	2	3.33	94	4	1.61
27	10	9.30	61	1	3.25	95	0	1.58
28	9	8.92	62	5	3.17	96	0	1.55
29	6	8.56	63	2	3.10	97	3	1.52
30	8	8.23	64	4	3.02	98	1	1.49
31	4	7.92	65	2	2.95	99	1	1.46
32	6	7.62	66	1	2.89	$\geq$100	88	90.69
33	5	7.35	67	3	2.82			
34	7	7.09	68	3	2.76			

$k = 0.1138$, $p = 0.0097$, $DF = 170$, $X^2 = 154.01$, $P = 0.81$, $N = 1299$.

トヨタ /toyota/), and place names (e.g. 東京 /toRkyoR/, 広島 /hirosima/) can exhibit different frequencies depending upon social, economical and biological factors. In contrast, compound common nouns denote concrete things or abstract ideas and matters so that these words might be associated with their linguistic and semantic nature. In any case, this

Table 1b. Frequencies of indegrees of Japanese two-kanji compound common nouns (positive Waring distribution).

X	f_x	NP_x	X	f_x	NP_x	X	f_x	NP_x
1	229	219.72	35	3	5.18	69	1	2.13
2	107	109.82	36	5	5.01	70	3	2.08
3	85	72.97	37	5	4.84	71	0	2.04
4	65	54.49	38	1	4.69	72	0	2.00
5	51	43.38	39	1	4.54	73	1	1.96
6	31	35.97	40	5	4.40	74	2	1.92
7	30	30.66	41	5	4.26	75	1	1.88
8	24	26.68	42	6	4.14	76	2	1.85
9	28	23.59	43	3	4.01	77	0	1.81
10	24	21.11	44	4	3.90	78	0	1.78
11	18	19.08	45	4	3.79	79	3	1.74
12	15	17.39	46	2	3.68	80	3	1.71
13	18	15.96	47	2	3.58	81	1	1.68
14	18	14.73	48	1	3.49	82	3	1.65
15	13	13.66	49	3	3.39	83	3	1.62
16	14	12.73	50	2	3.30	84	3	1.59
17	12	11.91	51	5	3.22	85	2	1.56
18	8	11.18	52	2	3.14	86	4	1.53
19	7	10.53	53	4	3.06	87	0	1.51
20	11	9.95	54	2	2.98	88	0	1.48
21	10	9.41	55	5	2.91	89	3	1.45
22	10	8.93	56	1	2.84	90	3	1.43
23	12	8.49	57	3	2.77	91	0	1.40
24	10	8.09	58	1	2.71	92	0	1.38
25	4	7.72	59	2	2.65	93	0	1.36
26	5	7.38	60	3	2.59	94	0	1.33
27	6	7.06	61	1	2.53	95	2	1.31
28	5	6.76	62	1	2.47	96	3	1.29
29	3	6.49	63	2	2.42	97	2	1.27
30	3	6.24	64	1	2.36	98	0	1.25
31	6	6.00	65	2	2.31	99	2	1.23
32	2	5.77	66	3	2.26	$\geq$100	99	97.18
33	3	5.57	67	2	2.22			
34	9	5.37	68	2	2.17			

$k = 0.0061$, $p = 0.0063$, $X^2 = 61.91$, $DF = 75$, $P = 0.86$, $N = 1136$.
Note: An even better fit can be achieved if one pools classes with $NP_x > 1$ (here $NP_x > 5$ was taken).

difference should be captured by different birth and death rates in the stochastic process. The process will be considered the same, namely a birth-and-death process, because this type of process underlies the

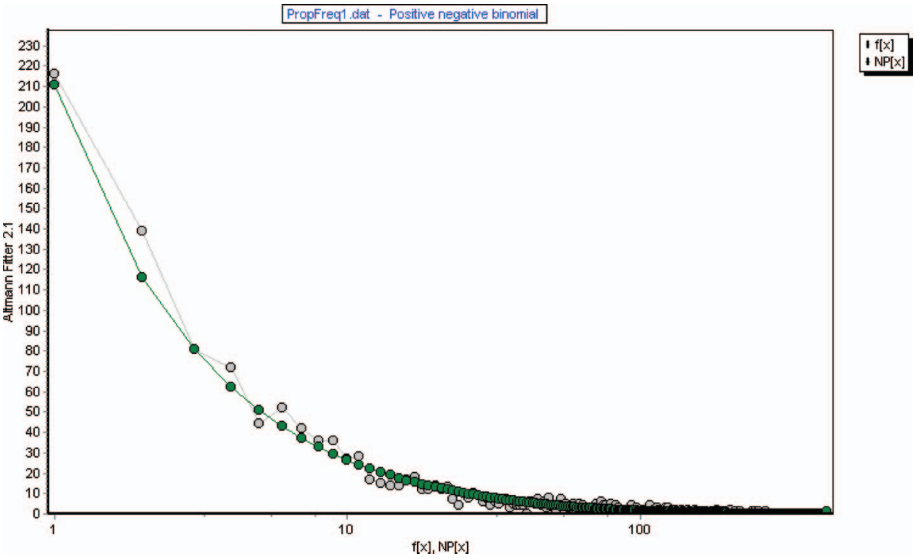

Fig. 1. Observed and calculated frequencies of outdegrees of Japanese two-kanji compound proper nouns.

background of any linguistic inventory building. However, we now try to determine the birth and death ratios as follows:

$$\text{birth rate:} \quad \lambda_x = b + x$$
$$\text{death rate:} \quad \mu_x = b + n + x.$$

Then we obtain the balancing equations

$$(b+1)P_1 = (b+n+2)P_2$$

$$(b+x+b+n+x)P_x = (b+x-1)P_{x-1} + (b+n+x+1)P_{x+1}, \quad x = 2,3,4,\ldots$$

whose solution results in a positive Waring distribution (= zero-truncated Waring distribution)

$$P_x = \frac{b}{n}\,\frac{n^{(x)}}{(b+n+1)^{(x)}}, \quad x = 1,2,3,\ldots$$

Even a simple estimator made from the first two frequency classes is sufficient to show the good fit of this model. We obtain from P_1 and P_2

$$\hat{b} = \frac{\hat{P}_1\hat{P}_2}{\hat{P}_1(1-\hat{P}_1)-\hat{P}_2} = \frac{f_1 f_2}{f_1(N-f_1)-Nf_2}$$

$$\hat{n} = \frac{N\hat{b}-f_1(1+\hat{b})}{f_1}.$$

Here we obtain $\hat{b} = 1.9421, \hat{n} = 22.9294$. The result of fitting for the distribution of indegrees is shown in Table 2a.

The fitting of the model to outdegrees is shown in Table 2b. Though the empirical frequencies do not decrease monotonously at the beginning, the fit is good. Alternative pooling of classes could have resulted in an even better chi-square.

If we ignore the direction of the edges (orientation) and merely consider the associativity of a noun, i.e. the degree of a vertex, we obtain a richer distribution both for proper nouns and for common nouns. Here again the zeroes are automatically dropped because a noun with zero degree does not belong to the network. As might be expected, the combined indegree and outdegree yield a mixed negative binomial distribution. A preliminary approximation has been performed with the aid of the one-displaced mixed negative binomial. Instead of tables we show only the graphs.

As can be seen in Figure 3, proper nouns have a more regular combination of degrees than common nouns. For example, many two-kanji compound proper nouns can be created with 山 /yama/ as the first kanji; combined with 本 /moto/, this yields one of the most often used family names, viz. 山本 /yamamoto/. In the same manner, other common family names are created, like 山下 /yamasita/, 山内 /yamauti/, 山口 /yamaguti/, all having the same kanji in the first position. On the other hand, as Figure 2 shows, there are some conspicuous tendencies towards multimodality. Common nouns can be created by a variety of kanji combinations. Further research regarding this distribution should be conducted.

3. RANK-FREQUENCY DISTRIBUTION

In linguistics it is common practice to consider the rank-frequency distribution of entities. Though it is possible to transform the two aspects

Table 2a. Frequencies of indegrees of Japanese two-kanji compound common nouns.

X	f_x	NP_x	X	f_x	NP_x	X	f_x	NP_x
1	137	137.00	35	12	10.87	68	1	2.93
2	122	122.00	36	10	10.34	69	2	2.84
3	125	109.12	37	12	9.85	70	1	2.75
4	89	98.00	38	16	9.39	71	1	2.67
5	88	88.35	39	9	8.96	72	1	2.59
6	85	79.93	40	11	8.55	73	2	2.51
7	65	72.55	41	10	8.17	74	5	2.44
8	77	66.06	42	11	7.81	75	1	2.36
9	60	60.32	43	6	7.47	76	4	2.30
10	57	55.23	44	11	7.15	77	1	2.23
11	58	50.70	45	3	6.85	78	1	2.17
12	49	46.65	46	7	6.57	79	2	2.10
13	47	43.03	47	4	6.30	80	1	2.05
14	31	39.77	48	5	6.04	81	2	1.99
15	33	36.84	49	9	5.80	82	2	1.93
16	42	34.19	50	4	5.57	83	0	1.88
17	37	31.78	51	6	5.36	84	1	1.83
18	27	29.60	52	4	5.15	85	3	1.78
19	26	27.62	53	1	4.96	86	0	1.73
20	33	25.81	54	6	4.77	87	2	1.69
21	28	24.15	55	5	4.60	88	2	1.64
22	18	22.64	56	4	4.43	89	0	1.60
23	17	21.24	57	5	4.27	90	1	1.56
24	20	19.97	58	0	4.12	91	1	1.52
25	12	18.79	59	6	3.97	92	1	1.48
26	23	17.70	60	4	3.84	93	0	1.45
27	22	16.70	61	2	3.71	94	4	1.41
28	14	15.77	62	4	3.58	95	1	1.38
29	15	14.91	63	7	3.46	96	2	1.34
30	17	14.11	64	7	3.35	97	1	1.31
31	12	13.36	65	5	3.24	98	1	1.28
32	10	12.67	66	5	3.13	99	0	1.25
33	8	12.03	67	2	3.03	$\geq$100	46	78.35
34	15	11.43						

$b = 1.9421$, $n = 22.9294$, $X^2 = 102.38$, $DF = 97$, $P = 0.33$, $N = 1825$.

into one another, the correspondence is not always satisfactory, especially in simpler cases. Thus we simply try to find an adequate ranking function for the in- and outdegrees ordered in decreasing order. For ranking problems, there are a great number of different distributions derived on the basis of different assumptions. However, they all hold true

Table 2b. Frequencies of outdegrees of Japanese two-kanji compound proper noun.

X	f_x	NP_x	X	f_x	NP_x	X	f_x	NP_x
1	109	127.60	33	11	13.57	65	0	2.90
2	113	116.90	34	12	12.83	66	4	2.79
3	128	107.26	35	13	12.14	67	2	2.67
4	100	98.56	36	16	11.49	68	1	2.57
5	96	90.69	37	12	10.88	69	5	2.47
6	85	83.56	38	13	10.31	70	1	2.37
7	90	77.10	39	13	9.78	71	5	2.28
8	75	71.22	40	9	9.28	72	3	2.19
9	78	65.88	41	8	8.81	73	1	2.11
10	65	61.01	42	11	8.37	74	3	2.03
11	51	56.56	43	8	7.95	75	1	1.95
12	61	52.50	44	5	7.56	76	3	1.88
13	54	48.78	45	7	7.19	77	0	1.81
14	41	45.38	46	5	6.84	78	3	1.74
15	38	42.25	47	4	6.52	79	3	1.68
16	51	39.38	48	10	6.21	80	0	1.62
17	26	36.74	49	7	5.92	81	1	1.56
18	36	34.32	50	10	5.64	82	1	1.50
19	26	32.08	51	6	5.38	83	1	1.45
20	25	30.01	52	2	5.14	84	2	1.40
21	24	28.10	53	6	4.90	85	2	1.35
22	18	26.34	54	5	4.68	86	1	1.30
23	23	24.70	55	4	4.48	87	0	1.26
24	14	23.19	56	2	4.28	88	3	1.21
25	19	21.79	57	2	4.09	89	0	1.17
26	12	20.48	58	3	3.92	90	0	1.13
27	14	19.27	59	5	3.75	91	0	1.09
28	12	18.15	60	2	3.59	92	2	1.06
29	20	17.10	61	4	3.44	93	1	1.02
30	20	16.12	62	4	3.29	$\geq$94	45	37.45
31	15	15.21	63	4	3.16			
32	12	14.37	64	2	3.03			

$b = 4.0032$, $n = 53.6646$, $X^2 = 90.23$, $DF = 91$, $P = 0.50$, $N = 1870$.

only for a restricted set of data. Recently, the negative hypergeometric distribution in particular has been successfully applied in capturing the frequency of occurrence of letters, words, phonemes, and musical units. It must be noted that none of the well known ranking distributions (Zipf-Mandelbrot, Lerch, Waring, Lavalette, etc.), except the negative hypergeometric, could be adequately fitted to our data. Using the negative hypergeometric in its 1-displaced form we obtained the

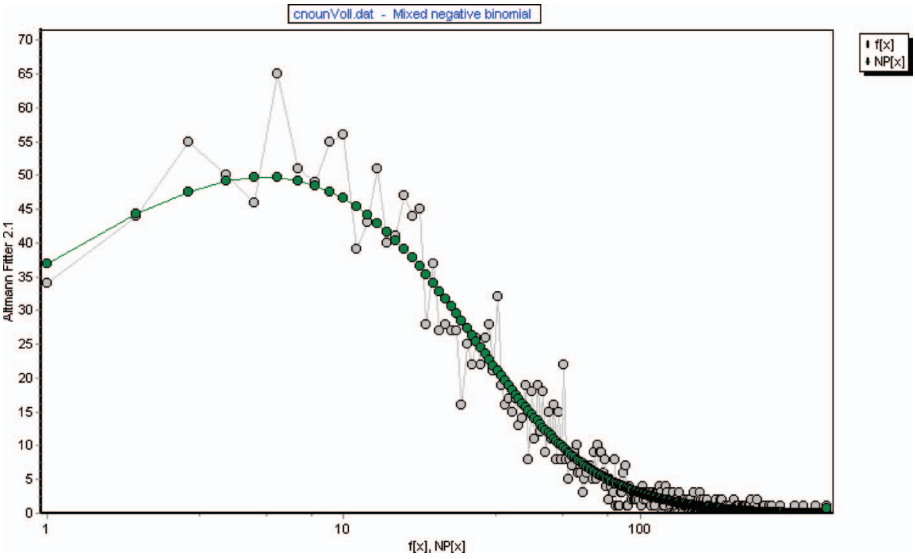

Fig. 2. Mixed negative binomial fitted to degrees of kanji (common nouns).

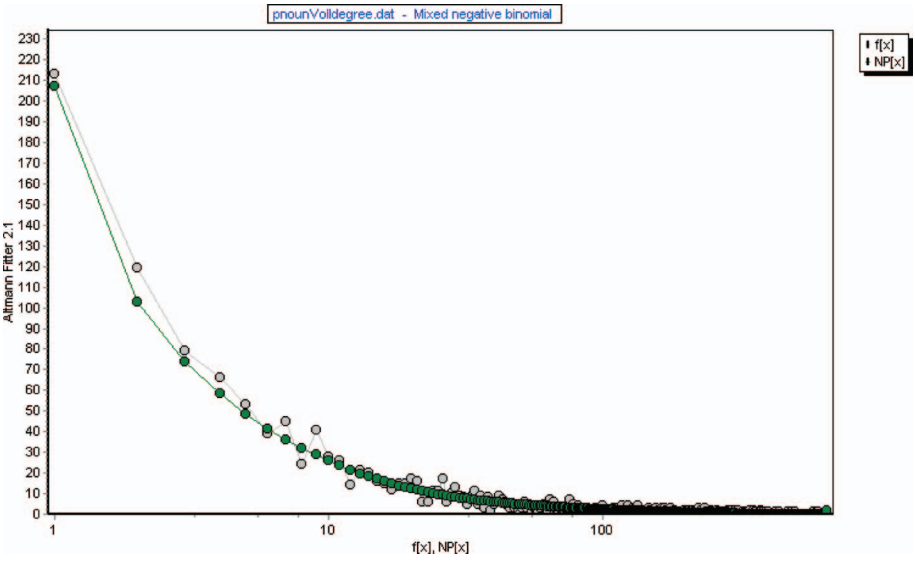

Fig. 3. Mixed negative binomial fitted to degrees of kanji (proper nouns).

following results (K, M, n are the parameters, C is the contingency coefficient which has been evaluated because of the enormous sample sizes):

Common nouns
outdegree: $K = 3.2175$, $M = 0.6223$, $n = 2100$, $X^2_{1830} = 318.33$, $C = 0.0089$ (Fig. 4a.)
indegree: $K = 3.9519$, $M = 0.6840$, $n = 2250$, $X^2_{1778} = 400.19$, $C = 0.0111$ (Fig. 4b.)

Proper nouns
outdegree: $K = 38.3551$, $M = 0.7744$, $n = 10000$, $X^2_{1199} = 268.05$, $C = 0.0081$ (Fig. 4c.)
indegree: $K = 949.8466$, $M = 0.7094$, $n = 200000$, $X^2_{1013} = 371.44$, $C = 0.0113$ (Fig. 4d.)

Instead of giving the tabular data, which is extensive, we simply present the graphs of observed and computed values in Figures 4a to 4d.

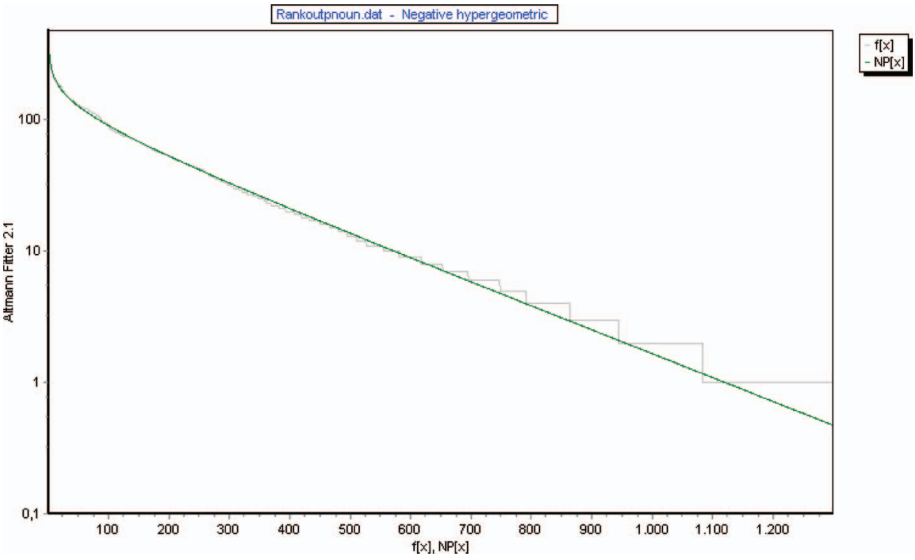

Fig. 4a. Rank-frequencies of outdegrees of Japanese two-kanji compound common nouns.

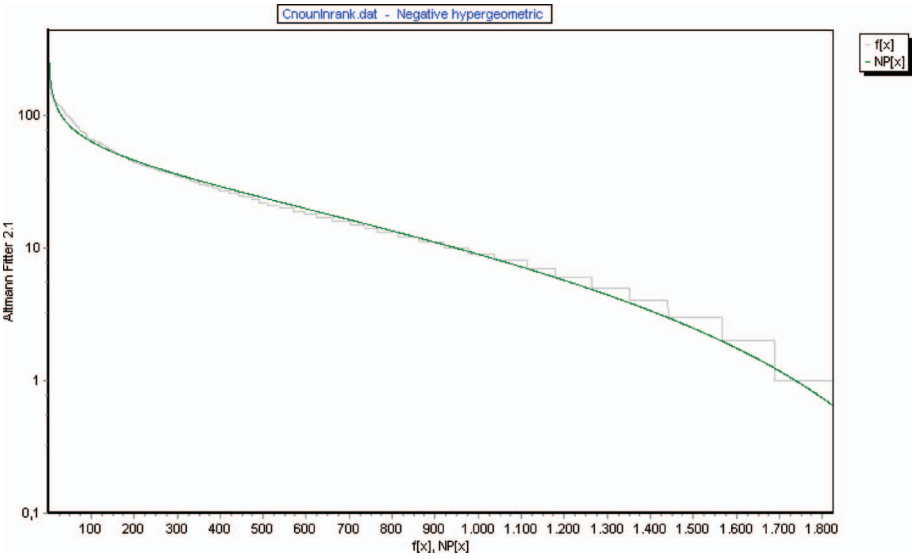

Fig. 4b. Rank-frequencies of indegrees of Japanese two-kanji compound common nouns.

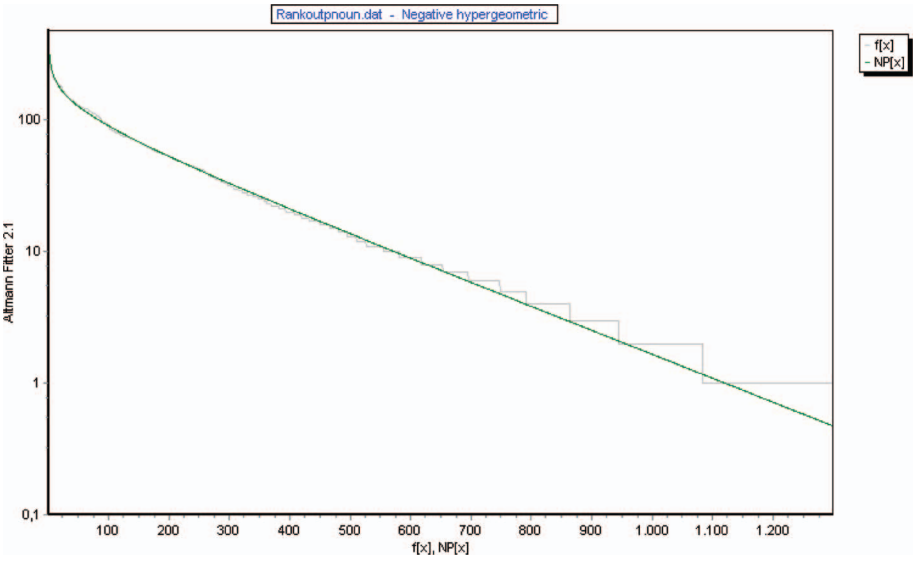

Fig. 4c. Rank-frequencies of outdegrees of Japanese two-kanji compound proper nouns.

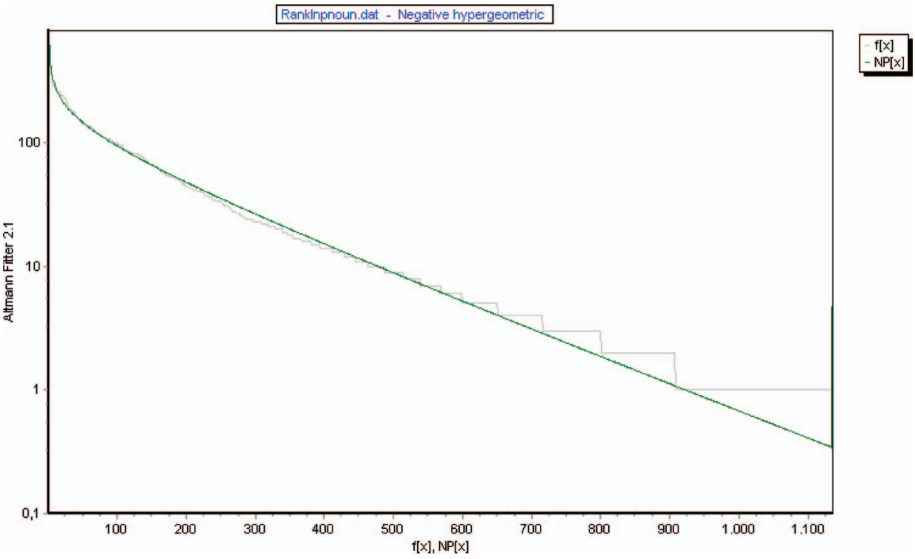

Fig. 4d. Rank-frequencies of indegrees of Japanese two-kanji compound proper nouns.

As can easily be seen, with proper nouns the parameters K and n have a very high value; thus the distribution converges to the negative binomial, which could be used as a limiting alternative. In any case we see that ranks of degrees behave very regularly and have very similar rank-frequency distributions for both proper and common nouns. This fact is also important for the general theory of ranking.

4. THE KIND OF MIXING

In this section, we consider possible combinatory tendencies in the building of kanji compounds. There are three known possibilities from network theory which are adapted here in directed graphs:

(1) Kanji with high outdegree are typically followed by kanji having a high indegree; in this case we speak of assortative compounding (also called assortative mixing in network theory).
(2) Kanji with high outdegree are usually followed by a kanji with low indegree, i.e. a commonly-used kanji prefers rather specialized

kanji as followers; in this case we speak of dissortative compounding.

(3) There are no such tendencies; the combining is random (neutral) (cf. Newman, 2002; 2003).

Since random combining is a logical possibility only with proper nouns, a specific kind of tendency is expected at least with common nouns. In order to ascertain this, we simply compute the linear regression between the outdegrees of kanji and the indegrees of those kanji which can be adjacent.

Since for every outdegree x there are a great number of different indegrees, we either consider their mean and obtain a simple curve, or we consider all values and perform a variance analysis. As a first step we chose to perform a variance analysis since the regression is possibly not linear and the correct curve must be derived theoretically. We remain at the level of an inductive approach. The results are given in Table 3a for proper nouns and in Table 3b for common nouns.

Table 3a. Regression analysis for mixing with proper nouns.

Source of variance	Sum of squares	Degrees of freedom	Mean squares
Slope of line	$0.3609101851(10)^8$	1	$0.3609101851(10)^8$
Variation of true group mean about the line	$0.977406021(10)^7$	153	63882.74647
Within groups	464733515	32839	14151.87780
Total	$0.5105985937(10)^9$	32993	

Table 3b. Regression analysis for mixing with common nouns.

Source of variance	Sum of squares	Degrees of freedom	Mean squares
Slope of line	232747.2877	1	232747.2877
Variation of true group mean about the line	360661.1676	119	3030.7661
Within groups	72801411	35787	2034.2977
Total	$0.7339481946(10)^8$	35907	

In the case of proper nouns, the test of linearity yields 63882.74647/14151.8778 = 4.51 which is distributed as F with 153 and 32839 degrees of freedom, practically $F_{\infty,\infty}$. Since this is greater than the critical value 1 (say at the 0.995 level), the null hypothesis must be rejected. Nevertheless, since classical tests with extremely large samples are problematic, we compute the straight line $y = 140.9687216 - 0.431678027x$. The results of both the F-test and the t-test indicate a highly significant regression coefficient. As it is negative, we can conclude that the relationship is not linear but the compounding is dissortative, that is, the greater the outdegree of a kanji, the smaller the indegree of those kanji which are combined with it to form a proper noun.

In the case of common nouns the linearity test gives us 7661/2034.2977 = 1.4898 which is distributed as $F_{119,\infty}$ and remains at the border of significance (at the 0.995 level). Even if we do not accept linearity, the straight line $y = 52.65144 - 0.0423916x$ signals a highly significant regression coefficient. Thus, also common names exhibit dissortative compounding.

The straight lines for both regressions are presented in Figure 5. Since straight lines are not adequate, we try to discover more realistic curves (dashed line = proper noun; solid line = common noun).

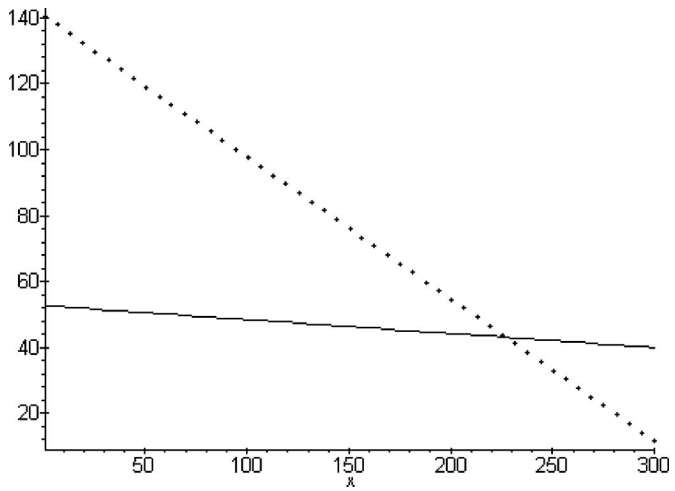

Fig. 5. Linear regression of outdegrees (x) and indegrees (y).

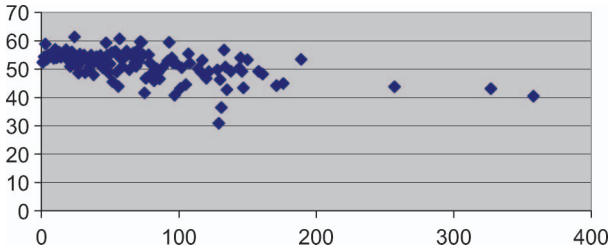

Fig. 5a. Regression for (means of) common nouns.

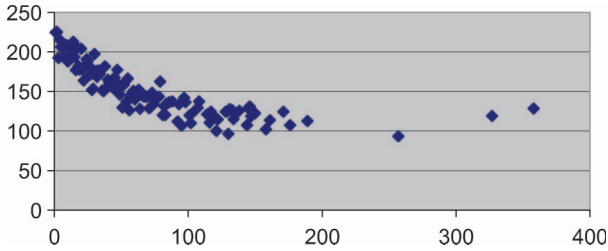

Fig. 5b. Regression for (means of) proper nouns.

To this end we merely compare the x values (outdegrees) with the respective mean indegrees (y) in order to see the course of the dependence (see Figure 5a and Figure 5b).

While a straight line would be visually acceptable for common nouns, proper nouns display a curvilinear trend, indicating that the dissortativity at the beginning, i.e. with smaller outdegrees, holds, but the trend slows down and changes direction with larger outdegrees. In other words, proper nouns beginning with commonly-used kanji preferably combine with commonly used kanji to build proper names.

5. CONCLUSIONS

Building compounds is not a chaotic process. Nevertheless, it is not necessarily straightforward, either. It depends on many factors, such as the character of a language, the requirements of its speakers, external realities, tradition, and the present state of language development. It is a process in which the formation of an individual compound can be a

random product of creativity; but as a whole it is a process displaying features of control. It is rule-governed in several regards, presenting conspicuous regularities, strategies and forms. If we consider this from a purely linguistic point of view, we obtain different classifications of compounds that show the construction techniques preferred or allowed and represent the contents of grammar textbooks. However, if we consider compounding as a mass phenomenon, we can arrive at some latent phenomena that are, perhaps, common to all languages that build compounds. Only a thorough analysis of individual languages can corroborate or weaken the assumption that some of the processes of compound building have a universal character. We are convinced that the processes, distributions, and curves found in Japanese will be different from those in other languages, but at the same time we believe that they boil down to a common principle which must, of course, be derived deductively, even if we have been forced to proceed in an inductive manner as in some cases here. Only future research can show the direction of a possible theory.

REFERENCES

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Review of Modern Physics*, *74*, 47–97.

Batagelj, V., Mrvar, A., & Zaveršnik, M. (2002a). Network analysis of texts. In T. Erjavec & J. Gros (Eds), *Jezikovne tehnologije/Language Technologies* (pp. 143–148). Ljubljana: http://nl.ijs.si/ijst02/zbornik/sdjt02-24batagelj.pdf

Batagelj, V., Mrvar, A., & Zaveršnik, M. (2002b). Network analysis of dictionaries. In T., Erjavec & J. Gros (Eds), *Jezikovne tehnologije/Language Technologies* (pp. 135–142). Ljubljana: http://nl.ijs.si/ijst02/zbornik/sdjt02-24batagelj.pdf

Ferrer i Cancho, R., & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society*, Ser. B, *286*, 2261–2265.

Ferrer i Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(3), 788–791. Retrieved March 14, 2008 from www.pnas.org/cgi/doi/10.1073/pnas.0335980100

Ferrer, I., Cancho, R., Solé, R.V., & Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E*, *69*, 051915 (Santa Fe Institute Working Papers 03-06-092).

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, *89*, 280701 (arXiv: cond-mat/0205405).

Newman, M. E. J. (2003). Mixing patterns in networks. *Physical Review E*, *67*, 026126 (arXiv: cond-mat/0209450).

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78.

Tamaoka, K., & Altmann, G. (2004). Symmetry of Japanese Kanji lexical productivity on the left- and right-hand sides. *Glottometrics*, *7*, 65–64.

Tamaoka, K., & Altmann, G. (2005). Mathematical modeling for Japanese kanji strokes in relation to frequency, asymmetry and readings. *Glottometrics*, *10*, 16–29.

Tamaoka, K., & Makioka, S. (2004). Frequency of occurrence for units of phonemes, morae, and syllables appearing in a lexical corpus of a Japanese newspaper. *Behavior Research Methods, Instruments & Computers*, *36*(3), 531–547.

Tamaoka, K., Kirsner, K., Yanase, Y., Miyaoka, Y., & Kawakami, M. (2002). A Web-accessible database of characteristics of the 1,945 basic Japanese kanji. *Behavior Research Methods, Instruments & Computers*, *34*(2), 260–275.

Zipf, G. K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Psychology*. New York: Houghton Mifflin.