Glottometrics 10, 2005, 16-29

Mathematical Modelling for Japanese Kanji Strokes in Relation to Frequency, Asymmetry and Readings[1]

Katsuo Tamaoka, *Hiroshima, Japan*[2]
Gabriel Altmann, *Lüdenscheid, Germany*

Abstract: The present study investigates the relationship between of Japanese kanji strokes and their printed-frequencies of occurrence, compositional asymmetry and kanji multiple readings. First, distributions of kanji strokes in both samples of the 1,945 basic kanji and of 6,355 kanji appearing in the *Asashi Newspaper* published between 1985 and 1998 followed a negative hypergeometric distribution as demonstrated by Figure 1. The distribution of strokes of the 1,945 kanji with their printed-frequencies is rather rhapsodic, as shown in Figure 2, but a rough-fitting model is drawn in Figure 3. Mathematical modelling for kanji strokes with lexical compositional asymmetry reveals the interesting tendency of *regressive compounding*; that is, that the greater the number of strokes in a kanji, the more it tends to produce two-kanji compound words by adding a kanji on the right side of the target kanji, as shown in Figure 4. A kanji may often have multiple readings; this study also examines the number of readings in relation to the number of kanji strokes. As shown in Figure 6, the greater the number of kanji strokes, the fewer the number of readings. In other words, the more visually complex the kanji is, the more specialised its reading becomes. As such, kanji strokes, as one of the central characteristics of kanji, are closely related to other properties such as frequency, asymmetry and readings. The present study uses mathematical modelling to indicate these relations.

Key words: mathematical modelling, kanji strokes, kanji frequency, lexical compositional asymmetry, multiple readings, regressive compounding, self-regulation cycle

1. A self-regulation cycle of linguistic properties

A single property of a lingustic entity can hardly be isolated from its other properties. The associated properties make up a self-regulating cycle in which one change creates a chain of alterations. The change comes gradually, and its reactions follow with some delay. These series of changes, as a whole, maintain a balanced 'equilibrium' state. Although this conjecture is considered to be applicable to linguistic properties, it is still a preliminary hypothesis which must be validated empirically. The present study on Japanese kanji strokes attempts to explore this conjecture.

A kanji represents the smallest unit of meaning, the 'morpheme'. Unlike alphabetic languages, Japanese kanji are drawn in horizontal and vertical dimensions (for details on kanji in general, see Kaiho & Nomura, 1983; Morton & Sasanuma, 1984; Tamaoka, 1991, 2003, 2005a, 2005b; Tamaoka, Kirsner, Yanase, Miyaoka & Kawakami, 2002; Tamaoka & Makioka, 2004; Tamaoka & Yamada, 2000; Wydell, Butterworth & Patterson, 1995; Wydell, Patterson & Humphreys, 1993). Among the 1,945 basic Japanese kanji, the simplest kanji is

[1] This study is supported by the Research Institute of Science and Technology for Society (RISTEX) at the Japan Science and Technology Agency (JST), Kawaguchi, Japan.
[2] Address correspondence to: Katsuo Tamaoka, International Student Center, Hiroshima University 1-1, 1-chome, Higashihiroshima, Japan 739-8524. E-mail: ktamaoka@hiroshima-u.ac.jp

drawn as a single horizontal line: 一 meaning 'one'; whereas the most complex kanji requires 23 strokes to write: 鑑 meaning 'appreciate'. As such, the number of strokes roughly indicates the visual complexity of a kanji. The question then is whether or not the number of kanji strokes (or kanji visual complexity) has some relation to other properties of a kanji. In the present study, three kanji properties are examined as candidates for possibly being related to the number of kanji strokes; namely, the printed-frequency of a kanji (hereafter, refer to 'kanji frequency'), asymmetry for two-kanji compound word production, and multiple readings.

2. Mathematical modeling for kanji strokes

The lexical database sampled from the *Asahi Newspaper* published between 1985 and 1998 (Amano & Kondo, 2000) contains a total of 6,355 different kanji. Kobayashi (1981) reports that the 1,000 most frequently used kanji represent 93.9 percent of all kanji printed in newspapers, and 90.0 percent of all kanji appearing in magazines. Knowledge of 2,000 kanji covers 99.6 percent of those found in newspapers and 98.6 percent in magazines. Knowledge of 3,000 kanji covers about 99.9 percent of kanji used in newspapers and magazines. Considering these figures, 6,355 kanji must encompass almost all the kanji seen in Japanese written texts. Then, the question is whether the number of kanji strokes has some influence on the kanji frequency.

As can be seen in Table 1, the distribution of the number of strokes is very regular. Using the possibilities furnished by the Wimmer-Altmann theory (for details, see Wimmer, Altmann, 2005) the 1-displaced negative hypergeometric distribution was chosen:

$$(1) \qquad P_x = \frac{\binom{M+x-2}{x-1}\binom{K-M+n-x}{n-x+1}}{\binom{K+n-1}{n}}, \qquad x = 1,2,3,...,n+1.$$

This formula describes the distribution of stroke numbers. Using the kanji database of 1,945 basic kanji (Tamaoka et al., 2002; Tamaoka & Makioka, 2004), the distribution of stroke numbers is obtained as seen with the above 1-displaced negative hypergeometric distribution with parameters K = 11.6517, M = 4.9490, n = 22, DF = 18, X^2 = 22.53, P = 0.21, C = 0.01. If the distribution is stable, it is automatically assumed that a larger sample must display the same behaviour. As can be seen in Table 1, iterative fitting to the inventory of 6,355 kanji shows that the distribution of strokes is, in fact, relatively stable and the fit is satisfactory. As the chi-square grows with increasing sample size, the values obtained for the larger sample were X^2 = 45.63 with 26 DF, corresponding to P = 0.01 and C = 0.0072. The result is displayed graphically in Figure 1. The increasing values of the parameters indicate that perhaps a limiting distribution would yield better results (e.g., Poisson or negative binomial) but an investigation of this issue will be postponed until data from other writing systems such as Chinese, Assyrian, Sumerian, and so on has also been examined. However, the chi-square value decreases for n between 29 and 40, and it increases again above 40. These figures indicate that the negative hypergeometric model is a very stable estimator for this purpose.

There exist different kinds of frequency counts: (a) using individual texts, (b) a list which represents a part of a dictionary, (c) the complete dictionary, (d) a frequency dictionary and a corpus representing a mixture of texts and being identical with a frequency dictionary. Here, we analyze case (b).

Table 1
The negative hypergeometric model for the distribution of strokes in 6,355 kanji

X (Strokes)	f_x (Number of kanji)	NP_x
1	5	3.61
2	29	17.46
3	57	48.11
4	109	98.91
5	146	168.45
6	202	250.76
7	312	336.85
8	436	416.84
9	469	481.90
10	539	525.74
11	579	545.25
12	577	540.53
13	537	514.34
14	432	471.26
15	427	416.76
16	377	356.37
17	300	295.05
18	205	236.71
19	168	184.12
20	136	138.88
21	110	101.57
22	67	72.00
23	47	49.43
24	38	32.84
25	20	21.08
26	12	13.06
27	8	7.79
28	6	4.46
29	3	2.44
30	2	2.43
	K = 24.6877, M = 6.8335, n = 41, DF = 26 X^2 = 45.63 , P = 0.0100, C = 0.0072	

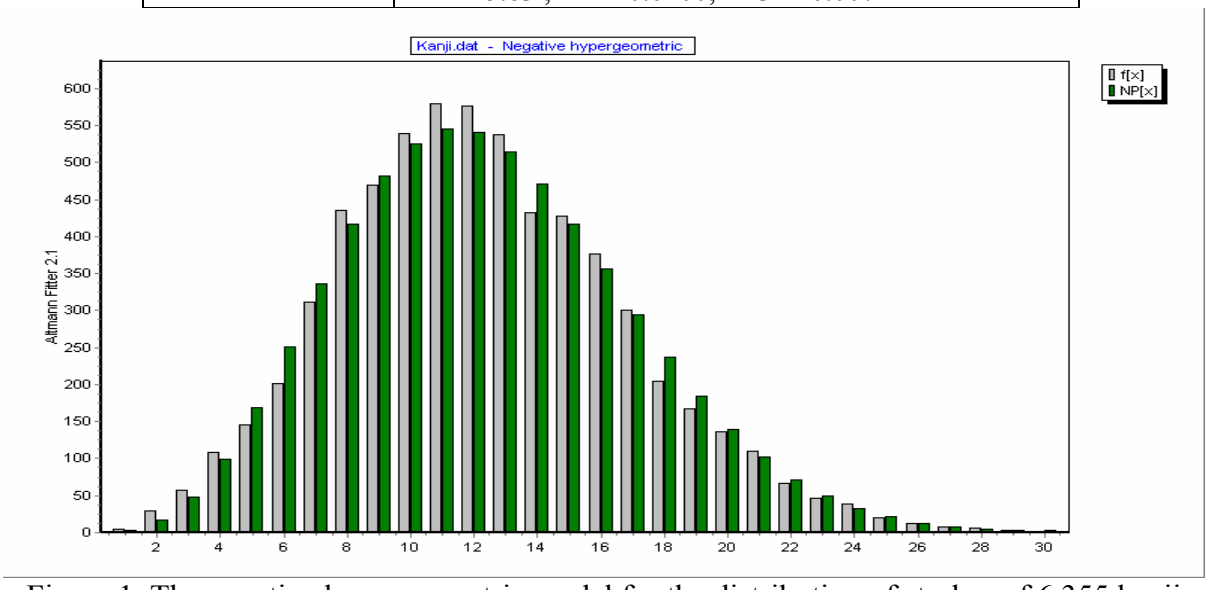

Figure 1. The negative hypergeometric model for the distribution of strokes of 6,355 kanji

3. Modelling the relationship between kanji strokes and kanji frequency

A further question is whether the above model continues to be suitable if the frequencies of the given kanji *in a corpus* are taken into consideration. This question is of a rather methodological nature. The first problem is the sample size. In our case, 6,335 kanji were taken from a very large corpus of 86,542,349 word occurrences where no classical statistical test can rescue the model from rejection. In other words, even if the model could perhaps be accepted, *cum grano salis*, the characteristics of the Chi-square test, destroy our hope. Its linear increase with increasing sample size is fatal for this investigation. The second problem is that language laws hold for homogeneous data. However, a corpus is as heterogeneous as it can be. It contains a mixture of frequencies which, in spite of statistical practice, do not level out with the increasing sample size. On the contrary, the irregularities can escalate. We might erroneously assume that a corpus is a sample from a population, but it has frequently been claimed that there are no populations in language (cf. Orlov, Boroda, Nadarejšvili, 1982). There is no population that can be called "Shakespeare" or "Akutagawa", nor any such population as "the language of the *Asahi Newspaper*" or "the language of the *Times*". Likewise, there is no population that can be called "English texts", "Japanese texts", "the word stock of German" or even "the word stock of Ainu". Great dictionaries contain about one to two hundred thousand lexical items, but the German word stock with all of its special dictionaries (terminology) is estimated to be approximately twenty million. A student of German found in one volume of a German magazine "*Der Spiegel*" more than 8,000 compounds that were not found in any German dictionary. Do they belong to the word stock if they are constructed ad hoc and not in regular use? Is the size of the word stock a real or a potential number? Even though some models yield good results if the lexicon size is considered infinite (cf. Kornai, 2002), the usual models of vocabulary growth are of an exponential type, having finite limits (e.g., Piotrowski, Bektaev, Piotrowskaja, 1985).

As can be seen in Figure 2, the frequency of kanji with a given number of strokes is very rhapsodic. It is multimodal. The main irregularities are in the middle of the distribution. Deviations from any "honest" distribution are so enormous that in this case, no usual model is adequate.

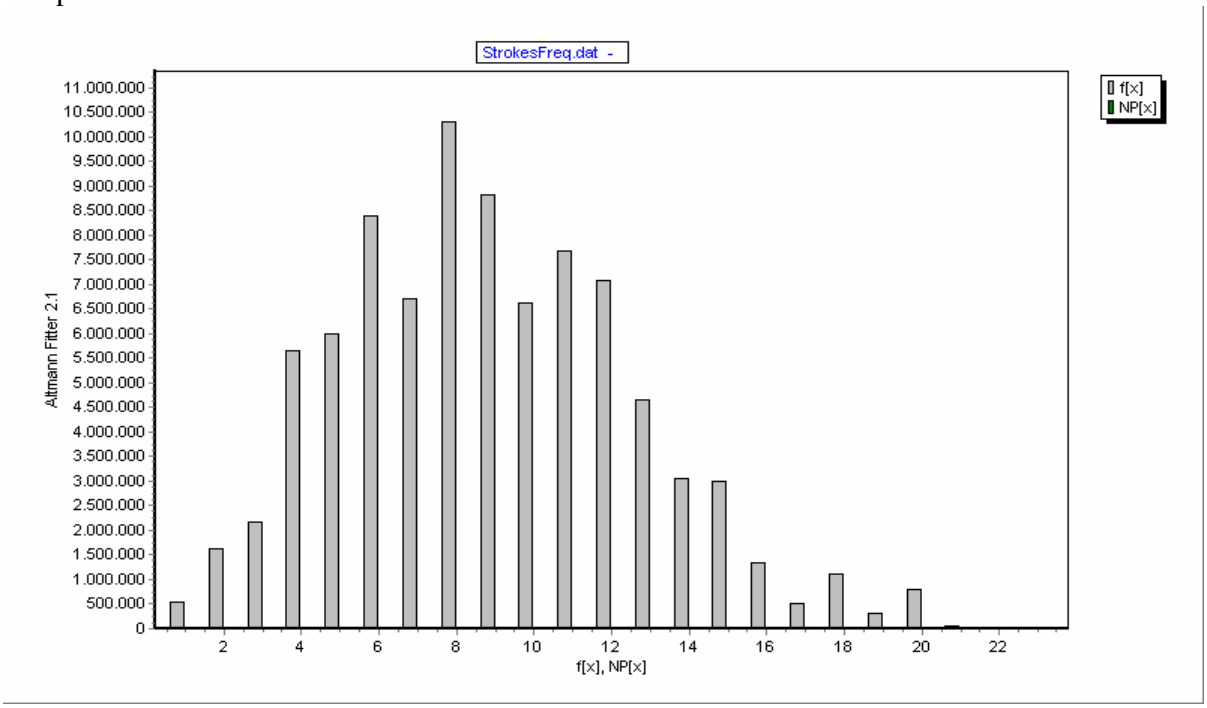

Figure 2. Frequency of kanji with X strokes

Of course, there are different possibilities to capture this irregularity. One of them is the mixing of distributions in different proportions. This technique, however, multiplies the number of parameters and it is not known how many components must be used. The second possibility is smoothing by pooling two or more frequency classes. In this case, the character of a discrete distribution is lost. This is, of course, merely a question of approximation, and in principle it is not relevant what kind of model – discrete or continuous – is used, as the data are distorted. Even if the law behind their generation was known, any model of this law would be forced to take account of many subsidiary conditions, which consequently yield a very special case. Therefore, it is best to consider modelling as a conceptual activity, taking into account simplifications, and even conscious distortions, at the beginning (Bunge, 1967, p. 388).

Table 2 presents the original data corresponding to Figure 2 and the data pooled by 2 using the mean X of the pertinent frequency classes (column 3). It can be seen that in the latter case the "curve" is smooth and can be well approximated by either a continuous distribution or a simple continuous curve.

The second case is preferable on several grounds. First, there is no need to derive a model for false, distorted data – since mixed data are false data. It would be more correct to take each article of the *Asahi Newspaper* separately. Second, the first approximation should be as concise as possible. Third, there is no problem in switching from discrete models to continuous ones or vice versa (cf. Mačutek, Altmann, in press). Since there is no software for a battery of continuous distributions, non-linear regression is simply used to find a preliminary model. Fitting is simpler if the absolute frequencies are transformed into relative ones, as shown in the fifth column of Table 2. Starting from the above mentioned Wimmer-Altmann theory (2001) the simplest case of its continuous version was chosen, namely

$$\frac{dy}{y} = \left(\frac{b}{x} - c\right)dx \tag{2}$$

which says that the relative rate of change of frequency (dy/y) is proportional to the relative rate of change of stroke number (bdx/x) to which a disturbing constant c is added, originating from distortions caused by mixing of texts and pooling of frequency classes. The solution of the differential equation (2) yields

$$y = ax^b e^{-cx} \tag{3}$$

and the results given in the last column of Table 2. As can be seen, the high value of $R = 0.97$ corroborates this kind of approach. Of course, this is merely the first, not the last word.

Based on values in Table 2, formula (3) is calculated as:

$$y = 0.0040x^{3.8342}e^{-0.5058x}$$

This formula is depicted graphically in Figure 3. It shows that after smoothing very heterogeneous data, the relation between stroke number and frequency becomes increasingly rigorous, and almost bell-shaped (with a slight asymmetry).

Table 2
Frequencies of kanji with *x* strokes

X (Stroke)	f_x (Kanji frequency)	x*	f_x*	p^*_x	p_{theor}
1	530246	1.5	2153263	.024881	0.008890
2	1623017				
3	2172880	3.5	7840187	.090594	0.083269
4	5667307				
5	5983444	5.5	14368733	.166031	0.171322
6	8385289				
7	6722157	7.5	17026988	.196747	0.204631
8	10304831				
9	8835897	9.5	15466409	.178715	0.184202
10	6630512				
11	7687320	11.5	14765449	.170615	0.139357
12	7078129				
13	4663567	13.5	7718553	.089188	0.093718
14	3054986				
15	3010830	15.5	4343000	.050183	0.057885
16	1332170				
17	520867	17.5	1647099	.019032	0.033523
18	1126232				
19	321022	19.5	1127029	.013023	0.018460
20	806007				
21	65617	21.5	77071	.000890	0.009761
22	11454				
23	8568	23.0	8568	.000099	0.005920
	86542349		86542349		a = 0.0040 b = 3.8342 c = 0.5058 R = 0.9712

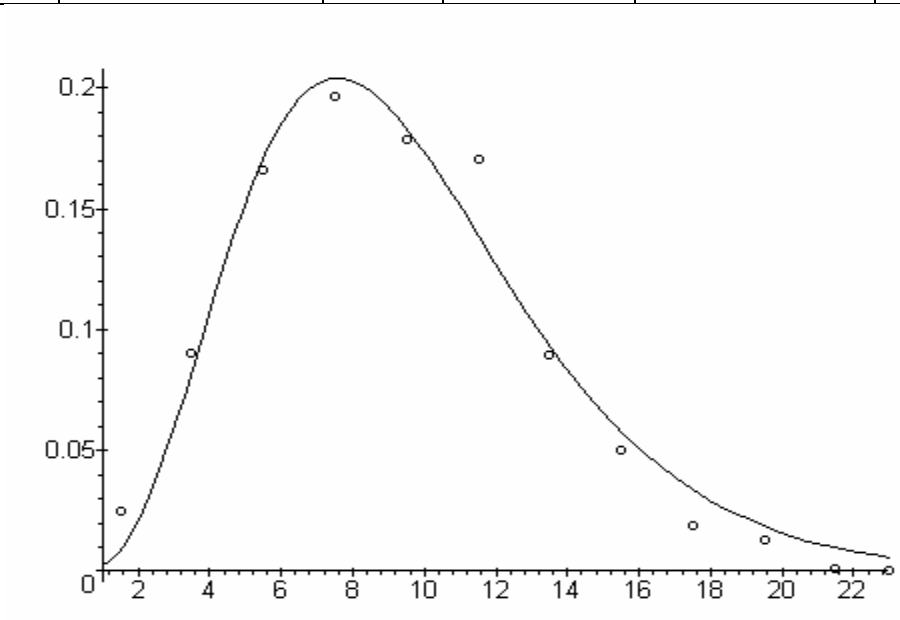

Figure 3. Fitting (3) to the relationship between number of strokes and kanji frequency

5. Modelling the relationship between kanji strokes and lexical compositional symmetry

In a previous study by Tamaoka and Altmann (2004), the symmetry between left-hand and right-hand side compounding has been examined both for individual kanji and for the whole field of 1,945 kanji. For instance, the kanji 学 /gaku/ meaning 'to learn' or 'learning' is combined with another kanji on the right-hand side position such as in 学校 (/gaQ koR/, 'school'), 学生 (/gaku sei/, 'student') and 学者 (/gaku sja/, 'scholar'). Combinations with other kanji on the left-hand side position are also possible such as in 入学 (/njuR gaku/, 'school admission'), 文学 (/bun gaku/, 'literature') and 私学 (/si gaku/, 'private school'). Tamaoka and Altmann investigated two questions regarding how two-kanji compound words were produced from a single kanji by adding another kanji on either the left or the right side, and how symmetric they are on both sides. Furthermore, the present study examines whether the number of strokes has some influence on the symmetry of compound building. Tamaoka and Altmann (2004) tested asymptotically the symmetry using the chi-square criterion

$$X^2 = \frac{(n_L - n_R)^2}{n_L + n_R} \tag{4}$$

which is distributed as a chi-square with 1 degree of freedom. Formula (4) does not show directly whether there is left or right symmetry. Therefore for use as a coefficient of symmetry we express (4) in the form

$$S = \frac{(n_L - n_R)}{\sqrt{n_L + n_R}} \tag{5}$$

indicating left asymmetry if $S > 0$, and right asymmetry if $S < 0$, without evaluating its significance. Of course, any other coefficient that shows the direction of asymmetry and does not allow 0 in the denominator would also be appropriate. Eliminating all cases of kanji having no compounding, we obtained 1,934 cases for which we computed the mean asymmetry at each $x = 1,\ldots,23$. As shown in Figure 4, the results indicated that the greater the number of strokes, the greater the left asymmetry of the kanji.

Table 3
Number of strokes of a kanji and its asymmetry

Number of strokes	Formula (3)	Lin.reg.	Lin.reg.
1	-9.52823359	-2.431076	--------
2	-0.93903791	-2.240516	-0.949802
3	-1.65643326	-2.049956	-0.851436
4	-0.81334384	-1.859395	-0.753069
5	-0.59039237	-1.668835	-0.654703
6	-0.37989045	-1.478275	-0.556337
7	0.02421970	-1.287714	-0.457970
8	0.20483855	-1.097154	-0.359604
9	-0.21886048	-0.906594	-0.261237
10	0.22139170	-0.716034	-0.162871

11	-0.18819254	-0.525473	-0.064504
12	0.05195306	-0.334913	0.033862
13	-0.00252829	-0.144353	0.132229
14	0.28074040	0.046208	0.230595
15	0.28992592	0.236768	0.328962
16	0.02195363	0.427328	0.427328
17	-0.00939287	0.617888	0.525695
18	0.50832419	0.808449	0.624061
19	0.85177389	0.999009	0.722427
20	1.13805872	1.189569	0.820794
21	0.59888953	1.380129	0.919160
22	0.65786894	1.570690	1.017527
23	1.78854380	1.761250	1.115893
		R = 0.3688 a = -2.6216 b = 0.1906	R = 0.7611 a = -1.1465 b = 0.0984

The values (number of strokes vs. formula (3)) are given in Table 3 (first three columns). Again, $x = 1$ displays a fully anomal value and can be left out from further computation. There are only two kanji with one stroke, 一 and 乙. The kanji 乙 is rather uncommon, occurring only 418 times in the *Asahi Newspaper* corpus of 287,792,797 words (Amano & Kondo, 2000). This kanji only produced 11 two-kanji compound word-types with a joint token frequency of 418 (this happened to be the same as the single kanji frequency); by contrast, the kanji 一 produces 381 two-kanji compound word-types with a joint token frequency of 529,828 (Tamaoka & Makioka, 2004). This discrepancy of the two kanji with one stroke created a deviated value of -9.52823359 in the formula (3). Excluding the case of these two one-stroke kanji, we obtain a rather clearer picture in the fourth column of Table 3 and Figure 4. The *F*-test for linear regression is highly significant in both cases (with or without $x = 1$), but we prefer the determination coefficient even if it is not so high ($R = 0.76$).

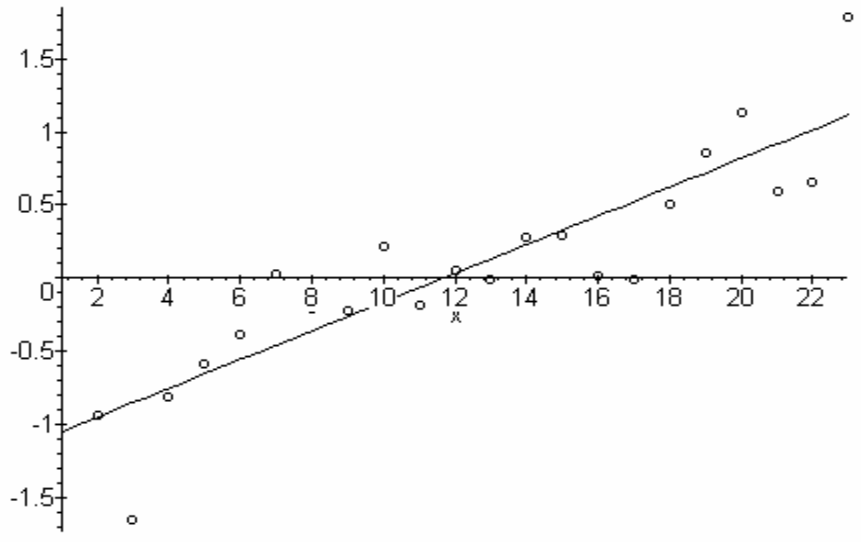

Figure 4. Kanji strokes vs. asymmetry (without $x = 1$)

This result shows that the greater the number of strokes in a given target kanji, the more it tends to produce two-kanji compound words by adding a kanji on the right side (with the target kanji positioned on the left side). This tendency was called *regressive compounding* in the kanji symmetry study of Tamaoka and Altmann (2004). In the case of a kanji with two strokes, the kanji 大 meaning 'big' appears 503,023 times in the *Asahi Newspaper* (Amano & Kondo, 2000). When this kanji is positioned on the right side, 399 two-kanji compound words are produced by adding other kanji on the left side, e.g. 巨大 meaning 'huge'. The same kanji produces 70 different two-kanji compound words by adding other kanji on the right side, e.g. 大学 meaning 'a university'. Roughly five times more compounds were produced when the target kanji is positioned on the right side (adding kanji on the left side) than on the left side (adding kanji on the right side). The token frequencies of these compounds show a similar difference; left-side compounds occur 374,376 times and right-side compounds 128,647 times.

In the case of a kanji with many strokes, this tendency reverses. For example, the kanji 警 meaning 'warn' or 'admonish' appears 91,377 times in the *Asahi Newspaper*. When this kanji is positioned on the right side, 10 different two-kanji compound words are produced by adding other kanji on the left side; for instance 夜警 meaning 'night watch'. The same kanji produces 27 different two-kanji compound words by adding other kanji on the right side, for example 襲来 meaning 'invasion'. The number of compounds for the left side position of the target kanji (adding kanji on the right side) was 2.7 times greater than for the right side position of the target kanji (adding kanji on the left side). The difference in token frequencies of these compounds is not so great but still maintains a similar trend: left-side compounds occur 66,454 times and right-side compounds occur 91,377 times.

Although this overall tendency of *regressive compounding* between kanji strokes and kanji compositional or compounding asymmetry is clearly shown in Figure 4, it is a surprising discovery which lacks a clear explanation.

5. Mathematical modelling for kanji strokes with multiple readings

Kanji pronunciations can be divided into two types: the On-reading derived from the original Chinese pronunciation, and the Kun-reading originating from the Japanese way of reading kanji (for details see Kaiho & Nomura, 1983; Morton & Sasanuma, 1984; Tamaoka, 1991, 2003, 2005a, 2005b; Wydell, Butterworth & Patterson, 1995; Wydell, Patterson & Humphreys, 1993). The mixture of material with two phonological origins created multiple pronunciations of a single Japanese kanji. In dispensing with the Chinese tones, and adapting three different sound systems from China, the Japanese created a great number of multiple readings for the kanji. According to the *Database for the 1,945 Basic Japanese Kanji (2nd edition)* produced by Tamaoka, Kirsner, Yanase, Miyaoka and Kawakami (2002), the total number of kanji which have only one pronunciation is 699 (667 kanji with a single On-reading and 32 kanji with a single Kun-reading). This is 35.94 percent of the 1,945 basic Japanese kanji. Kanji which have only On- or only Kun-readings, regardless of the number of pronunciations, total 779 (739 for On-reading and 40 for Kun-reading), or 40.05 percent of the 1,945 basic Japanese kanji. In other words, although it is a commonly-held notion that a kanji has both an On-reading and a Kun-reading, only 1,166 kanji or about 59.95 percent of the 1,945 basic Japanese kanji have both types of pronunciations.

Starting from the well-known dependence between word length and word polysemy, we conclude that the same relation must hold between the number of strokes (length) of a kanji and the number of readings, which may be an analogue to polysemy. In Table 4 we see the

relevant numbers based on 1,945 kanji (Tamaoka, et al., 2002). The usual relation is here (cf. Köhler 1986); it can be derived from the differential equation (2) setting c = 0 :

(6) $y = ax^{-b}$.

In order to obtain a more solid result, we leave out all mean values based on less than 10 kanji in each case of strokes, i.e. in Table 4, two kanji with 1 stroke, 5 kanji with 21 strokes, 2 kanji with 22 strokes and 1 kanji with 23 strokes. A total of 10 kanji were excluded from the calculation of Formula (6). Iterative computing of (6) yields

$$y = 3.6382x^{-0.2454}$$

giving a preliminarily satisfactory result. Again, we would like to point out that the choice of 1,945 kanji is not random but is based on specific criteria. The graph of observed and computed values is in Figure 6.

Table 4
Dependence of the number of readings on the number of strokes

Number of strokes	Number of kanji	Number of readings (On + Kun)	Mean number of readings	Computed means (4)
1	2	5	2.50	-
2	12	36	3.00	3.07
3	30	95	3.17	2.78
4	66	153	2.32	2.59
5	93	220	2.37	2.45
6	111	258	2.32	2.34
7	141	305	2.16	2.26
8	187	398	2.13	2.18
9	179	379	2.12	2.12
10	199	410	2.06	2.07
11	195	396	2.03	2.02
12	199	431	2.17	1.98
13	147	296	2.01	1.94
14	105	199	1.90	1.90
15	104	180	1.73	1.87
16	68	142	2.09	1.84
17	34	57	1.68	1.82
18	33	64	1.94	1.79
19	21	30	1.43	1.77
20	11	21	1.91	1.74
21	5	9	1.80	-
22	2	5	2.50	-
23	1	1	1.00	-
			a = 3.6382, b = 0.2454, R = 0.81	

In order to show that this is parallel to Köhler's approach, we draw the respective parts of his scheme using our variables *L-Strokes* meaning the logarithm of stroke number and *L-readings* meaning the logarithm of reading numbers. Since the relationship has a decreasing

character, we use a negative proportionality coefficient *–b* and add a factor *A* representing the specification requirement. The scheme is shown in Fig. 5.

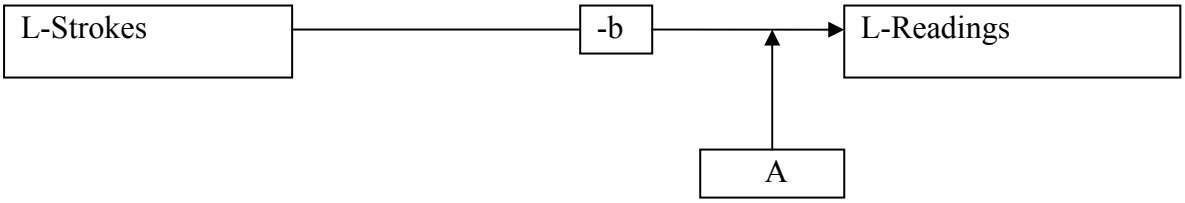

Fig. 5. Systems theoretical scheme of the relationship between the number of strokes and the number of readings of a kanji

The above scheme is analogous to that of the relationship between word length and meaning: in order to specify the meaning, one adds to the word an affix or another word (to build a compound) or reduplicates it. In case of kanji, the reading will be more specific the more complex the kanji is, i.e. the more elementary strokes it contains.
Using the above scheme we have

$$L\text{-}Readings = A - b(L\text{-}Strokes)$$

or $$\log y = A - b \log x.$$

Taking antilogarithms and denoting $e^A = a$, we obtain formula (6).

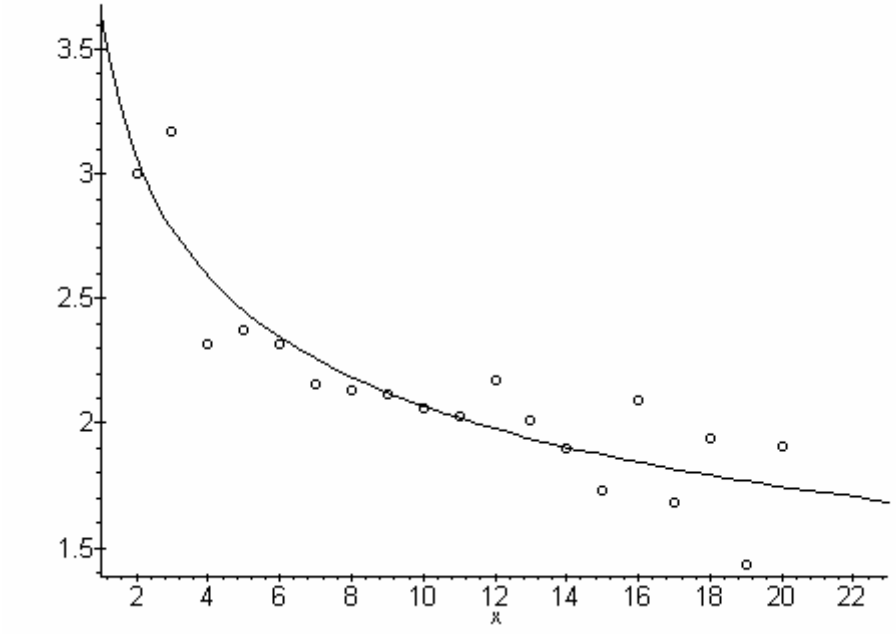

Figure 6. Kanji strokes with mean numbers of kanji readings

Thus, taking pairs of kanji properties and examining their links, we can discover a partial coincidence with Kohler's cycle; but in general we must expect also other types of relationships which are characteristic of kanji only. A more general view of this issue might be achieved by incorporating analyses of Korean and even Assyrian script, both having analogous properties. In Köhler's original self-regulating cycle, only four properties are set in relation: frequency, length, polysemy and polytexty; and all are linked in the way given by

formula (6). With the aid of Köhler's cycle, formulas like (3) and (6) can be set up without the use of differential equations; at the same time they yield a systems theoretical explanation. The formulae express all the requirements imposed on language by its users (cf. Köhler 1986, 1987, 1990a,b,c, 2005). Usually they are captured by the parameters, which must be interpreted; later on their value must be stated *a priori*.

6. Conclusions

It is known that word length is strongly related to frequency in many languages (e.g., Köhler 1986; Breiter 1994; Hammerl 1991; Leopold 2000; Miyajima 1992; Strauß, Grzybek, Altmann 2005). In the case of Japanese kanji, word length can be described as *visual complexity* represented by kanji strokes which require drawing a whole picture of a kanji. The number of kanji strokes is one of a kanji's central properties. This study has investigated the relationships of the number of strokes in Japanese kanji to their printed frequencies of occurrence, compositional asymmetry and kanji multiple readings. Distributions of kanji strokes in *both* samples (the 1,945 basic kanji, and the 6,355 kanji appearing in the *Asashi Newspaper* between 1985 and 1998) produced a negative hypergeometric distribution. The distribution of strokes of the 1,945 kanji with their printed frequencies is rather rhapsodic. Looking at the relationship between kanji strokes and the kanji's lexical compositional asymmetry uncovers the interesting tendency of *regressive compounding*, which fits in with a previous proposal by Tamaoka and Altmann (2004). It has been ascertained that the greater the number of strokes in a kanji, the more strongly it tends to produce two-kanji compound words by adding a kanji on the right side (i.e. with the target kanji positioned on the left side). Furthermore, we have also examined the number of readings of a kanji in relation to the number of kanji strokes. A kanji often has multiple On- and Kun-readings; but we found that the larger the number of kanji strokes, the fewer the number of readings. In other words, the more visually complex the kanji is, the more specialised its reading becomes. As such, the number of strokes, as one of the central characteristics of the kanji, has close relationships with other properties of the kanji: frequency, asymmetry and reading. The present study has demonstrated these relationships. Since the number of kanji strokes can be considered an integral feature of the system of morphemes and lexical properties, we suggest that further investigations into this area will strongly enhance our overall picture of the mutual relationships among different properties of the kanji.

References

Amano, N., & Kondo, K. (2000). *Nihongo-no goi tokusei [Lexical properties of Japanese]*. Tokyo: Sanseido.

Breiter, M. A. (1994). Length of Chinese words in relation to their other systemic features. *Journal of Quantitative Linguistics 1, 224-231.*

Bunge, M. (1967). *Scientific research I.* Berlin: Springer.

Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells.* Trier, WVT.

Kaiho, H., & Nomura, Y. (1983). *Kanji joohoo shori no shinrigaku [Psychology of kanji information processing]*. Tokyo: Kyouiku Shuppan.

Kobayashi, I. (1981). *Kanji kyooiku-no kiso kenkyuu [Fundametal studies on teaching kanji]*. Tokyo: Meiji Tosho.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R. (1987). Systems theoretical linguistics. *Theoretical Linguistics 14, 241-257.*

Köhler, R. (1990a). Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation. *Glottometrika 11, 1-18.*

Köhler, R. (1990b). Zur Charakteristik dynamischer Modelle. *Glottometrika 11, 39-46.*

Köhler, R. (1990c). Elemente der synergetischen Linguistik. *Glottometrika 12, 179-187***.**

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Handbook of Quantitative Linguistics: 760-774*. Berlin: de Gruyter.

Kornai, A. (2002). How many words are there? *Glotometrics 4. 61-86.*

Leopold, E. (2000a). Length-distribution of words with coinciding frequency. In: *Proceedings of the fourth conference of the International Quantitative Linguistic Association, Prague, August 24-26: 76-77.*

Mačutek, J., Altmann, G. (in press). Discrete and continuous modeling in quantitative linguistics. (to appear in Journal of Quantitative Linguistics 2006).

Miyajima, T. (1992). Relationship in the length, age and frequency of Classical Japanese words. *Glottometrika 13, 219-229.*

Morton, J., & Sasanuma, S. (1984). Lexical access in Japanese. In L. Henderson (Ed.), *Orthographies and reading: 25-42*. London: Lawrence Erlbaum Associates.

Orlov, J.K., Boroda. M.G., Nadarejšvili, I.Š. (1982). *Sprache, Text, Kunst. Quantitative Analysen.* Bochum: Brockmeyer.

Piotrowski, R.G., Bektaev, K.B., Piotrowskaja, A.A. (1985). *Mathematische Linguistik*. Bochum: Brockmeyer.

Strauss, U., Grzybek, P., Altmann, G. (2005). Word length and word frequency. In: Grzybek, P. (ed.), *Word length studies and related issues: 255-272*. Boston/Dordrecht: Kluwer.

Tamaoka, K. (1991). Psycholinguistic nature of the Japanese orthography. *Studies in Language and Literature* (Matsuyama University), *11(1), 49-82***.**

Tamaoka, K. (2003). Where do statistically-derived indicators and human strategies meet when identifying On- and Kun-readings of Japanese kanji? *Cognitive Studies*, 10(4), 1-28.

Tamaoka, K. (2005a). Meimei kadai ni oite kanji 1-ji no shoji to on'in no tan'i wa itti suru ka [Is an orthographic unit of a single Japanese kanji equivalent to a kanji phonological unit in the naming task?] *Cognitive Studies*, *12(2), 47-73.*

Tamaoka, K. (2005b). The effect of morphemic homophony on the processing of Japanese two-kanji compound words. *Reading and Writing*, *18, 281-302.*

Tamaoka, K., & Yamada, H. (2000). The effects of stroke order and radicals on the knowledge of Japanese kanji orthography, phonology and semantics. *Psychologia*, *43, 199-210.*

Tamaoka, K., & Altmann, G. (2004). Symmetry of Japanese kanji lexical productivity in the left- and right-hand sides. *Glottometrics*, *7, 65-84.*

Tamaoka, K., & Makioka, S. (2004). New figures for a Web-accessible database of the 1,945 basic Japanese kanji, fourth edition. *Behavior Research Methods, Instruments & Computers*, *36(3), 548-558.*

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.*

Wimmer, G. & Altmann, G. (2001). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Handbook of Quantitative Linguistics: 791-807.* Berlin: de Gruyter.

Wydell, T.N., Patterson, K.E., & Humphreys, G.W. (1993). Phonologically mediated access to meaning for kanji: Is a rows still a rose in Japanese Kanji? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, *491-514*.

Wydell, T. N., Butterworth, B., & Patterson, K. E. (1995). The inconsistency of consistency effects in reading: The case of Japanese kanji. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, *1155-1168*.