On the relation between types and tokens of Japanese morae

Katsuo Tamaoka

1 Type and token frequencies of Japanese mora

The fundamental set of Japanese sounds, which are described in phonological scripts of two types of kana symbols: *hiragana* or *katakana*, consists of 101 basic kana including the special sound of the nasal /ŋ/ (written in hiragana as ん and in katakana as ン). Each Japanese kana symbol fundamentally corresponds to a specific moraic sound. For example, a kana symbol か in hiragana or カ in katakana represents a moraic sound /ka/ consisting of consonant /k/ and vowel /a/. Using the word printed-frequency database created by Amano and Kondo (2000, and 2003 for CD-ROM version)[1], Tamaoka & Makioka (2004) studied the frequency (token frequency) and the context of Japanese morae, consisting of words in which they occur (type frequency).

Since type frequency represents here the property of *cotextuality*, i.e. the number of different contexts of a mora, and token frequency represents the *frequency*, the present study proposes the hypothesis that there is some relationship between these two properties known from synergetic linguistics (cf. Köhler 1986, 2002). Köhler tried to express this hypothesis by means of a straight line and obtained a significant correlation coefficient signaling the existence of a linear relationship. However, the fitted straight line is merely a first exploratory approximation because the lower frequency values of the computed straight line are negative, which is empirically impossible. The present study therefore tries to find a more adequate expression of this re-

1. A very large lexical corpus of 341 771 morphemes was established from newspapers containing 287 792 797 morphemes, all of which were taken from the *Asahi Newspaper* printed from 1985 to 1998. At present, this is the largest and the most up-to-date lexical database created from calculating the word frequency of occurrence in Japanese written texts. There are some coding errors in this database. For instance, Tokyo, written in two kanji as 東京, is phonologically stored as /higasikyoR/ (/R/ refers to a long vowel) instead of /tokyoR/. In addition, a long vowel /R/ is not clearly distinguished in the database. Thus, I did not include the sounds of the long vowel /R/ and the geminate sound /Q/ in this study.

lationship.

The present study starts from the general theory of language laws (Wimmer & Altmann, 2005) stating that the majority of relationships between linguistic variables can be expressed by means of the relation between their relative rates of change, i.e. between dy/y and dx/x. Here y is frequency (tokens), x is cotextuality (types). This assumption is reasonable because linguistic variables cannot attain infinite sizes and their changes tend to slow down depending on the already attained value. Thus we postulate a very simple relation

$$\frac{dy}{y} = \left(c + \frac{b}{x}\right) dx \tag{1}$$

where b is the proportionality coefficient and c is interpreted either as a kind of "noise" or disturbance, or it represents the *ceteris paribus* condition, i.e. a constant state of all other possible factors. The solution of this equation is

$$y = ax^b e^{cx} \tag{2}$$

where $a = e^C$ with integration constant C.

This relationship follows directly from the respective part of the synergetic control cycle joining polytexty with frequency. Since in synergetic linguistics one considers the variables usually in their logarithmically transformed form, we speak about *L*-polytextuality and *L*-frequency. In the cycle presented below (Köhler, private communication, cf. Köhler 1986; 2002) one can see that polytextuality affects frequency proportionally, with an additive constant coefficient of usage requirement *L-Appl* ("communicative relevancy" of the given meaning); at the same time it affects the frequency by means of an exponential operator with a constant C, which is assumed to represent the mean sign complexity (or length). The cycle is a new variant in Köhler's theory, allowing a non-logarithmic, "plain" effect of one variable on another, using the exponential operator.

Thus, the formula becomes

$$\ln(F) = R\ln(Appl) + B\ln(PT) - C\exp(\ln(PT)),$$

i.e.

$$\ln(F) = R\ln(Appl) + B\ln(PT) - C(PT)$$

from which it follows that

$$F = Appl^R PT^B e^{-C(PT)}.$$

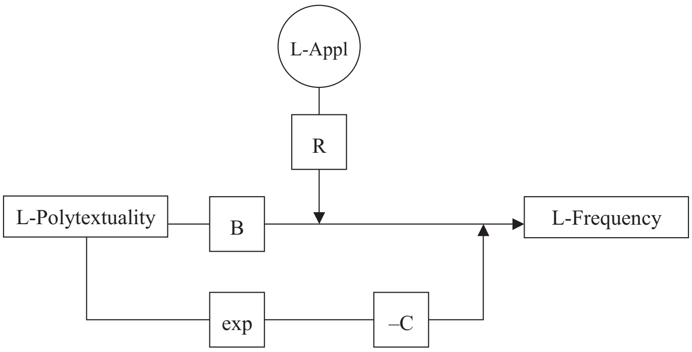

Figure 1: The relationship between polytextuality and frequency in general

Since $Appl^R$ can, in the framework of a synchronic study, be considered a constant, for example, A, while $PT = x$, and $F = y$, we can obtain

$$y = Ax^b e^{-cx}$$

which is identical to the above solution of the differential equation. Iterative computing of this curve yielded

$$y = 26.57366832x^{1.31502554}\exp(-0.0000125937521x)\,. \qquad (3)$$

The determination coefficient is $R^2 = 0.93$, indicating that relationship (2) can be preliminarily accepted. The observed and the computed values are shown in Table 1 . The computation has been performed in logarithmic transformation, thus R^2 holds for the logarithmic values in the given interval. Though in many cases the deviation is considerable, the variability of token frequencies is so enormous that it renders R^2 very high and the F-test very significant.

2 Conclusion

The problem of type and token frequencies can in this case be reduced to that of cotextuality and frequency, a relationship that has been very lightly studied in quantitative linguistics. In any case, the result confirms Köhler's

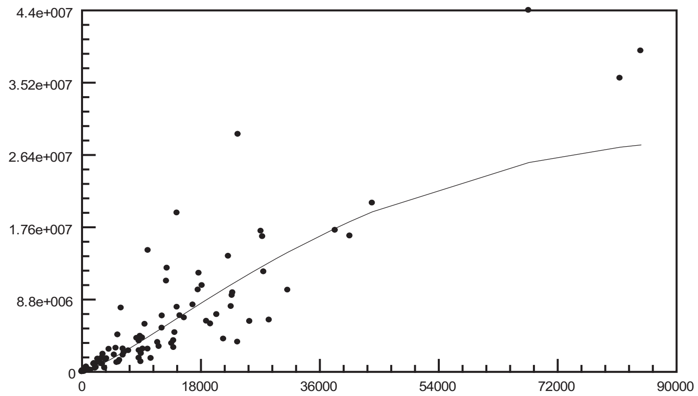

Figure 2: Relation between types and tokens of Japanese morae

assumption. In the majority of linguistic cases, frequency is considered as an independent variable but in the case of Japanese morae it is exactly the other way round. This can be explained by the fact that frequency, as most of the other linguistic quantities in the synergetic control cycle, has direct as well as indirect effects on other variables and is, at the same time, a dependent variable, e.g. depending on polytextuality.

References

Amano, Shigeaki; Kondo, Tagahisa

2000 *Nihongo-no goi tokusei. Dai-7-kan.* [Lexical properties of Japanese. Volume 7.] Tokyo: Sanseido.

2003 *Nihongo-no goi tokusei. Dai-2-ki CD-ROM-ban.* [Lexical properties of Japanese. The second volume of CD-ROM version.] Tokyo: Sanseido.

Ferrer i Cancho, Ramon; Solé, Richard V.; Köhler, Reinhard

2004 "Patterns in syntactic dependency networks." In: *Physical Review*, E 69, 051915.

Köhler, Reinhard

2002 *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik.* [`http://ubt.opus.hbz-nrw.de/volltexte/2004/`]

2005 “Synergetic linguistics”. In: Köhler, Reinhard; Altmann, Gabriel; Piotrowski, Rajmund G. (Eds.), *Quantitative Linguistics. An International Handbook*. Berlin: de Gruyter, 760–774.

Tamaoka, Katsuo; Makioka, Shogo
2004 “Frequency of occurrence for units of phonemes, morae, and syllables appearing in a lexical corpus of a Japanese newspaper.” In: *Behavior Research Methods, Instruments & Computers*, 36(3); 531–547.

Wimmer, Gejza; Altmann, Gabriel
2005 “Unified derivation of some linguistic laws”. In: Köhler, Reinhard; Altmann, Gabriel; Piotrowski, Rajmund G. (Eds.), *Quantitative Linguistics. An International Handbook*. Berlin: de Gruyter, 791–807.

Table 1: Types (cotextuality) and tokens (frequency) of Japanese morae in a corpus

#Types	#Tokens	(3)	#Types	#Tokens	(3)
9	95	477.81	8688	3740881	3603412.66
45	511	3964.81	8694	2577386	3606409.18
69	3304	6953.57	8857	4303939	3688009.24
69	3087	6953.57	8951	1210586	3735140.99
82	125380	8724.03	9007	2206487	3763249.39
83	16016	8864.09	9174	4118162	3847166.66
113	14646	13294.87	9263	2758994	3891956.70
117	15343	13916.47	9566	5777889	4044762.26
147	29907	18781.27	10011	2749654	4269991.40
170	53284	22731.02	10038	14757618	4283689.66
175	10406	23612.77	10491	1620066	4513869.51
183	50769	25039.87	11507	3554252	5032502.51
201	15494	28321.35	11707	3067681	5134888.29
207	59424	29436.08	12158	5306817	5366028.81
400	33782	69829.51	12180	6791877	5377309.08
504	162751	94506.13	12822	11037767	5706765.14
551	166750	106199.38	12919	12594714	5756559.53
700	577939	145211.16	13622	3432278	6117584.83
738	446886	155590.11	13922	3777201	6271623.81
1122	214108	268616.69	13925	2927552	6273169.18
1165	182551	282083.06	14103	4767153	6364539.70
1166	188263	282397.6	14413	19326674	6523625.18
1216	199251	298240.76	14435	7858931	6534915.86
1343	192847	339317.3	14873	6823149	6759522.22
1814	929206	500865.36	15529	6550231	7095418.24
1898	995900	531023.82	16836	8127590	7762131.57
2034	411356	580621.35	17622	9953427	8160955.32
2084	937044	599085.76	17737	11996376	8219151.77
2183	507561	635993.97	18207	10498607	8456526.23
2442	1541224	734627.19	18898	6149909	8804141.22
2633	1254559	809153.65	19482	5808199	9096493.22
2937	984460	930614.9	20442	6947107	9574022.17
3024	1646454	965975.25	21459	3972466	10075369.91
3125	936779	1007342.14	22195	14053377	10435061.19
3208	2138138	1041580.88	22624	7932086	10643439.60
3480	445557	1155274	22763	9271734	10710752.90
3623	1482387	1215911.6	22853	9612206	10754271.59
3769	1591698	1278397.76	23598	3607169	11112889.60
4139	2721853	1439209.44	23656	28899891	11140687.70
4929	2041928	1792936.88	25431	6116529	11981880.94
5175	2858438	1905609.9	27131	17102180	12769738.39
5317	1104998	1971142.4	27384	16442465	12885432.49
5464	4484240	2039337.01	27551	12154255	12961581.65
5569	1177824	2088262.07	28364	6289027	13329688.48
5736	1384993	2166436.68	31183	9926212	14571893.89
5965	7744822	2274317.01	38359	17211261	17480465.76
6249	2009223	2409134.24	40585	16518016	18306022.84
6260	2771906	2414376.53	43982	20530065	19494944.36
6422	2328945	2491787.48	67643	43985426	25488953.21
7088	2549499	2813332.09	81463	35719268	27348656.72
8336	4064902	3427819.93	84624	39052254	27630516.29
8670	1644763	3594413.22			