Word length distribution of Japanese dialects

Wenchao Li
School of International Studies Zijingang Campus,China

Citation: Wenchao Li (2022) Word length distribution of Japanese dialects, *International Journal of Quantitative and Qualitative Research Methods*, Vol.9, No.3, pp.7-16

ABSTRACT: *This study applies a mathematical linguistic approach to explore word length distribution of Japanese dialects to cluster dialects at a lexical level. Data were extracted from spoken recordings of native speakers from 47 areas. The findings revealed that the further south the area was, the longer the mean word length (MWL) became. In majority of dialects, MWL ranges from one to nine. The Saga dialect has the longest MWL (3.26). Further analysis of the MWL-frequency relationship via the Altmann-fitter reveals that MWL-frequency of all dialects fit more than 30 distribution models, including the binomial and Poisson families.*

KEYWORDS: Japanese, word length, dialect, distribution, power law function

INTRODUCTION

It has been suggested that word length is associated with stylometrics (Mendenhall 1887, 1901; Williams 1975), usage frequency (Zipf 1949), word age, ambiguity, and language acquisition (Miyajima 1990; Sanada 1997; Ishii 1990; Ogino 1980; Minami et al. 2013). This study applies the mathematical linguistic approach to investigate the word length of 47 dialects in Japan to understand (i) whether word length is sensitive to the diversity of dialects (ii) how the distribution of length follows a specific regularity (iii) and what parameters indicate a trend.

Background

Japan has 43 standard prefectures, two urban prefectures (Osaka and Kyoto), one metropolis prefecture (Tokyo), and one circuit prefecture (Hokkaido). Numerous studies have investigated dialects, focusing on the variations in accent (Hattori 1931; Hirayama 1936; Kindaichi 1937), lexicon (Kushihiki 2008), case system (Nitta 1992), and conjunction (Toyota 1985). However, there is room for further exploration, such as the diversity of lexical complexity among dialects with dynamic data drawn from corpus from a mathematical linguistic approach. The rest of this paper is organised as follows. The methodology (including the data, measurement of word length) is outlined in section 3. The results and discussion are provided in section 4. Finally, concluding remarks are presented in section 5.

DATA AND METHODS

Corpus

This study explored the sensitivity of Japanese word length to the diversity. To this end, a self-built database comprising a dynamic spoken dialect corpus, including 4,000 hours of recordings from 47 areas, was examined. Details of the materials used are provided in Table 1.

Table 1. Materials

Dialects	Recorded year	Recording length	Male speaker	Female speaker
Hokkaidoo	1978	0:37:16	1	2
Aomori	1979	1:13:38	1	2
Iwate	1980	0:46:59	1	1
Miyagi	1977	1:21:42	6	3
Akita	1977	0:25:52	1	2
Yamagata	1980	0:54:39	1	2
Fukushima	unknown	0:59:52	1	2
Ibaraki	1982	1:04:07	5	2
Tochigi	1979	0:34:44	2	2
Gunma	1983	1:23:51	3	2
Saitama	1981	0:37:57	2	1
Chiba	1978	0:41:27	1	2
Tokyo	1982	1:19:34	3	4
Kanakawa	1983	0:51:55	6	2
Niigata	1980	0:36:44	4	1
Toyama	1981	0:34:57	4	4
Ishikawa	1977	0:21:31	2	1
Fukui	1983	1:12:19	4	6
Yamanashi	1978	0:26:37	3	2

Nagano	1978	1:01:17	4	2
Gifu	1979	0:46:30	7	3
Shizuoka	1979	0:23:48	3	1
Aichi	1981	1:28:24	6	3
Mie	1981	1:05:59	11	2
Shiga	1981	0:39:33	4	1
Kyoto	1984	1:40:44	7	14
Oosaka	1977	0:44:58	10	15
Hyoogo	unknown	1:28:32	9	12
Nara	1981	1:47:25	51	27
Wakayama	1981	0:31:43	2	2
tottori	1984	0:05:28	1	1
Shimane	1980	0:35:19	1	2
Okayama	1979	0:27:42	1	1
Hiroshima	1977	0:38:13	2	4
Yamaguchi	1978	0:37:17	4	2
Tokushima	1981	0:36:50	2	2
Kakkawa	1978	0:36:05	1	2
Ahime	1981	0:32:56	1	1
Koochi	1977	0:33:47	1	1
Fukuoka	1981	1:38:08	16	8
Saga	1978	0:20:58	1	2
Nagasaki	1983	0:24:36	1	1
Kumamoto	1980	2:01:02	32	18
Ooita	1978	0:40:35	2	2
Miyazaki	1981	0:22:39	1	1
Kagoshima	1977	0:34:37	2	1
Okinawa	1978	0:58:52	1	6

Procedures

A computer programme is created for the calculation and fitting to models. The following procedures are carried out.

Step 1: Obtain raw data from the dialect corpus.
Step 2: Parse each sentence via the GiNZA v4 Parser (NINJAL and Megagon Labs).
Step 3: Romanise the Japanese scripts using a python programme.

Step 4: Calculate the dynamic mean word length distance from the parsed outputs based on syllable unit.

The associations between word length and dialects were determined via Euclidean distance. Taking D_1 and D_2 as vectors representing the compared dialects, the distance between L_1 ($D_{1,1}, \ldots, D_{1,n}$) and D_2 ($D_{2,1}, \ldots, D_{2,n}$) was calculated using the following formula:

$$d(D_{1,}D_2\ ) = \sqrt{(D_{1,1} - D_{2,1})^2 + (D_{1,2} - D_{2,2})^2 + \cdots + (D_{1,n} - D_{2,n})^2} = \sqrt{\sum_{i=1}^{n}(D_{1,n}\ - D_{2,n})^2)}$$

RESULTS AND DISCUSSION

This study analysed the data drawing on the aforementioned methodology. An overview of word length in the 47 areas is presented in section 4.1. The probability distribution and the parameters that may indicate a trend in the diversity of word length are discussed in section 4.2.

Word length in Japanese dialects

The MWL of each dialect is reported in Table 2.

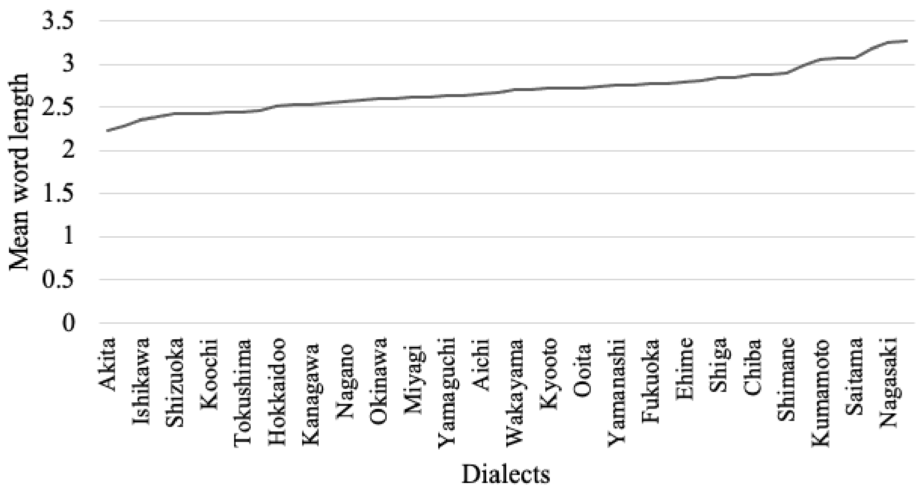
Table 2. Mean word length of different dialects in Japan

Dialects	Mean word length	Dialects	Mean word length	Dialects	Mean word length
Akita	2.24	Yamaguchi	2.63	Miyagi	2.62
Iwate	2.29	Hyoogo	2.64	Tochigi	2.62
Ishikawa	2.36	Aichi	2.65	Gifu	2.99
Kagawa	2.39	Okayama	2.67	Kumamoto	3.05
Shizuoka	2.43	Wakayama	2.71	Saitama	3.08
Fukui	2.43	Tottori	2.71	Oosaka	3.18
Koochi	2.43	Kyooto	2.72	Chiba	2.88
Ibaraki	2.44	Yamagata	2.73	Miyazaki	2.89
Tokushima	2.44	Ooita	2.73	Shimane	2.90
Niigata	2.46	Kagoshima	2.74	Gunma	3.07
Hokkaidoo	2.52	Yamanashi	2.76	Nagasaki	3.25
Toyama	2.53	Hiroshima	2.76	Saga	3.26

Kanagawa	2.53	Fukuoka	2.77	Nara	2.85
Fukushima	2.54	Aomori	2.78	Shiga	2.84
Nagano	2.57	Ehime	2.80	Okinawa	2.60
Mie	2.59	Tokyo	2.82		

The trend of MWL is illustrated in Figure 1.

Figure 1. Mean word length in Japanese dialects



(1) provides the distribution of word length in the dialects, which suggests that word length ranges from 1-9 in 42.55% of dialects, from 1-10 in 21.27%, and 1-8 in 17.02%. Additionally, there are five tokens ranging from 1-11 and two from 1-12.

Dialect groups based upon the distribution of word length

(a) Word length ranges from 1-8: Niigata, Kagoshima, Mie, Kochi, Okinawa, Okayama, Ehime, Shizuoka, Tottori, Fukuoka, Hokkaido.

(b) From 1-9: Akita, Yamagata, Saitama, Kagawa, Hiroshima, Kyoto, Miyagi, Iwate, Ibaraki, Aichi, Kanagawa, Aomori, Ishikawa, Chiba, Osaka, Oita, Nagano, Tokyo, Nara, Toyama.

(c) From 1-10: Yamanashi, Kumamoto, Yamazaki, Gifu, Saga, Miyazaki, Shimane, Tokushima, Tochigi, Yamaguchi.

(d) From 1-11: Shiga, Gunma, Fukui, Fukushima, Hyogo.

(e) From 1-12: Nagasaki, Wakayama.

Essentially, the word length distribution-based classification is in accordance with MWL-based classification. As shown in (3), the Saga dialect has the longest MWL (3.26), Nagasaki the second (3.25), and Saitama the third (3.08). We have three groups of dialects based on MWL: (i) long MWL group of dialects with MWL extending to 3, including Saga, Nagasaki, Saitama, Gunma, and Kumamoto, (ii) middle MWL group of dialects with MWL between 2.3 and 3.0, and (iii) the short MWL group of dialects with MWL less than 2.49, cf. (2), including Gifu and Kagoshima.

Dialect groups based on MWL
(a) Group with long MWL (longer than 3): Saga, Nagasaki, Saitama, Gunma, Kumamoto.
(b) Group with middle MWL (between 2.5-3.0): Gifu, Kagoshima, Shiga, Yamanashi, Kanagawa, Yamagata, Mie, Hiroshima, Kyoto, Miyagi, Miyazaki, Okinawa, Okayama, Ehime, Aichi, Aomori, Chiba, Osaka, Oita, Nagano, Tottori, Shimane, Tokyo, Tochigi, Nara, Toyama, Fukuoka, Fukushima, Hyogo, Wakayama, Yamaguchi.
(c) Group with short MWL (between 2.0-2.49): Niigata, Akita, Kochi, Kagawa, Iwate, Ibaraki, Shizuoka, Ishikawa, Tokushima, Fukui, Hokkaido.

Euclidean distance-based clustering was carried out and the findings are illustrated in Figure 2.

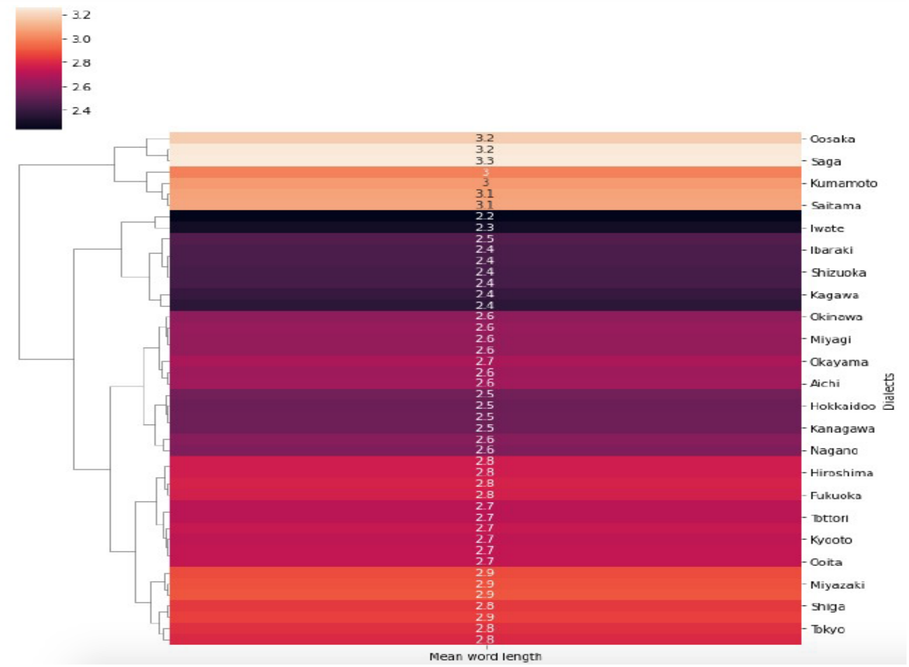


Figure 2. Euclidean distance-based clustering of Japanese dialects

Figure 3 is a self-drawn map, clustering dialect groups following MWL. The long, middle, and short MWL groups are indicated by blue, red, and yellow, respectively.

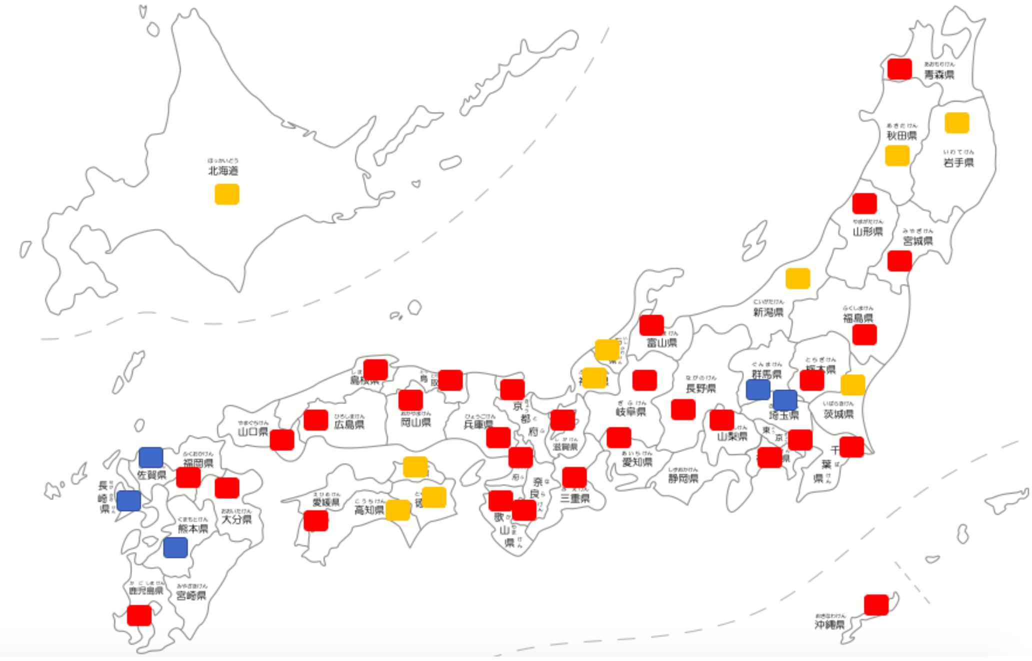


Figure 3. Clustering of Japanese dialects based on MWL

Probability distribution of word length

This section explores whether the distribution of word length followed a specific regularity. A computer program was produced to fit the power law function ($y=ax^b$). The fittings of the data showed that MWL and frequency could be demonstrated by the power law function with acceptable results, with the highest value of the determination coefficient R^2 of 0.8984 ($R^2 > 0.90$, very good; $R^2 > 0.80$, good; $R^2 > 0.75$, acceptable; $R^2 < 0.75$, unacceptable). The frequency of MWL being 1 was lower than that of MWL, which was 2. This study selected fixed fitting results, Aichi, Akita, Ehime, Ishikawa, Kyoto, and Saga dialects (see Figure 4).

ISSN 2056-3620(Print)
ISSN 2056-3639(Online)

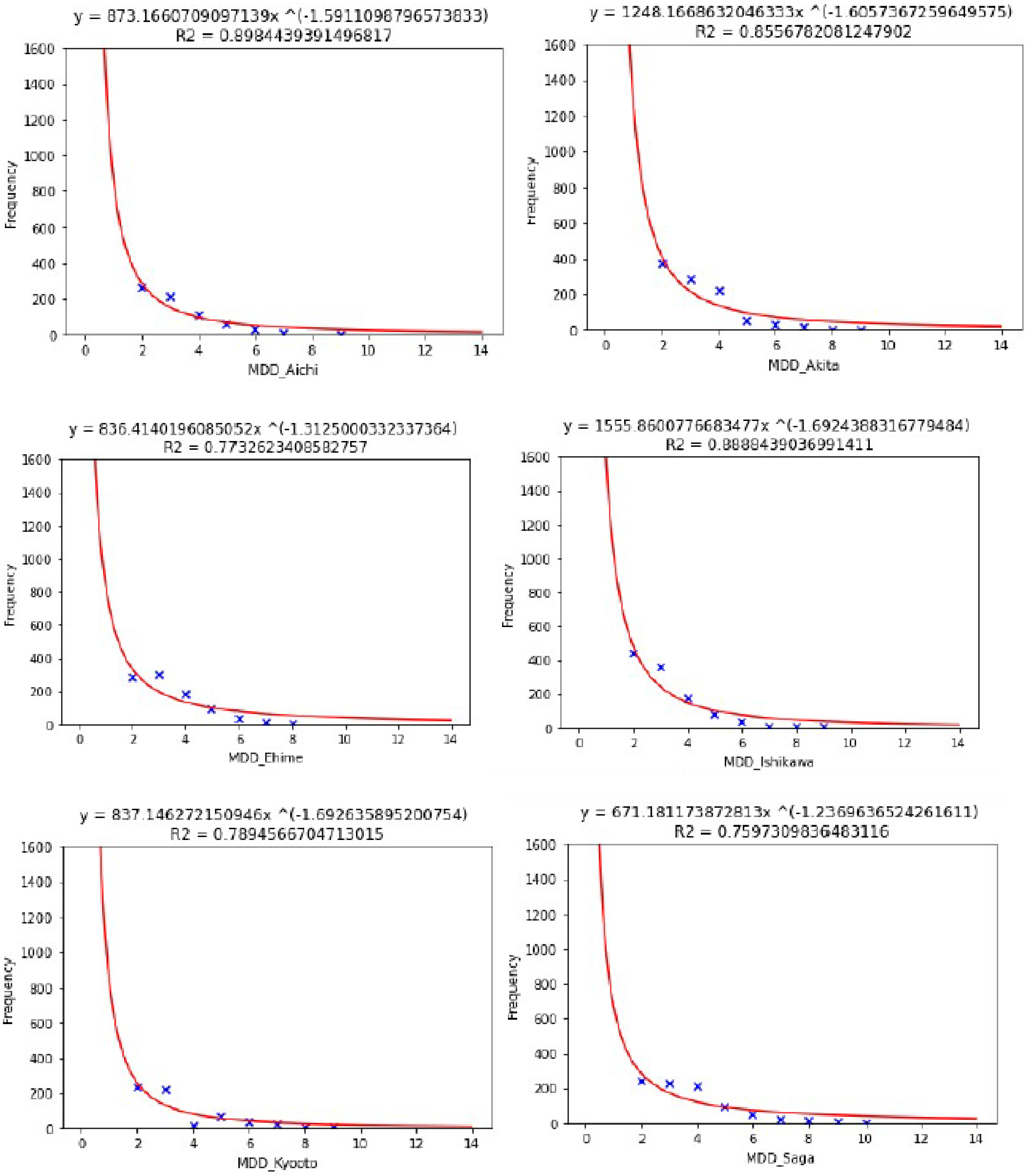
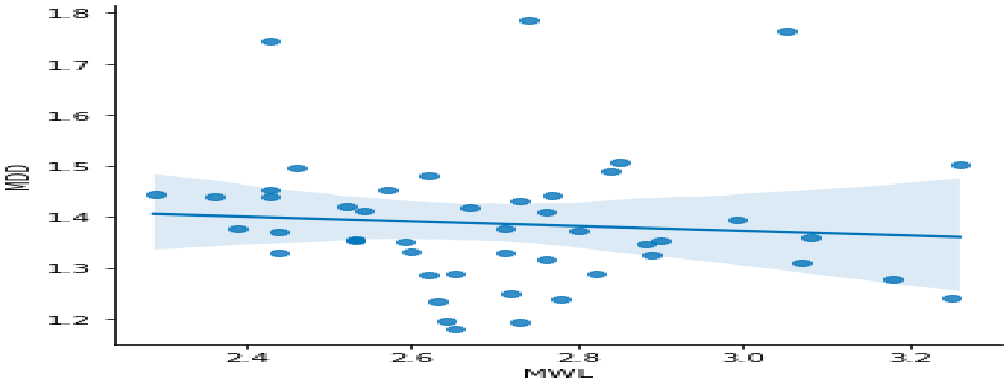


Figure 4. Fitting results of power law function to word length and frequency

Analysis is followed by MWL-frequency relationship via the Altmann-fitter. The findings indicated that MWL_frequency of all the dialects fit more than 30 distribution models, including the binomial and Poisson families. Furthermore, Li (under review) calculated the mean dependency distance (MDD) of the 47 dialects. The Spearman correlation coefficient was used to examine the relationship between MWL and MDD. The findings are presented in Figure 5. It could not be confirmed if with an increase (decrease) of MWL, MDD increased (decreased).



The results revealed a diversity of word length in the 47 dialects. A slight trend suggested the further south the region was, the longer the MWL was. A Euclidean distance-based clustering revealed three groups of dialects in terms of MWL. The group with a long MWL included Saga, Nagasaki, Saitama, Gunma, and Kumamoto. Apart from Saitama, the dialects were found in southern Japan. The group with middle MWL covered the Kanto, Tokai, Kinki, and Chugoku areas. The Shikoku area seemed to be characterised by a shorter MWL. The majority of dialects presented word length ranging from 1-9. Nagasaki and Wakayama exhibited the richest distribution of word length, ranging from one to 12. The Saga dialect had the longest MWL (3.26), followed by Nagasaki (3.25). A further analysis of the MWL-frequency relationship via the Altmann-fitter showed that MWL_frequency of all the dialects fit more than 30 distribution models, including the binomial and Poisson families. The finding slightly differed from Zipf's (1949) principle of least effort, which stated that shorter words tended to be used more frequently. In all Japanese dialects, the frequency of MWL increased along with an increase of MWL from one to two. Starting from two, the frequency decreased along with an increase in MWL.

References

Hattori, Shiroo. 1933. Accent and dialect. [Akusento to hoogen]. Meijishoin.

Hirayama, Teruo. 1936. The accent of south Kyuushuu 1, 2. [Minamikyuushuu akusento no kenkyuu]. Dialectology [Hoogen] 6 (4, 5).

Ishii, Hisao. 1990.Words in Chuuoo Kooron 1986 [Chuuoo Kooron: 1986 nen no yoogo]. *National Institute for Japanese Language and Linguistics Research Reports* [Kokuritsukokugokenkyuusho Hookoku] 101, 1-40.

Kindaichi, Haruhiko. 1937. A comparision of Japanese dialects and the accent in Heian Period: a focus upon two-syllable noun. [Gendai shohoogen no hikaku kara mita Heianjidai no akusento: tokuni nionsetsumeeshi ni tsukite].

Koshigaya, Gozan. 1775. Monorui koshoo.

Kushihiki, Yukiko. 2008. A shift of semantic meaning in Japanese dialects. [Nihongo hoogen goi no imihenka ni kansuru kenkyuu]. Doctoral thesis, Tohoku University, Japan.

Mendenhall, T. A. 1887. The characteristic curves of composition. *Science*, 11, 237-249.

Mendenhall, T. A. 1902. A mechanical solution to a literary problem. *Popular Science Monthly*, 60, 97-105.

Minami, Yasuhiro and Kobayashi, Tessee. 2013. Correlations between Word Lengths and Word Acquisition Times and Periods of Infants and Toddlers. [Go no nagasa to yooji no goi shuutoku jiki, kikan to no sookan]. *Journal of the Phonetic Society* [Onsee kenkyuu] 17 (3), 44-53.

Nitta, Yoshio. 1992. On the case markers: a comparison with Tohoku dialect. [Kaku hyooji no arikata o megutte: Toohoku hoogen to notaisyoo no motoni]. *Journal of Guugo linguistics* [Guugo Gogaku] vol. 11. Meiji Shoin.

Sanada, Haruko. 1997. The shift of Chinese translation in Meiji Period: a comparison of Philosophy lexicon and multiple vocabulary list [Meijiki kanyakugo no nagare:

Tetsugaku jii to kakusyugoihyoo to no hikakuchoosa]. Presented at *The 142nd Conference on Modern Japanese Language Research* [Kindaigo kenkyuukai Dai 142 kai kenkyukai].

Shinbo, Kaku and Tsunemi, Chisato. 1932. Japanese accent dictionary. [Kokugo hatsuon akusento jiten].

Yamada, Bimyoo. 1893. Japanese Dictionary [Nihongo daijiten].

Toyota, Toyoko. 1985. An investigation to Japanese conjunctions “to, ba, tara, nara” [To, ba, tara, nara no yoohoo no choosa to sono kekka]. *Journal of Japanese language teaching* [Nihongo kyooiku] 56.

Zipf, G. K. 1949. *Human behaviour and the principle of least effort*. Cambridge: Addison-Wesley.