Lexical variation in Japanese dialects revisited: Geostatistic and dialectometric analysis

Péter Jeszenszky [a, *], Yoshinobu Hikosaka [b], Keiji Yano [a]

[a] *Department of Geography, Ritsumeikan University, Kyoto, Japan, pjeszenszky@gmail.com; yano@lt.ritsumei.ac.jp*
[b] *College of Letters, Ritsumeikan University, Kyoto, Japan, atsumi053145@gmail.com*

* Corresponding author

Keywords: Spatial analysis, geostatistics, GIScience, linguistic geography, dialectology, geovisualisation

Since the end of the 19th century in Japan, the official language policy enforced using Standard Japanese, based on the variety spoken in Tokyo (formerly Edo), in all official situations and in schools. Since then, Japanese dialects have been dwindling and 'flattening' (i.e., they retain less regional variation). Nevertheless, differences of language varieties keep being important topics and they reinforce the feeling of belonging and group formation in Japan, similarly to most languages with dialects. This study explores the spatial patterns in Japanese lexical variation based on digitised dialectal survey data (using the Linguistic Atlas of Japan) and presents first results of a dialectometric analysis, quantifying a number of factors assumed to affect lexical variation in Japanese.

Although several different research directions have been explored using the data from the Linguistic Atlas of Japan (LAJ) and other dialect surveys, lacking digitised data many of these directions could not have been discovered very deeply (Takada, 1969; Hondo, 1980; Inagaki, 1980; Kasai, 1981; Ichii 1993; Inoue 2001, 2004; Kumagai, 2013, 2016). One of dialectometry's most important foci is the quantitative expression of linguistic differences across the surveyed locations (e.g., Seguy, 1971; Goebl, 1982; Nerbonne, 2010) and attributing these '*linguistic distances'* to some geographical measures which usually account for the possibility of contact and isolation between speakers of a language (such as geographic distances, travel times, and gravity-like urban hierarchy relations). Most of the research on Japanese dialects focused, however, on the linguistic relations themselves, and did not often account quantitatively for the underlying factors assumed to affect dialectal variation.. Based on previous work in linguistic geography (e.g., Gooskens, 2005; Spruit, 2006; Szmrecsanyi, 2012; Jeszenszky et al., 2017; Sieber, 2017) it is possible to hypothesise the following patterns with regards to explaining linguistic variation based on language-external factors. Geographic distances will explain a considerable amount of the variation, due to the fact that usually the linguistic variables queried in dialect atlases are assumed to exhibit spatial variation. The logarithm of the geographic distances is expected to have a larger explanatory power, as linguistic distance reaches a sill (the maximal linguistic distance, lack of linguistic similarity), beyond which it cannot grow anymore, while the geographic distance constantly grows. The pace of language change in dialects is different area by area and with regards to various linguistic aspects. It is generally acknowledged that historical contact paths and isolation patterns might have contributed more to today's language variation than the contact paths and isolation patterns visible today. With the recent digitisation of the Linguistic Atlas of Japan (LAJ), in this study we focus on the research gaps of the 1) survey site level dialectometric analysis of Japanese dialect data, 2) the associations across dialectal features, 3) the geostatistical account for dialect areas and 4) the discovery of the effects of some historical and geographical factors on the variation.

This preliminary study uses digitised and publicly available data from the Linguistic Atlas of Japan (LAJ), provided by Yasuo Kumagai at the National Institute of Japanese Language and Literature (NINJAL) (also see Kumagai, 2016). The atlas contains 285 questions (termed *variables* in this abstract), mostly about lexicon (variation in vocabulary, mostly common nouns, verbs and adjectives). The survey providing the data was conducted between 1957 and 1968. Throughout Japan, 2400 locations were surveyed, with one elderly male respondent at each survey site. We use 37 digitally available questions' data at the time of the submission, 37 questions. According to the concept of '*apparent time'* (Bailey et al. 1993), mother tongue is mostly acquired until the late teenage, after which one's language variation is more resistant to change. Therefore, it is assumed that one's language bears the signs of the environment of their early life. As LAJ-respondents were born between 1879 and 1903, their language usage is supposed to be representative of the late 19th, early 20th century. We hypothesise this variation to be affected by the earlier Japanese administrative system of so called *domains* (Japanese: *han*) as their boundaries restricted the movement of their inhabitants.

At the beginning of the dialectometric analysis of the lexical data, first we discovered the associations across the dialectal variants used. As for many dialectal variables a great number of *lexical variants* (~10 to ~500) are in use by respondents, we categorised the variants that are similar to each other based on the originally published LAJ maps (NLRI, 1966–1974), resulting in 3-15 categories for each variable. We tested corresponding usage across these categories by calculating a

Abstracts of the International Cartographic Association, 1, 2019. © Authors 2019. CC BY 4.0 License.
29th International Cartographic Conference (ICC 2019), 15–20 July 2019, Tokyo, Japan | https://doi.org/10.5194/ica-abs-1-148-2019

parity-based distance for the usage of each variant category pair. Respondents using the same categories for several variables, if spatially clustered, can be associated with dialectal areas. Finding the associations this way, importantly, helps avoid the subjectivity posed by drawing dialect areas on maps, traditionally used to discover dialect area formation. Figure 1 shows the association graph of the variant categories. Based on the 37 variables used, two smaller clusters are associated with variations in the southern archipelago of Okinawa and the large central cluster represents the standard variants, usually found spread across large areas in the main island Honshu.

Linguistic distances have been calculated for each pair of survey sites based on the variant categories. The linguistic distance between a survey site pair is defined by the sum of differing answers for the 37 questions (variables). The resulting distance can be mapped from any survey site. Figure 2 maps the linguistic distance from Tokyo, Osaka, Sendai, Kagoshima, a rural site in Aomori and an island in Okinawa. Usually the closer a survey site is to the central site, the smaller their linguistic distance seems, but the Okinawa archipelago always tends to show larger linguistic distances. It is interesting to see that the linguistic distances to Tokyo (the birthplace of Standard Japanese) tend to be smaller throughout Honshu, the biggest island of Japan. The average linguistic distance map (Figure 3) plots for each survey sites the average of linguistic distance towards all sites. As Okinawan variations are even argued to be a distinct language, unsurprisingly they seem to be the most different on average. Japanese speakers settled Hokkaido mainly from the 19th century from several parts of then Japan, resulting in the most flattened, similar dialects, which is reflected on the map by small average linguistic distances. Based on the 37 variables, it is surprisingly not the Tokyo area, the source of standardised Japanese, that shows the greatest average similarities, but the midwestern *Kansai* area, which contains the former capital Kyoto and economical hub Osaka. We used multidimensional scaling to reduce the 2400*2400 linguistic distance matrix into its three most representative vectors. Projecting these vectors into the RGB colour space makes it possible to give each survey site a colour, practically mapping areal dialect similarities. From Figure 4 at least 5 different dialect areas can be defined, the distinct nature of Okinawan dialects becomes visible. Besides, the immediate effect of the Tokyo variation can be tracked, while we can interpret the mixed picture in Hokkaido as mixed dialects.

Using great circle distances between all pairs of survey sites, we calculated the correlation between the linguistic distance and the geographic distance, and its logarithm, respectively, to show the amount of variance they explain in the linguistic distance matrix. Figure 5 shows the association between geographic distance and linguistic distance in a hexplot with the linear and logarithmic regression lines overlaid. Pearson's product-moment correlation coefficients reveal that the logarithm of geographic distances explains slightly more variance in the linguistic distance (r =0.6493 and r =0.6708, respectively).

The effect of administrative boundaries was tested using the non-parametric Mann-Whitney U test, testing the overlap of two groups of values. We tested the effect of *prefecture* boundaries (today Japan is divided into 47 prefectures), and the boundaries of the 68 *domains* in 1868. The Mann-Whitney U test was done with the following groups: 1) survey site pairs, where both sites are inside the same prefecture or domain and 2) survey site pairs, where one site is in the prefecture/domain in question and the other in another prefecture/domain, but less than 200 km away. Vargha and Delaney's A (2000) is a related effect size statistic, showing the probability that a value sampled from one group will be greater than a value sampled from the other group, unaffected by sample size. The value A=0.2233647 for prefectures means that links across prefectural boundaries have a large chance to have greater linguistic distance than links within prefectures. On the other hand, A=0.33705 for domains means only a smaller chance that links across domain boundaries have a greater linguistic distance than links within domains. This is somewhat surprising as a larger effect of domain boundaries was expected due to their historical isolating role. However, the prefecture system is largely based on the previous domain boundaries, but prefectures have usually larger size, allowing for larger distances between survey sites that are divided by prefecture boundaries. Figure 6 shows the density plots of linguistic distances for groups *1* and *2* for the prefecture boundaries.

In the present state of the research, the 37 variables and the methods used so far deliver exploratory results in the dialectometric analysis of Japanese lexicon and show us possible directions for measuring the association within linguistic variables and contrasting them to external, geographic factors. We are in the process of extending the methodology of this revisited analysis of dialectal differences is. Beside the solidification of the above results through involving more digitised variables, the following measures will be calculated. Digital elevation models will be used to find the most probable natural contact paths between survey sites, which are supposed to have been used for hundreds of years before the industrial revolution, assumed to form the local linguistic differences for a long time. Correlation with modern travel times, an informed guess about today's possible contact paths will also be calculated (similarly to Gooskens, 2005 and Jeszenszky et al., 2017), along with historical travel times (sourced from georeferencing historical maps and network analysis of Edo-era port data), fitting more the time when the respondents grew up, acquiring their mother tongue. We will implement Trudgill's (1974) linguistic gravity theory, which supposes that linguistic influence between communities is arranged in a way similar to gravitational interaction, with population playing the role of mass.

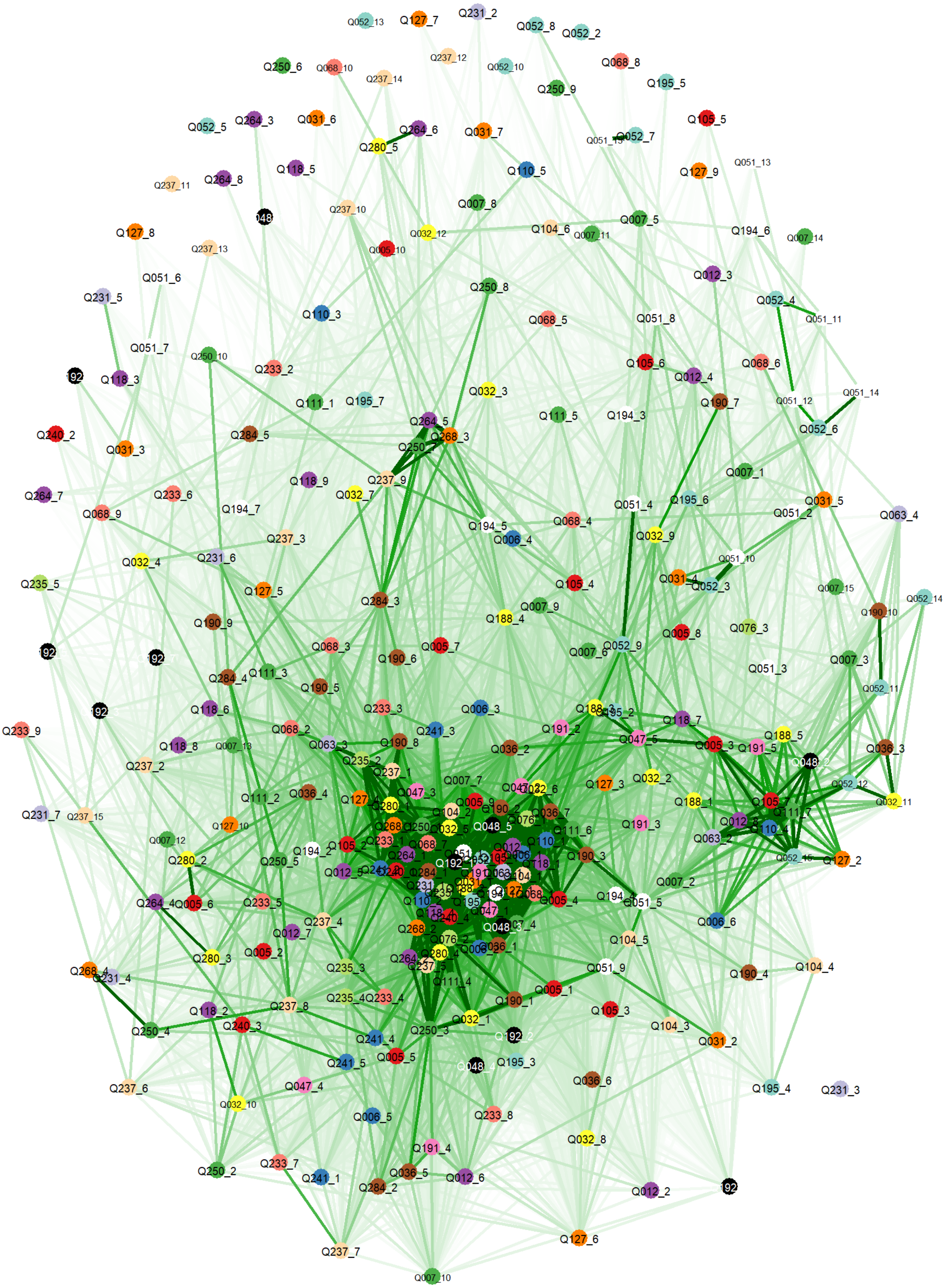

Figure 1. Association graph of the variant categories linked to the 37 questions used in the study. The thickness and saturation of the links' green colour stands for the strength of the association (the spatial overlap) across the variant categories. Variant categories that belong to the same variable have the same colour.

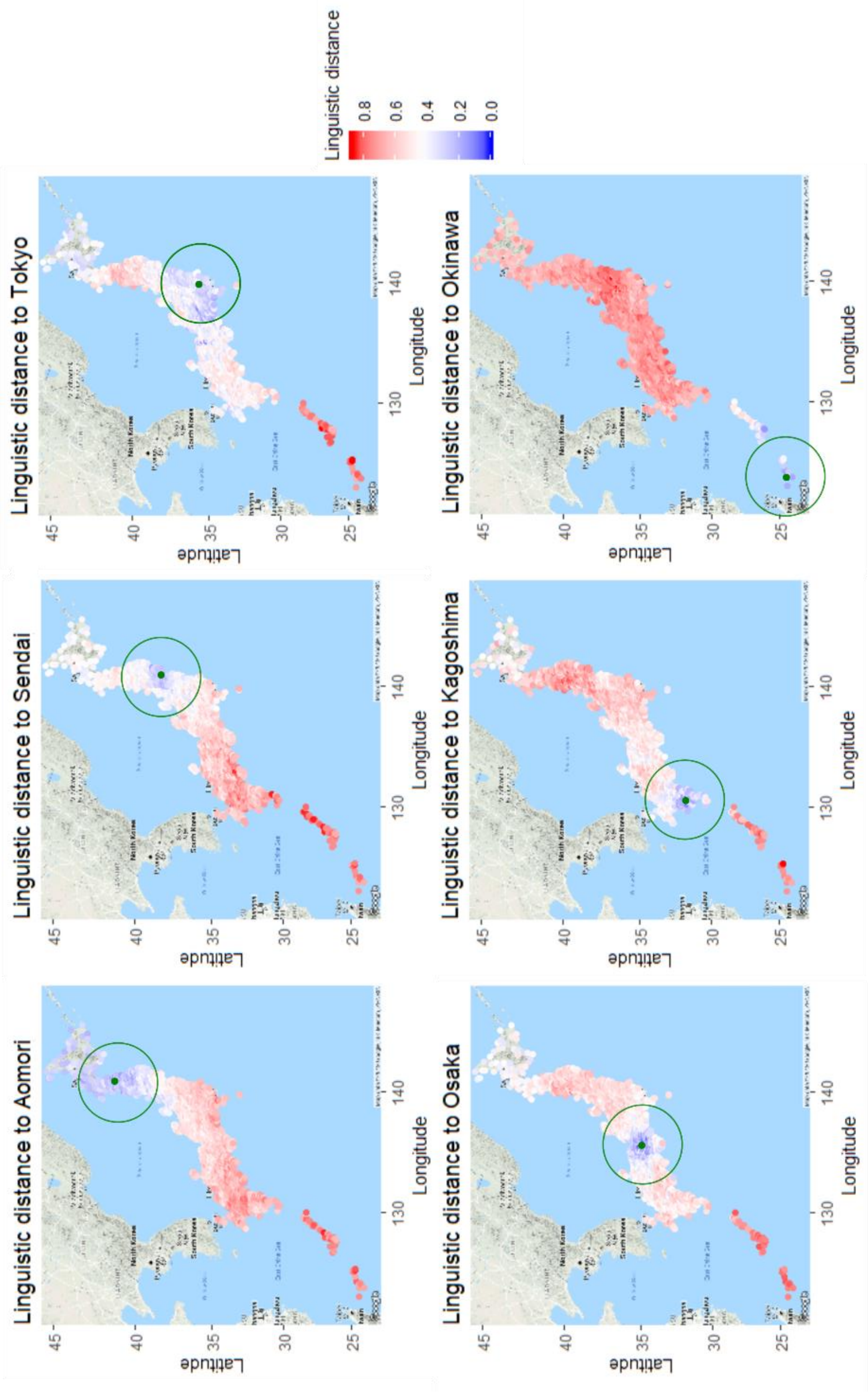
Linguistic distance to Aomori
Linguistic distance to Sendai
Linguistic distance to Tokyo
Linguistic distance to Osaka
Linguistic distance to Kagoshima
Linguistic distance to Okinawa
Latitude
45
40
35
30
25
Longitude
130
140
Linguistic distance
0.8
0.6
0.4
0.2
0.0

Figure 2. Linguistic distance maps with central survey sites in various areas of Japan. The encircled green dots mark the central survey sites.

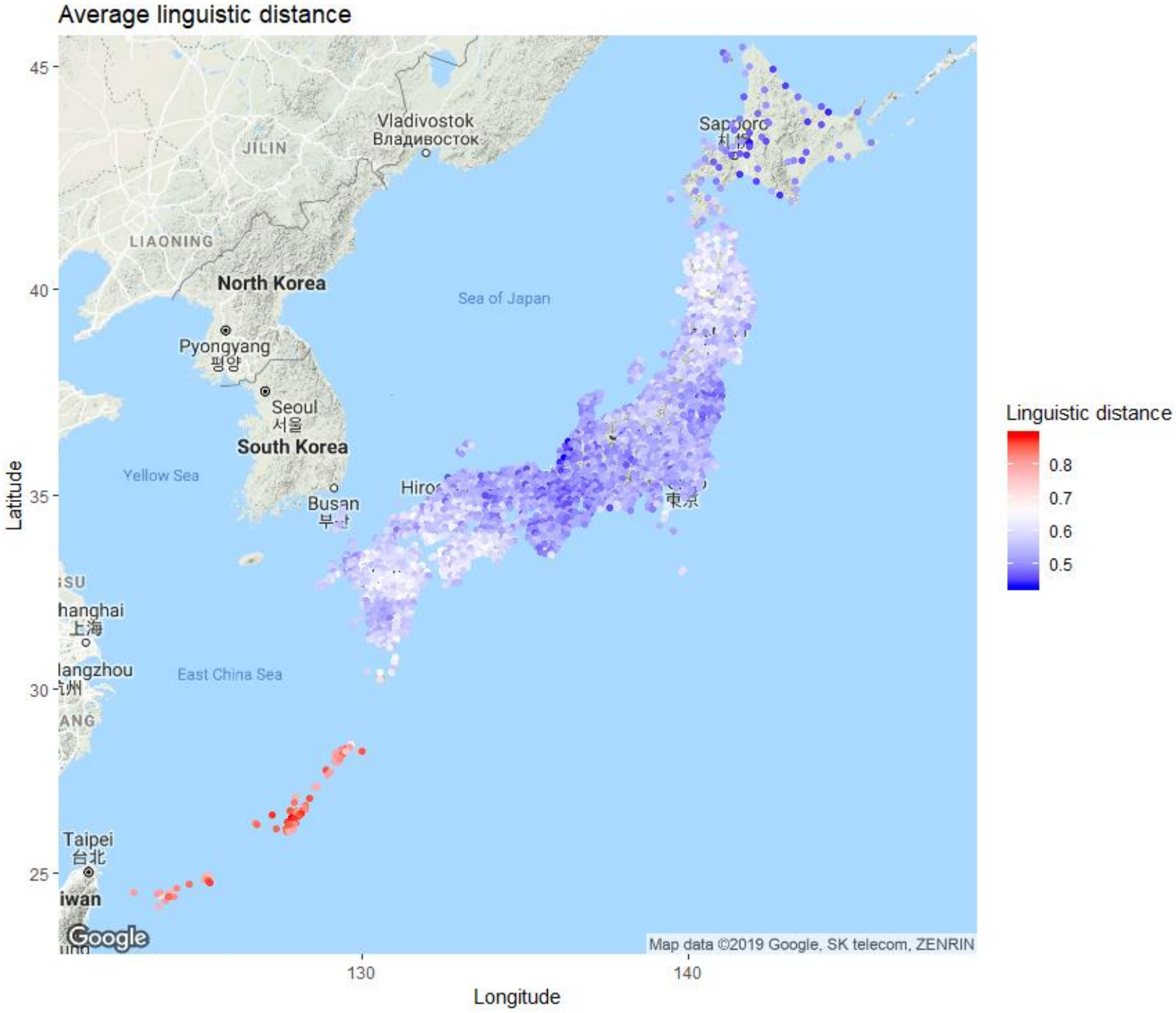

Figure 3. Average linguistic distance map based on 37 digitised variables in the Linguistic Atlas of Japan. At each survey site, the colour represents the average of the linguistic distances to all other survey sites.

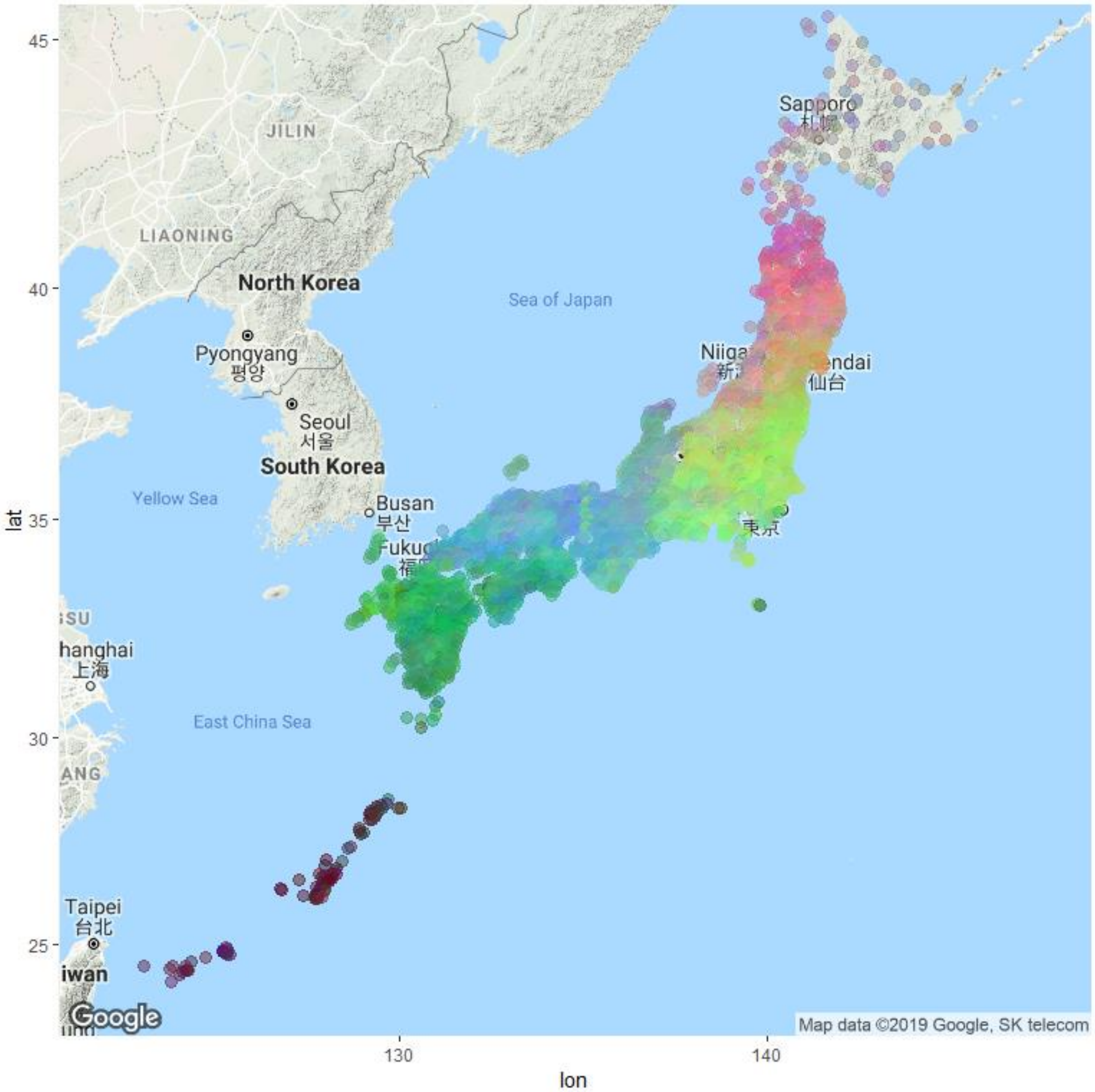

Figure 4. RGB vectors based on the results of multidimensional scaling mapped. Similar colours represent dialectal similarity with regards to the 37 variables.

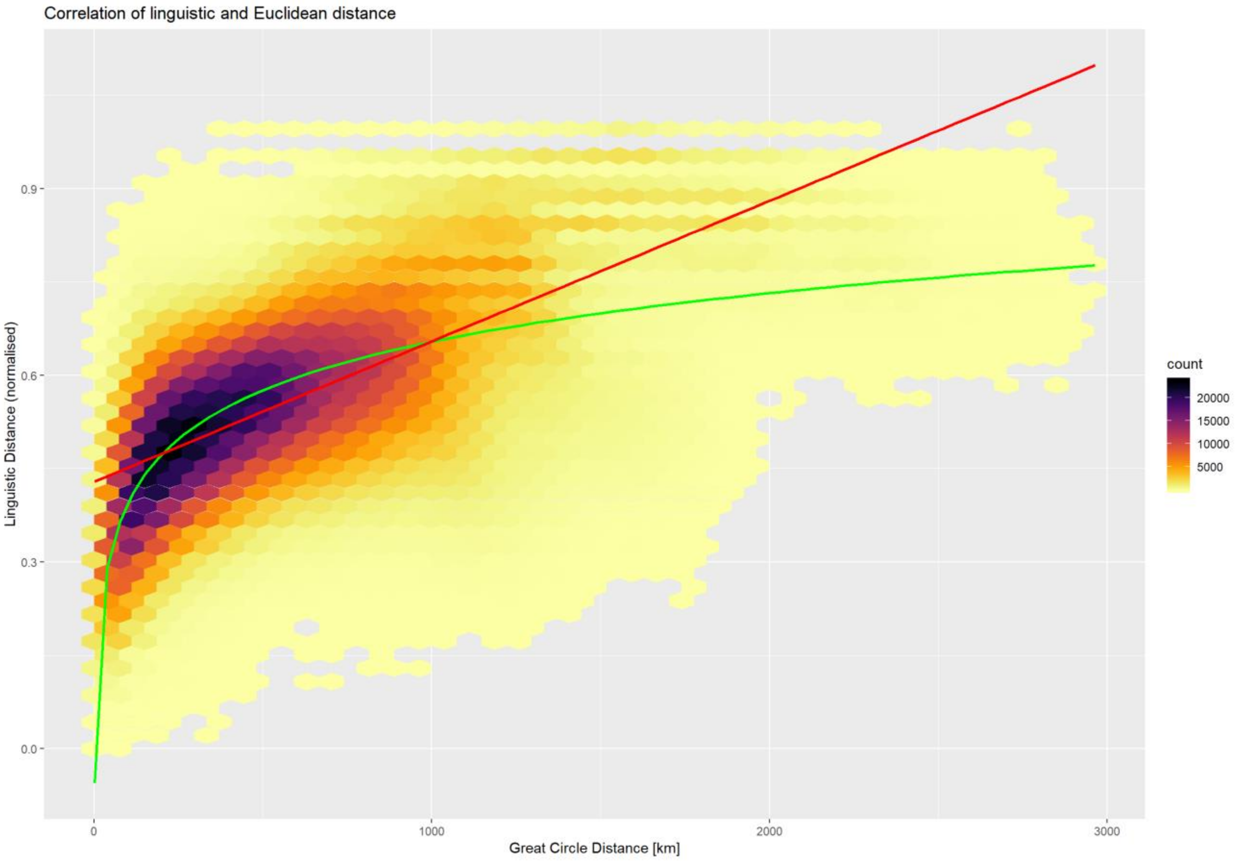

Figure 5. Correlation plot of the great circle distance and linguistic distance with the linear and logarithmic regression line overlaid.

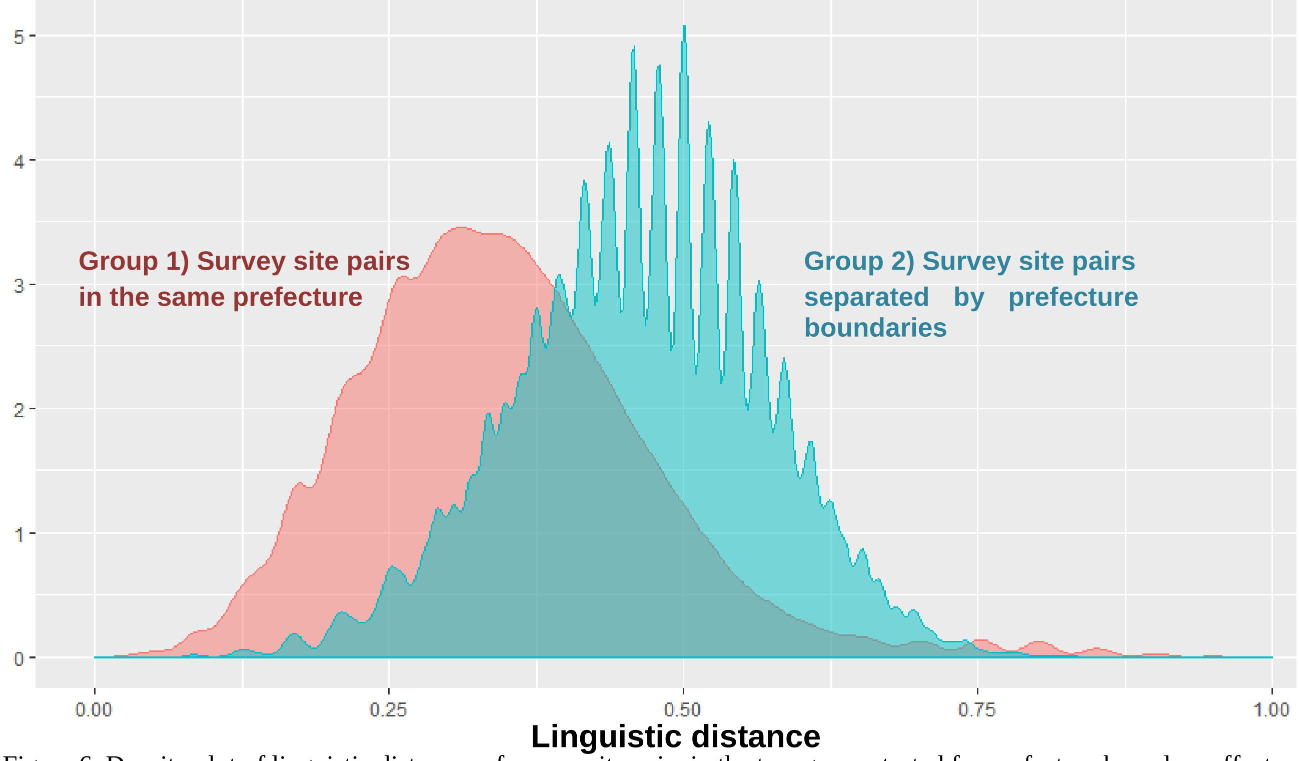

Figure 6. Density plot of linguistic distances of survey site pairs in the two groups tested for prefecture boundary effects. It is visible that the survey site pairs that are isolated by a prefectural boundary often have larger linguistic distances than those that are both within a prefecture.

References

Bailey, Guy, Wilke, Tom, Tillery, Jan, & Sand, Lori. 1993. Some patterns of linguistic diffusion. *Language Variation and Change*, *5*, 359–390.

Goebl, Hans. 1982. *Dialektometrie* (Philosophisch-historische Klasse, Denkschriften, 157. Band). Vienna: Österreichische Akademie der Wissenschaften.

Gooskens, Charlotte. 2005. Travel time as a predictor of linguistic distance. *Dialectologia et Geolinguistica*, *13*, 38–62. http://doi.org/10.1515/DIALECT.2006.003

Hondô, Hiroshi. 1980. Gendai hyôjun nihongo no bunpu: Nihon gengo chizu de mite [Distribution of modern standard Japanese: An observation by using the LAJ]. In Shigeru Satô (ed.), *Sato shigeru Kyoju taikan kinen ronshu kokugogaku*. 479–498. Tokyo: Ohfusha.

Ichii, Tokiko. 1993. Hôgen to keiryô bunseki [Dialect and quantitative analysis]. Tokyo: Shintensha.

Inagaki, Shigeko. 1980. Hôgen sesshoku to gokei heiyô: "Nihon gengo chizu" no bunpu kara [Dialect contact and doublets: Some examples from the distributions in the LAJ]. *Tôkyô toritsu daigaku hôgen kenkyûkai kaihô* 92. 1–10.

Inoue, Fumio. 2001. Keiryôteki hôgen kukaku [Quantitative dialect division]. Tokyo: Meiji shoin.

Inoue, Fumio. 2004. Heiyô genshô to gengo genshô no chûkan dankai: Kasai data 3 kurasutâ no hukyû katê [Joint usage of forms and intermediate stages of linguistic change: Process of diffusion of 3 clusters of Kasai data]. *Gogaku kenkyûjo ronshû (Journal of the Institute of Language Research)* 9. 1–19.

Jeszenszky, Péter, Stoeckle, Philipp, Glaser, Elvira, & Weibel, Robert. 2017. Exploring global and local patterns in the correlation of geographic distances and morphosyntactic variation in Swiss German. *Journal of Linguistic Geography*, *5*(2), 86–108. http://doi.org/10.1017/jlg.2017.5

Kasai, Hisako. 1981. Hyôjun gokei no zenkoku bunpu [Nationwide distribution of standard forms]. *Gengo seikatsu* 354. 52–54.

Kumagai, Yasuo. 2013. *Development of a Way to Visualize and Observe Linguistic Similarities on a Linguistic Atlas* (Working Papers from NWAV Asia-Pacific 2.) (Vol. 3).

Kumagai, Yasuo. 2016. Developing the Linguistic Atlas of Japan Database and advancing analysis of geographical distributions of dialects. In M.-H. Cote, R. Knooihuizen, & J. Nerbonne (Eds.), *The Future of Dialects. Selected Papers from Methods in Dialectology XV.* (pp. 333–362). Berlin: Language Science Press. http://doi.org/10.17169/langsci.b81.159

Nerbonne, John. 2010. Mapping aggregate variation. In *Language and Space. An international Handbook of Linguistic Variation. Vol 1. Theories and Methods* (pp. 476–495). Berlin/ New York: Mouton de Gruyter.

National Language Research Institute - Kokuritsu Kokugo Kenkyûjo (NLRI). 1966-1974. *Nihon gengo chizu (Linguistic atlas of Japan)*. Tokyo: Printing bureau, Ministry of Finance.

Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane*, *35*(138), 335–357.

Sieber, Christian D. 2017. *Einfluss von scharfen und unscharfen Grenzen auf syntaktische Dialektunterschiede in der deutschen Schweiz*: Master's thesis.

Spruit, Marco R. 2006. Measuring syntactic variation in Dutch dialects. *Literary and Linguistic Computing*, *21*(4), 493–505. http://doi.org/10.1093/llc/fql043

Szmrecsanyi, Benedikt. 2012. Geography is overrated. In S. Hansen, C. Schwarz, P. Stoeckle, & T. Streck (Eds.), *Dialectological and Folk Dialectological Concepts of Space - Current Methods and Perspectives in Sociolinguistic Research on Dialect Change* (pp. 215–231). Berlin, Boston: De Gruyter.

Takada, Makoto. 1969. Kotoba no chiri: Nihon gengo chizu kara [Geography of words, Kyûshû district: An observation by using the LAJ]. *Gengo seikatsu* 216. 30–38.

Trudgill, Peter. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, *2*, 215–246.

Vargha, András, & Delaney, Harold D. 2000. A Critique and Improvement of the CL Common Language Effect Size Statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*(2), 101–132. http://doi.org/10.3102/10769986025002101