Appendix B : Details of the Research That Follows

journal or publication title	Senri Ethnological Studies
volume	108
page range	163-172
year	2022-03-09
URL	http://hdl.handle.net/10502/00009908

Senri Ethnological Studies 108: 163–172 ©2022
Fijian Languages, Cultures, and Their Representation
Edited by Kikusawa Ritsuko and Sano Fumiya

Appendix B: Details of the Research That Follows

A Joint International Research (B) program, "Integrating Language Change in Time and Space: Applying Geographical Information System (GIS) and Statistic Modeling to Historical Linguistics" has been funded by Grants-in-Aid for Scientific Research by the Japanese Society for the Promotion of Science (JSPS) for 2019–2023, PI: Kikusawa. This is a translation of the original Japanese grant application submitted in 2017.

1. Research Objective and Methods

In this research, large-scale data (approximately 300 dialects and 5,800 words) of Fijian dialects will be used as test cases to clarify the relationship between temporal variation of language and spatial propagation. It will further verify the correlation between the results and non-linguistic information (e.g. topography and the distance traveled between points) to comprehensively capture linguistic variation in association with human activity. As a developmental history of Fijian dialects, a linear bifurcation model (East/West Fiji) reflecting time variation and a dialect chain reflecting spatial diffusion have been proposed, but there is no method to compare and integrate them. Here, geographic location is regarded as an intersection of language change, and views of time-space are associated using a Geographic Information System (GIS) to determine specific developmental processes. To enable this, the necessary language data will be prepared and GIS construction and utilization will be adopted. To enable this, a research team will be formed in the collaboration of The University of the South Pacific (Fiji), known for historical linguistic research in Oceania, and Massey University (New Zealand), known for GIS research. In addition, young Japanese researchers and graduate students will perform statistical modeling and verification from the perspective of other languages to generalize the analysis results and will participate in the writing of an international collaborative paper while being involved in data manipulation and analysis, thereby linking to future international research project development.

(1) Academic Background of the Study and the Question at the Core of the Research Agenda

By simultaneously looking at spatiotemporal variation in a language, this study will address the following research questions. How can we clarify the historical variation of language? Based on that, how can we elucidate prehistoric human social activities? To answer these questions, high-quality, large-scale data of Fijian languages that are available will be used to establish a methodology. Behind this are emerging problems that cannot be solved in the micro-level comparison and reconstruction (meaning, comparison and reconstruction of genetically and geographically close languages). Such problems occur as a result of not

having a way to generalize and incorporate the effects of spatial propagation in the traditional comparative method. This research will tackle the following more specific questions with the general goal of finding a more advanced and effective way to clarify linguistic history.

1) How does time-series variation in language correlate with variation from space propagation? In historical linguistics, elements of direct inheritance from the parent language that are based on sound correspondences are represented in the form of a phylogenetic tree model by specifying genetic relationships among languages based on shared innovation. In contrast, geographical distribution serves as one of the clues to deal with language change that involves borrowing and language contact. These two methodologically distinct variations are known to progress in relation to each other in the actual language. Clarifying and modeling how temporal-spatial language changes influence one another leads to a comprehensive understanding of language change.
2) How do we elucidate the developmental history of language with both time-series variation and space transmission? To elucidate the developmental history of languages without textual records, a common method is to compare and reconstruct data of the currently spoken language on a time axis. For spatial variations, it is not possible to compile a map showing the distribution of variations in the past, for the lack of the documentation. To capture such distribution, we need to refine the methodology, as questioned in 1), and incorporate spatial analyses in the reconstruction of language development.
3) What environmental, social, and cultural factors show correlation with temporal-spatial language changes? It has long been known that various factors related to human activity, such as environment, society, and culture, influence language change. However, correlation with topography and transportation network is verified by the abstract concept of "bundling dialect boundary lines." In actual language change, this correlation shows different categories that depend on the language event category (e.g. the nature of the syntactic structure, the type of morpheme, and the semantic category of vocabulary), time, social background, etc. Clarifying correlation between individual language events revealed in Question 2) and non-linguistic events leads to the elucidation of language change and human social activity. Further, it is expected to lead to an understanding of the histories of human movement and exchange through goods/things at the micro-level.
4) What is the deductive approach for understanding language change and the method for quantitative verification? Historical linguistic methods, which are the premise of Questions 1) through 3) above, are inductive and qualitative approaches wherein experts analyze large amounts of data based on knowledge and experience. Visual analysis is limited to the applicable language groups, and there is no method to model and objectively validate analytical findings. In contrast, quantitative and statistical approaches to languages, which have been developed and have gained popularity in recent years, are currently limited to typological classifications of a language, and even for time variation, the data are often linguistically inappropriate, which does not lead to the elucidation of language change. Mathematical modeling of variation based on historical

linguistic methods and descriptive linguistics is an essential task to link statistical analysis to actual linguistic change.

(2) Purpose and Academic Originality of This Research

The originality of this research resides in adopting the methods developed in geography, statistics, and cultural anthropology in relation to language change and developing inter-disciplinary research in linguistics. The specific characteristics are as follows:

1) Integrating temporal and spatial factors using the latest technology to comprehensively capture language change by employing a geographically positioned base (the Republic of Fiji), using GIS, using tools that can perform dynamic expressions (e.g. zoom in, zoom out). Outputs to do these are techniques that are original to this study and have not been previously applied in historical linguistics. This will allow classification based on sound correspondences reflecting temporal and vocabulary change, capturing language change in the context of space and time by dinamicly combining geographical distributions of linguistic elements that reflect spatial propagation.
2) Categorizing developmental patterns for each category of language elements by using data from a wide range of categories. This is without narrowing the target to directly inherited words so as to combine conventional comparative methods with indirectly inherited words, in addition to borrowing and contact analysis after bifurcation is original to this study. Specifically, all 5,800 words will be analyzed, including, i) 100 basic vocabulary words; ii) a 100-word comparative list, in which the differences between Fiji dialects are observed; and iii) vocabulary, grammatical function words, animal and plant names, and technical and cultural terms belonging to various semantic categories.
3) Elucidating language change in relation to human activity while objectively supporting it by using GIS. This includes attempts to quantitatively and objectively examine the correlation between language change and non-linguistic elements through the calculation function of GIS and to capture language change in the context of human activity are both original for this research. To use GIS analysis tools, data with no gaps in the positional relationship are required. Detailed linguistic data of Fijian dialects and information on the residence and social behavior of the speakers are rare examples that satisfy this criterion. This allows different combinations of trials with various parameter settings.
4) Inductive analytical findings verified from a deductive point of view by introducing statistical models to make the method versatile. The method proposed in this application can be applied to other language groups. It is linked with a systematic statistical model being developed in the field of natural language processing, thereby understanding spatial elements by newly developing the spatial Gaussian process and factor models. Such are original and creative aspects to be brought into historical linguistics.

(3) Clarifying Questions Such as "What?" "How?" and "To What Extent?"

As a sample, data of about 300 Fijian dialects "communalects" in Fijian Linguistics are

collected, and a methodology for clarifying the complete developmental linguistic history will be established. Modern Fijian languages took about 2,000 years after the split from Proto-Central Pacific, their parent language, to typologically form two language groups of East and West. Through micro-level analysis, details of the language formation process of each dialect group and the socio-cultural factors that influenced it are clarified. To this end, change (the development history of sound correspondences and vocabulary) is projected in accordance with the time line on a map and compared with the data distribution of the 5,800 words. A distinction is made between words that match, do not match, and match in part. Next, combinations and spatial distribution patterns of the sounds contained in each word are analyzed and categorized. The developmental history is traced, and correlation between each type and relevant non-linguistic information is verified and modeled. Through this process, syntactic structure and pragmatic features are methodologically applied to the point that they are recognized as the elucidation of their developmental history. Furthermore, based on the knowledge obtained about Fiji, clarification will be linked to human movement across prehistoric Oceania. Research will be advanced by collaborating with members by using internet tools. Research seminars will be held every year in Fiji or New Zealand to improve trajectory modification and techniques. The work schedule is presented in Figure B-1 below. For each milestone, the results will be made public by reporting them at international academic conferences and by writing in international collaborative papers.

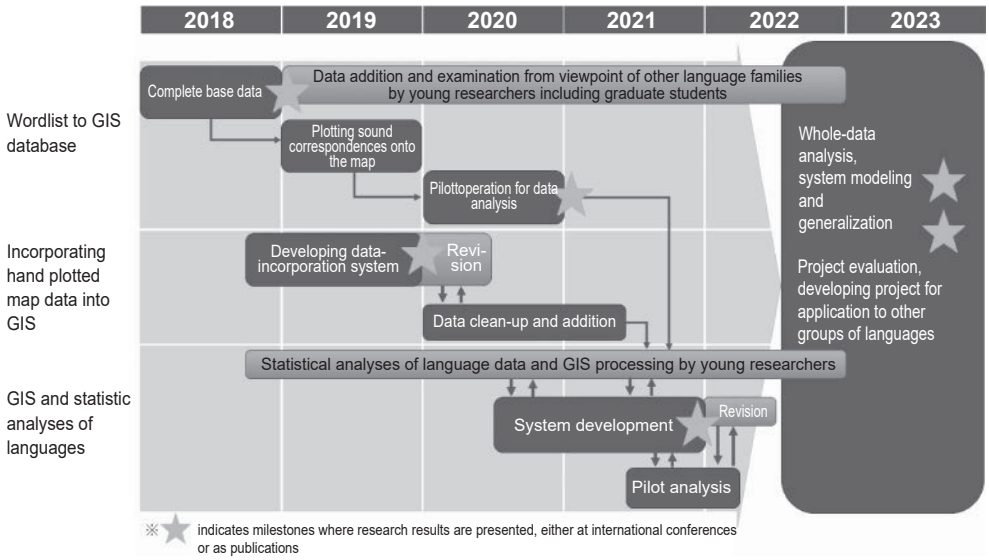

Figure B-1 Work schedule

(4) Specific Roles of the Research Representative, Researcher, and Research Collaborator

Research representative Kikusawa Ritsuko is responsible for the overall review and coordination between the three fields. Murawaki Yugo, analyzes and statistically processes language data in accordance with the time axis, and Mochihashi Daichi analyzes and statistically processes the spatial distribution of language data. Yoshioka Noboru, evaluates the validity from the perspective of linguistics in general, and the validity and versatility of the analysis and methodology from the perspective of other language groups, and makes suggestions for trajectory modification as necessary. Paul Geraghty performs Fijian language data preparation and analysis from a comparative linguistic perspective, collaborating with Apolonia Tamata, Okamoto Susumu, and Kikusawa. John Lowry proposes GIS-based language processing, and Murawaki is in charge of development. Okamoto and Sano Fumiya oversee system development, and Teramura Hirofumi provides technical support. The analysis combines the linguistic perspectives of Kikusawa, Tamata and Geraghty; the cultural and social perspectives of Tamata, Sano, and Niwa Norio; the geographical perspective of Lowry; and the mathematical process modeling of Mochihashi and Murawaki.

2. Significance and Necessity of International Collaborative Research

(1) How the Idea Developped

In recent years, cases have been reported wherein phylogenetic relationships of languages that do not have written records cannot be expained. Most of them relate to the comparison and reconstruction of languages that are geographically close. Kikusawa, having seen this, pointed out that causes are that, as a result of shared changes caused by language contact, linguistic features that have been inherited from their remote ancestor language(s) are obscured in such cases (Kikusawa 2018). She noted also that the Fijian dialect data, compiled by Geraghty, thorough and of languages that are geographically close to each other can be useful in the search for new research methods for such a case. She happens to have conducted descriptive work of some Fijian dialects herself, and therefore, this appeared to be a good place to start. To find a way to integrate space and time, she also thought that using a GIS has a potential as a new tool. The cooperation of Lowry, a GIS expert with rich experience in interdisciplinary research, has been obtained, and the execution of this task because feasible. To secure objective verification of the results by mathematical processing; the cooperation of Mochihashi and Murawaki (Fujii et al. 2017; Murawaki 2015a, 2015b, 2017; Murawaki et al. 2018; Nakamura et al. 2017; Noji et al. 2013; Ohishi et al. 2014; Shinozaki et al. 2017; Uchiumi et al. 2015; Yamauchi et al. 2016), who specialize in the mathematical processing of time and geographical language data, and that of Yoshioka (Yoshioka 2016, 2017, 2018a, 2018b) has been obtained. He is interested in spatial distribution and the historical variation of other language groups. With this team with various specialists in relevant fields, in addition to Geraghty, the compiler of the original data, and Tamata, a Fijian linguist and Anthropologist, the research project appeared to be ready to launch.

(2) Research Trends in Japan and Abroad and the Position of the Proposed Research

1) In the comparative reconstruction of languages, especially languages without written records, a method has been developed and applied to identify indirectly inherited words and loanwords in time variation (Kikusawa 2015). In addition, research focusing on the distribution of regional and social variants and tracking historical variation behind it is widely developed. This study is a new attempt to integrate the two using GIS tools, based on accumulation in these fields.
2) Currently, there is no statistical model that can spatially represent discrete linguistic data, but one has just appeared in the deep learning study (Gal et al. 2016) that deals with data with uncertainty and has great potential. Additionally, the systematic method does not consider such spatial elements, and it is thought that incorporating a spatial statistical model of language as a starting point will lead to the development of new research fields.
3) For Fijian phylogenetic and dialect classification, Pawley and Sayaba (1971), Geraghty (1983), and Kikusawa (2002), each proposed different hypotheses: a bifurcation diagram, that is a phylogenetic diagram including a dialect chain, and a partial bifurcation diagram. In this study, verifying the validity of each hypothesis and consolidating Fijian language data and history analysis in the process will contribute to future research on Fijian language.
4) The use of GIS in historical linguistics is limited to the display of linguistic data on a map, and there is almost no usage that takes advantage of the characteristic feature that the map information and database are linked and designed for efficient analysis (Hoch and Hayes 2010; Luebbering 2013). Luo *et al*. (2018) attempted to quantify differences in linguistic forms and measure their geographical correlation, but did not reflect on linguistic knowledge. They did not distinguish between loanwords and directly inherited words. This study is new in that it will utilize the database function of GIS while using the traditional comparative reconstruction method.
5) Thus far, language data have been represented by points on a map, which is actually inadequate for considering contact and propagation between languages distributed on a plane. On the contrary, it is difficult to accurately grasp the range of language distribution, and there is no precedent for research linking language data to accurate linguistic analysis using GIS. Relatedly, it is difficult to combine the qualitative data of language form with the quantitative calculation tool of GIS, which may also be a fundamental reason that GIS has not been used as a linguistic analysis tool until now.
6) When using GIS as a tool in linguistics, the output of visualized distribution information is analyzed according to our visual expertise. This study is unique in that it will identify whether the same conclusion can be obtained when the same data are mathematically processed by the same principle, and it will attempt to verify the results.
7) In linguistic data analysis in statistics, machine-readable data sets have been published in recent years, and it is believed that they accelerate considerable research. This study will also contribute to statistical research in that it will provide a data set to quantitatively test hypotheses about spatiotemporal variation.

(3) Significance and Necessity of International Collaborative Research

Collaboration that spans the three fields of historical linguistics, geography (especially GIS), and statistical mathematics, that are required in this study, is possible only in international collaborative research.

1) Data preparation by using The University of the South Pacific as a base will maximize the collection and monitoring of Fijian language data that are analyzed and will enable quality control. Additionally, to obtain information on non-linguistic information from the Fijian government, geographical bases are required.
2) The establishment of a base geographical system will create a collaborative structure with Massey University, which is known for its GIS research. By receiving feedback from related experts centered on research collaborator Lowry, it will be possible to have content that reflects research results using the latest GIS. In addition, collaborating with these institutes can accelerate the writing of accomplishment reports through international collaborative writing, and having young researchers and graduate students actively involved provides a basis for each of them to develop their own research into international joint research in the future.

(4) Past Research Activity (as of September 2017)

1) Kikusawa has been promoting comparative research on Austronesian languages since 1995. In addition to the comparison of sound correspondences and vocabulary, progress has been made in the development of methods for syntactic structures and the development history of sign language.
2) From 2002 to 2003, through collaborative research with Satoshi Kinugasa, Madagascar dialects were plotted and analyzed on a map using GIS, and the results were reported to the International Historical Linguistics Society conference (Kikusawa et al. 2005).
3) From 2013 to 2018, regarding the relationship between the limit of the phylogenetic tree model and the wave model, an international symposium was held to deepen discussions and was published as a collection of papers.
4) Mochihashi and Murawaki discussed how to combine statistical models and linguistics in addition to the possibility of modeling geographical information and language change in time at The Institute of Statistical Mathematics Sciences' FY2017 inter-institutional collaboration and literature and science fusion project titled "Lineage, variation, and diversity in language and its mathematics."
5) From FY2017 to FY2018, a GIS-based project for location information of Fiji dialects was promoted. We plan to report on the research as an international collaboration study at the International Austronesian Linguistics Society (July 2018) and the Australasia Geography Society (July 2018).

(5) Preparation Status and Feasibility (as of September 2017)

The preparation is as follows, and there are no problems with feasibility.

【Language data】Two 100-word lists in 300 dialects are planned to be completed in GIS by the end of 2018. For the distribution data of approximately 5,800 hand-written vocabulary and morphemes at 90 sites, the method of data collection has been considered. In addition,

five of the members have conducted fieldwork in Fiji, which provides a basis for collecting additional data as needed.

【GIS】The geographical information of Fijian dialects already obtained from the government in Fiji is currently being developed into GIS data and is expected to be completed by the end of March, 2019. Additionally, preliminary research on the correlation between language and non-linguistic information has begun targeting Kadavu Island.

【Mathematical modeling】Mochihashi, a research collaborator, has been researching geographical information of languages, and Murawaki has been conducting research on mathematical models related to phylogenetic relationships and time changes; hence, they have the bases to analyze the data necessary for this project.

References

Fujii, R., R. Domoto, and D. Mochihashi
2017 Nonparametric Bayesian semi-supervised word segmentation. *TACL* 5: 179–189.

Gal, Y. and Z. Ghahramani
2016 Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. arXiv:1506.02142v6. (Last access February 7, 2022)

Geraghty, P. A.
1983 *The History of the Fijian Languages* (Oceanic Linguistics Special Publication 19). Honolulu: University of Hawai‘i Press.

Hoch, S. and J. J. Hayes
2010 Geolinguistics: The Incorporation of Geographic Information Systems and Science. *The Geographical Bulletin* 51(1): 23–36.

Kikusawa, R.
2002 *Proto Central Pacific Ergativity: Its Reconstruction and Development in the Fijian, Rotuman and Polynesian Languages* (Pacific Linguistics 520). Canberra: Pacific Linguistics.
2015 The Austronesian Language Family. In C. Bowern and B. Evans (eds.) *The Routledge Handbook of Historical Linguistics*, pp. 657–674. London: Routledge and Taylor & Francis.
2018 What the Tree Model Represents: Language Change, Time Depth, and Visual Representation. In R. Kikusawa and L. A. Reid (eds.) *Let's Talk about Trees: Genetic Relationships of Languages and Their Phylogenic Representation* (Senri Ethnological Studies 98), pp. 171–193. Osaka: National Museum of Ethnology.

Kikusawa, R. and S. Kinugasa
2005 An Application of GIS to Historical Linguistics. Paper presented at the 17th International Conference on Historical Linguistics (ICHL17), Wisconsin, August 2, 2005.

Luebbering, C. R.
2013 Displaying the Geography of Language: The Cartography of Language Maps. *The Linguistics Journal* 7(1): 39–67.

Luo, W., J. Hartmann, F. Wang, H. Pingwen, V. Sysamouth, J. Li, and X. Cang
2018 GIS in Comparative-historical Linguistics Research: Tai Languages. In B. Huang (ed.)

Comprehensive Geographic Information Systems, pp. 157–180. Amsterdam: Elsevier.

Murawaki, Y.

2015a Spatial Structure of Evolutionary Models of Dialects in Contact. PLOS ONE 10(7): e0134335. DOI: 10.1371/journal.pone.0134335

2015b Continuous Space Representations of Linguistic Typology and Their Application to Phylogenetic Inference. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 324–334. DOI: 10.3115/v1/N15-1036

2017 Diachrony-aware Induction of Binary Latent Representations from Typological Features. In G. Kondrak and T. Watanabe (eds.) *Proceedings of the Eighth International Joint Conference on Natural Language Processing* (*Volume 1: Long Papers*), pp. 451–461. Taipei: Asian Federation of Natural Language Processing.

Murawaki, Y. and K. Yamauchi

2018 A Statistical Model for the Joint Inference of Vertical Stability and Horizontal Diffusibility of Typological Features. *Journal of Language Evolution* 3(1): 13–25. DOI: 10.1093/jole/lzx022

Nakamura, T., N. Takayuki, D. Mochihashi, I. Kobayashi, H. Asoh, and M. Kaneko

2017 Segmenting Continuous Motions with Hidden Semi-Markov Models and Gaussian Processes. *Frontiers in Neurorobotics* 11: 67.

Noji, H., D. Mochihashi, and Y. Miyao

2013 Improvements to the Bayesian Topic N-gram Models. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* pp. 1180–1190. Seattle, Washington, USA, 18–21 October 2013.

Ohishi, Y., D. Mochihashi, H. Kameoka, and K. Kashino

2014 Mixture of Gaussian Process Experts for Predicting Sung Melodic Contour with Expressive Dynamic Fluctuations. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 3714–3718. DOI: 10.1109/ICASSP.2014.6854295

Pawley A. and T. Sayaba

1971 Fijian Dialect Divisions: Eastern and Western Fijian. *The Journal of the Polynesian Society* 80(4): 405–436.

Shinozaki, T., S. Watanabe, D. Mochihashi, and G. Neubi

2017 Semi-supervised Learning of a Pronunciation Dictionary from Disjoint Phonemic Transcripts and Text. *Interspeech 2017*, pp. 2546–2550. Stockholm: The International Speech Communication Association.

Uchiumi, K., H. Tsukahara, and D. Mochihashi

2015 Inducing Word and Part-of-Speech with Pitman-Yor Hidden Semi-Markov Models. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (*Volume 1: Long Papers*), pp. 1774–1782. Beijing: Association for Computational Linguistics.

Yamauchi, K. and Y. Murawaki

2016 Contrasting Vertical and Horizontal Transmission of Typological Features. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 836–846. Osaka: The COLING 2016 Organizing Committee.

Yoshioka, N.

2016 Domaaki as a Language of Northern Pakistan: From a Geolinguistic Point of View. *Papers from the Third International Conference on Asian Geolinguistics*, pp. 38–45. Tokyo: Research Institute for Languages and Cultures of Asia and Africa (ILCAA).

2017 Nominal Echo-Formations in Northern Pakistan. *Bulletin of the National Museum of Ethnology* 41(2): 109–125.

2018a Tone/accent in South Asia (Aryan, Iranian, Nuristani, Dravidian, Andamanese, and Isolates). In S. Shirai and M. Endo (eds.) *Studies in Asian Geolinguistics VIII*, pp. 19–20. Tokyo: Research Institute for Languages and Cultures of Asia and Africa (ILCAA), Tokyo University of Foreign Studies.

2018b It rains: South Asia (IE [Aryan, Iranian], Dravidian, Andamanese, and Burushaski). In S. Shirai and M. Endo (eds.) *Studies in Asian Geolinguistics VIII*, pp. 48–49. Tokyo: Research Institute for Languages and Cultures of Asia and Africa (ILCAA), Tokyo University of Foreign Studies.