Conducting Research on the Geographical Linguistics by Utilizing the Data Comprising Twitter Postings

Shinsuke Kishie[1], Shuichi Matsunaga[2], Takashi Kirimura[3], Shin Abe[4], Kota Hattori[5], Yukako Sakoguchi[6]
[156]Tokushima University, [2]Jumonji University, [3]The University of Tokyo, [4]Nagoya University of Foreign Studies

Background

・Social Networking Sites (SNS) have become a common tool for our daily communication. They have also become a powerful tool for researchers to investigate various issues in language variation. (e.g., Mocanu et al., 2012; Gonçalves and Sánchez, 2014)

・Geo-tagged SNS data may not reflect dialectal variations within a language well, because the SNS data's mobility may lead to mismatch between the distribution of the geographical data and the location of where a speaker's dialect is spoken.

・In this study, we focused on a specific word, '**driving school**', whose dialectal variations are already known by former survey data **(Fig.1)**, and we examined whether dialectal variations in Twitter data correspond with our survey data.

Data

This study is based on these data including the word, *driving school*.

・**Twitter data** : 297,393,787 geo-tagged Japanese tweets including the word, 'driving school' which were collected from February 2012 to January 2016.

・**Survey data** : A nationwide survey data. We have conducted this survey with 4,838 university students from January 2014 to December 2015.

Methods

・Focusing on the dialectal variations of the word, 'driving school', we calculated the proportion of the number of tweets including each dialectal form per prefecture dividing by total number of tweets in each prefecture.

・We also calculated the proportion of the number of students using each dialectal form per prefecture dividing by the total number of students in each prefecture.

・We created choropleth maps for each dialectal form and compared the locations of SNS points (where Twitter users tweeted) and survey data points (where students were born).

Results

- **Over all :** Having compared maps from **Fig.3a** to **Fig.7b**, it seems that the distribution based on Twitter data and our survey data are generally matched.
- ***JIDOSHAGAKKO*** is used all over the places in Japan. An official name, ***JIDOSHAKYOSHUJO*** is distributed mainly in the Kanto, Kinki and Shikoku regions. **(Fig.2/3)**
- There are various abbreviated forms of ***JIDOSHAGAKKO*** and ***JIDOSHAKYOSHUJO*** throughout Japan. Almost of them have geographical distribution.
  - (a) ***JISHAKO*** appears in the Tohoku region. **(Fig.4)**
  - (b) ***SHAGAKU*** appears in Niigata and Kochi Prefecture. **(Fig.5)**
  - (c) ***SHAKO*** is widely distributed in Western Japan, mainly the Kyushu, Chugoku and Shikoku region. In addition to these area, it is also distributed in Western Chubu and Hokuriku region. **(Fig.6)**
  - (d) The very unique form, ***JIREN***, appears only in Okinawa Prefecture. This is an abbreviated form of ***JIDOSHARENSHUJO*** [ʒidoːʃareɴʃɯːʒo]. **(Fig.7)**

However, the distribution of ***JIDOSHAGAKKO***, standard form analyzed from Twitter data does not necessarily match with our survey data, because it is commonly used throughout Japan. **(Fig.2)**

Linguistic forms with local attributes show regional distribution where, linguistic forms without the attributes show nationwide distribution.

These results suggest that SNS data is valid and reliable information resource to further investigate issues in sociolinguistic studies.

Future Research

We will reveal the distribution of other words considered as standard language forms in Japan by comparing Twitter data and survey data.

SNS data like Twitter data have a great potential to improve the modern Japanese studies. We will explore future possibilities and promote our study, and disseminate the results into the all over world.

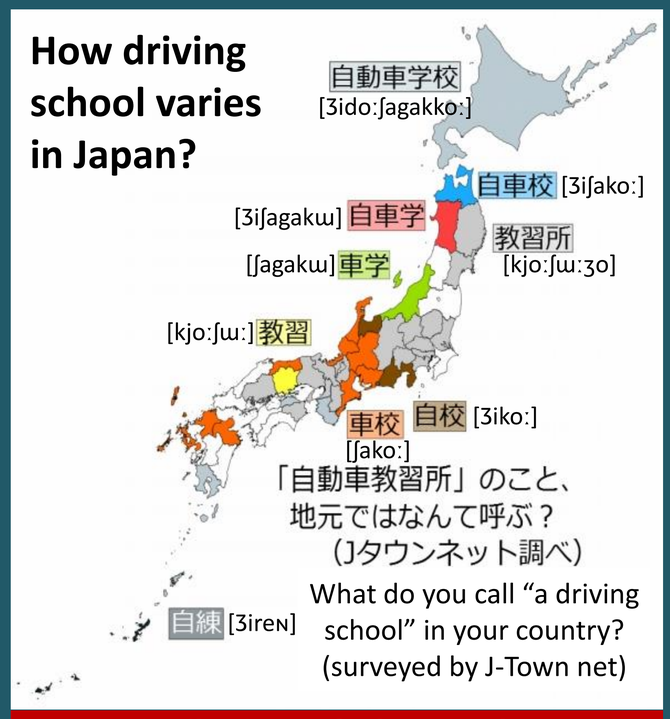

Fig.1 Distribution of the word meaning "driving school" based on J-Town net survey*1

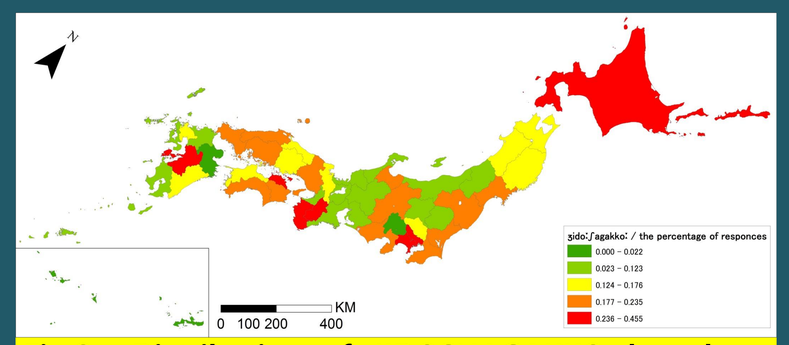

Fig.2a Distribution of JIDOSHAGAKKO based on our survey data

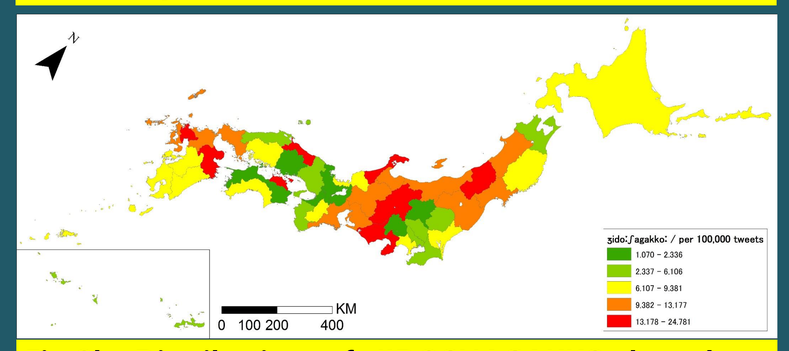

Fig.2b Distribution of JIDOSHAGAKKO based on Twitter data

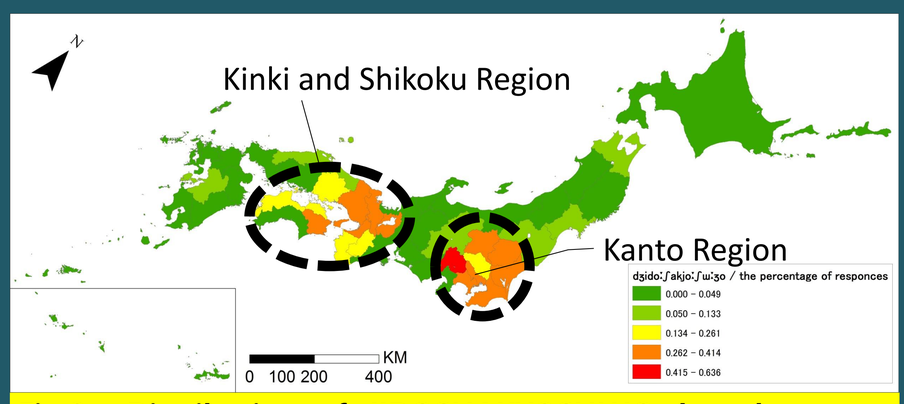

Fig.3a Distribution of JIDOSHAKYOSHUJO based on our survey data

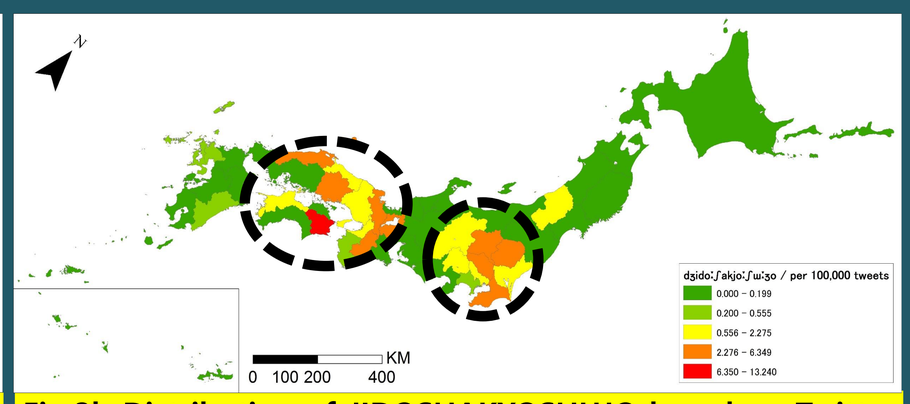

Fig.3b Distribution of JIDOSHAKYOSHUJO based on Twitter data

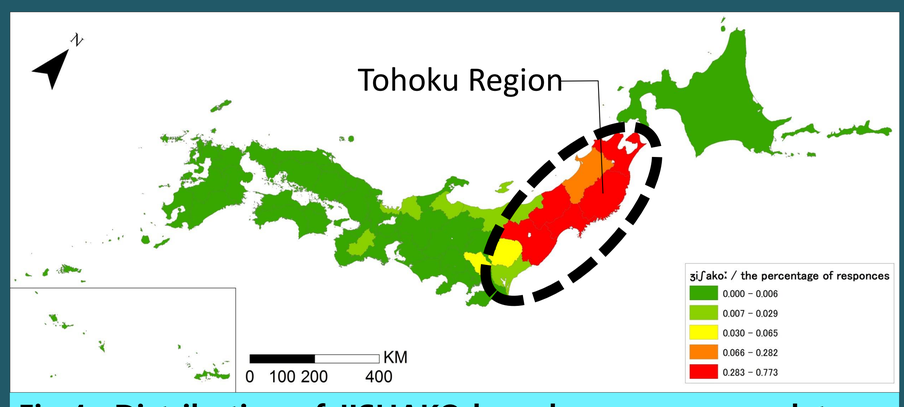

Fig.4a Distribution of JISHAKO based on our survey data

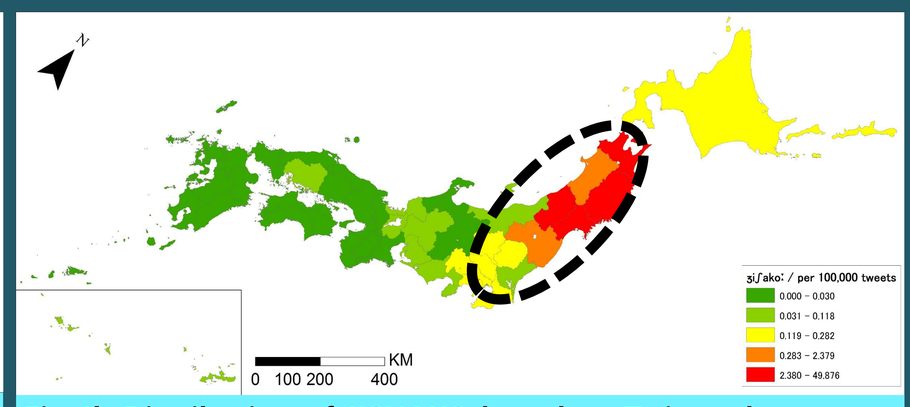

Fig.4b Distribution of JISHAKO based on Twitter data

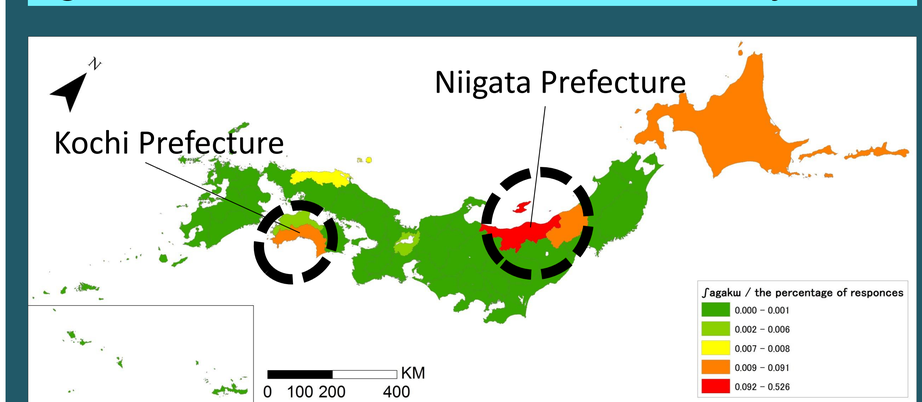

Fig.5a Distribution of SHAGAKU based on our survey data

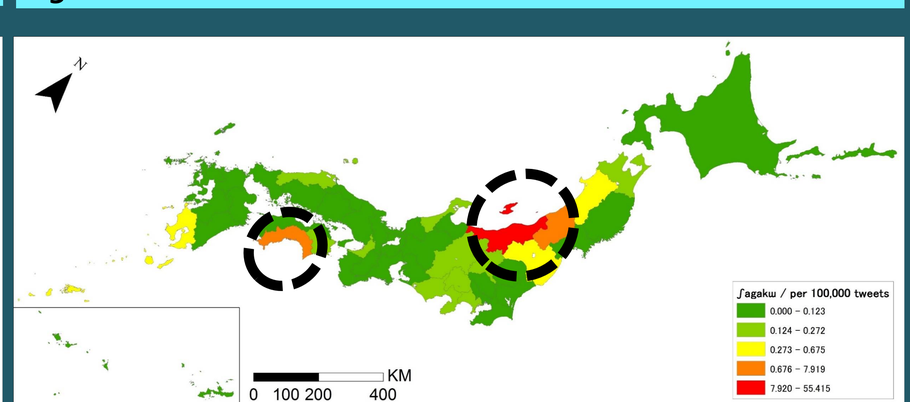

Fig.5b Distribution of SHAGAKU based on Twitter data

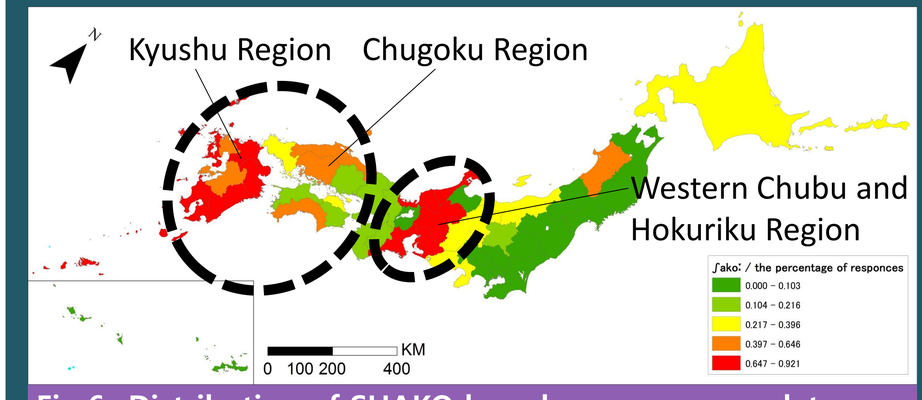

Fig.6a Distribution of SHAKO based on our survey data

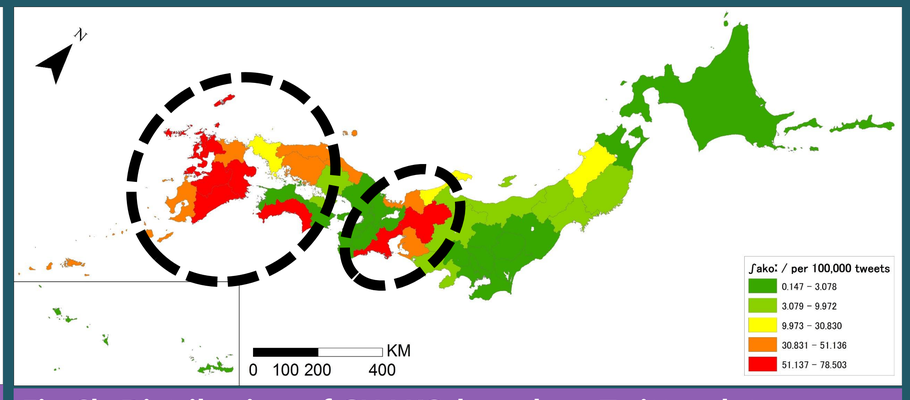

Fig.6b Distribution of SHAKO based on Twitter data

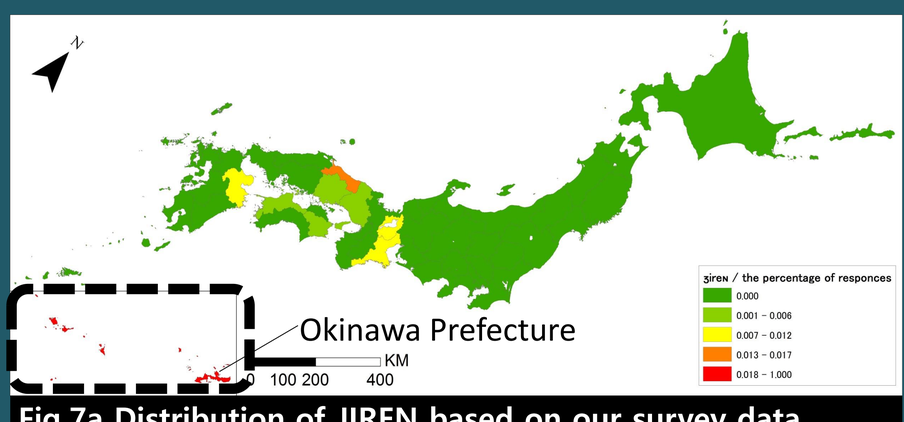

Fig.7a Distribution of JIREN based on our survey data

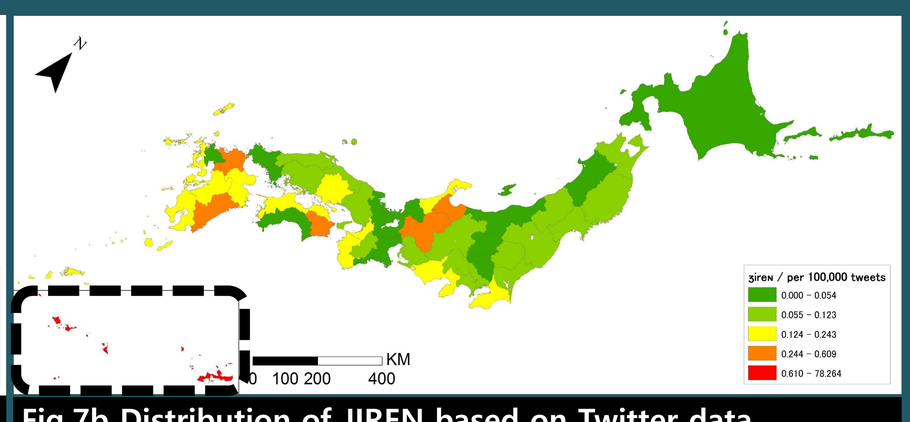

Fig.7b Distribution of JIREN based on Twitter data

This work was supported by JSPS KAKENHI Grant Numbers 15H01712.

*1 http://j-town.net/tokyo/research/results/123976.html?p=all（2016/4/13 ACCESS）