How fast would it be?
- Observing the distributions of emerging words through Twitter

Kota Hattori[1], Shinsuke Kishie[2], Takashi Kirimura[3], Yukako Sakoguchi[4], & Nanami Shiokawa[5]

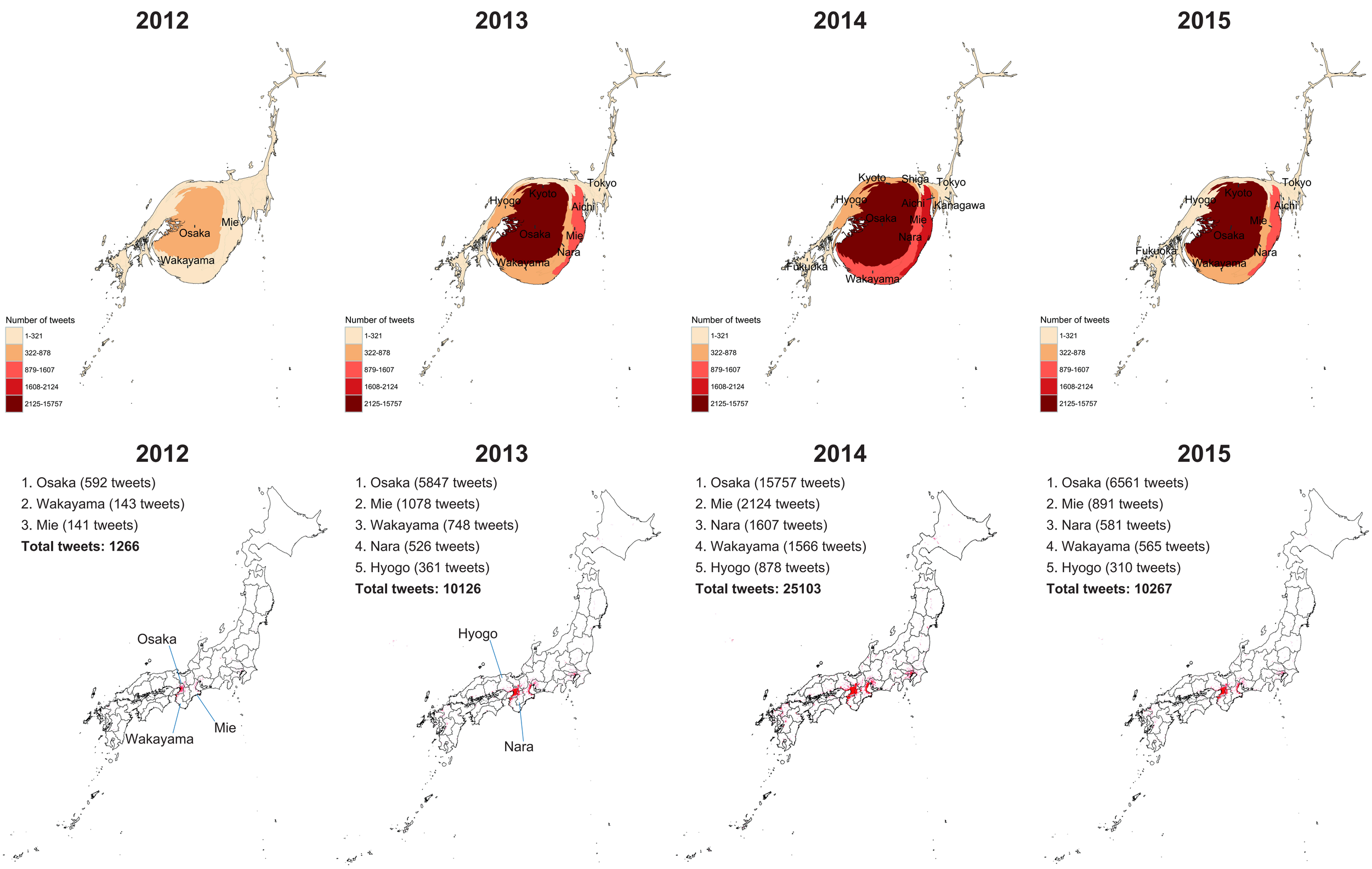
Tokushima University[1,2,4,5], University of Tokyo[3]

Introduction

The present study examined whether we could observe how newly emerged words from a local dialect spread using Twitter. Specifically, we observed how ***koyan***, which means 'not come' , has spread so far in Japan. We collected data between November, 2012 and December, 2015. We cleaned up our data with R and manual checks, leaving approximately 47,000 tweets. The results demonstrated that the word has been heavily used in Osaka area, where the word emerged. The results also demonstrated that the word has been spreading to adjacent areas as well as distant areas.

Background

It is belive that ***koyan*** is based on a traditional local word in Mie and Wakayama prefectures (*kiyan*), which means 'not come' . They say *kiihen* for 'not come' in Osaka dialect. Young folks in Osaka somehow chose the negation part (*-yan*) from the Mie/Wakayama dialect and added it to the verb, 'come' (*ki*). This is how ***koyan*** emerged in Osaka area.



2012

1. Osaka (592 tweets)
2. Wakayama (143 tweets)
3. Mie (141 tweets)

Total tweets: 1266

2013

1. Osaka (5847 tweets)
2. Mie (1078 tweets)
3. Wakayama (748 tweets)
4. Nara (526 tweets)
5. Hyogo (361 tweets)

Total tweets: 10126

2014

1. Osaka (15757 tweets)
2. Mie (2124 tweets)
3. Nara (1607 tweets)
4. Wakayama (1566 tweets)
5. Hyogo (878 tweets)

Total tweets: 25103

2015

1. Osaka (6561 tweets)
2. Mie (891 tweets)
3. Nara (581 tweets)
4. Wakayama (565 tweets)
5. Hyogo (310 tweets)

Total tweets: 10267

Figure1. Cartograms show which prefectures had large proportion of tweets. For example, Osaka had the largest proportion of tweets, and that is why its territory got expanded a lot. Dot maps show where Twitter users tweeted with ***koyan***. In general, ***koyan*** has rapidly spread in Kansei area (i.e., Osaka, Hyogo, Wakayama, and Kyoto) and Mie prefecture. The word seems to have been gradually spreading to some distant areas such as Aichi, Tokyo, and Fukuoka. Note that prefecture names appearing in the cartograms have more than 100 tweets with ***koyan***.

Data

We have been collecting geo-tagged Twitter data all the time since February 2012 with a PHP script. After processing the data, we had approximately 47,000 tweets including ***koyan***.

Data processing

We inspected all tweets using the following three patterns and wrote an R script with regular expressions to identify Tweets with ***koyan***. Then, we manually checked them.

- One *hiragana* + ***koyan*** (each hiragana + ***koyan***)
  e.g., grep(pattern = "あこやん" , x = mydf$tweets, value = TRUE))
- *Katakana* + ***koyan***
  e.g., grep(pattern = "[\u30A0-\u30FF] こやん ", x = mydf$tweets, value = TRUE)
- *Kanji* + ***koyan***
  e.g., grep(pattern = "[\u4E00-\u9FFF] こやん ", x = mydf$tweets, value = TRUE)
- Te + *koyan* & de + ***koyan***
  e.g., grep(pattern = " でこやん | てこやん ", x = mydf$tweets, value = TRUE)

Challenges for data processing

- Morphological analysis
  Tweets are written in very casual or broken Japanese (e.g., omitting postpositions and using slangs).Therefore, it is hard to parse strings and obtains words.
- Ambiguity
  When tweets are written in *hiragana*, we cannot determine whether we can keep tweets in our data. For example, あやこやん (***ayakoyan***) can be either 'Aya does not come.' or 'Oh that is Ayako.'
- Misspelling
  Some people seem to purposely use misspelling (e.g., 今 日 わ こ や ん). This requires manual checks.

Potential solution

We could develop customized filters with regular expressions based on real tweets. In a long run, this will allow us to filter tweets with good accuracy. We need to constantly update customized filters.

Conclusion

We demonstrated that ***koyan*** has been spreading to some areas in Japan, particularly in Mie-Aichi area, Tokyo area, and Fukuoka area. Yet, the number of tweets in these areas is still small. This may be because many tweets are not geo-tagged. We will keep collecting tweets and track down how ***koyan*** will further spread. We will also examine how other verb forms with the negation from, *-yan* (e.g., *miyan*, which means 'not see') spread through Japan. This would reveal how the negation form would be spreading in Japan.

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