(submit in MS Word form)

Form6

Grant No.	２０－Ｇ５２

Research Report

Name: Maria Telegina

Organization (at the start of the grant period)**:**

Tokyo University

Title of Research:

Towards an Understanding of the Fundamental Domains of Japanese Mental Lexicon.
Why hima (‘free time’) is bad and ushiro (‘back’) is scary?

Purpose of Research:（200 words）

This project investigates culturally and socially specific features of connections between the concepts in the Japanese mental lexicon based on the crowd-sourced study on word relations and word similarity. The resulting dataset of this study will consist of the descriptions of relations and similarity scores of pairs of associations or words co-occurring in corpus produced by the volunteers on a crowd-sourcing platform. The resulting dataset will have a potential for implementation in a number of fields – from lexicography and language learning to natural language processing and understanding. I expect this project to contribute to the global discussion on the mental lexicon and to play an important role in forming a basis for science-supported Japanese language learning and teaching. This study will not only allow detection of semantic relations and similarity within the data set validated by native speakers but also will create a novel form of public outreach and help to form a community of Japanese native-speaking volunteers interested in research on the Japanese language.

Content/Methodology of Research:（400 words）

This study was conducted as a people-powered study on semantic relations, aiming to obtain information on word similarity and types of relations. The page of the study and survey work-flows – one on similarity, another one on word relations were created via the online citizen science platform Zooniverse. As the first step of the project, a set of pairs of word associations was uploaded to the platform. The first set consisted of associations tagged as culturally specific together with hapax legomena (individual associations) from my doctoral project dataset. As the study progressed new association pairs were added using the newly processed data from the Japanese version of the Small World of Words project - a free-word association data collection project we launched in parallel with this project. In total, more than 2000 pairs of word associations were added to the data set and annotated by the volunteers. The previous studies(e.g., Vankrunkelsven et al., 2018) suggest that the simultaneous analysis of word association data and co-occurrence-based data gives the fullest picture of the mental lexicon. Therefore, further in the project words frequently co-occurring with the stimuli will be also added to the stimuli set from the corpus.

Participants, volunteering to take part in the study first choose the workflow they want to participate in – word similarity or word relations, then they read through a tutorial explaining the process of the annotation. They are also instructed that they can choose a number of pairs of words they are willing to annotate.　In word similarity workflow, the participants are shown one word pair at a time and asked to evaluate their similarity on a scale from 0 (not similar) to 6 (very similar). In the relations workflow, participants are shown one word pair at a time and asked to choose or to describe the relations between two words. The set of choices that participants are given consists of 12 options: similar meaning (synonym), opposite meaning (antonym), part, material, actions, superior concept, subordinate concept, environment, attribute, culturally specific concept, other (with an option to describe in their own words), not known.

The study is still ongoing, but the resulting data are in the process of qualitative analysis and quantitative analyses (De Deyne et al., 2016).

De Deyne, S., Perfors, A., & Navarro, D. J. (2016). Predicting human similarity judgments with distributional models: The value of word associations. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1861–1870.
Vankrunkelsven, H., Verheyen, S., Storms, G., & De Deyne, S. (2018). Predicting lexical norms: A comparison between a word association model and text-based word co-occurrence models. *Journal of Cognition*, *1*(1).

Conclusion/Observation （200 words）

To conduct the study, I have created two resources: a data collection page (https://www.zooniverse.org/projects/mariatelegina/go-to-go) and a Facebook page (https://www.facebook.com/kotoba.no.kankei) to promote the project, discuss it with the citizen scientists, and publish updates and news.

The study is still ongoing, but in the process of the data collection 3 stimulus sets (2,191-word pairs in total) were annotated, and the community of the citizen science annotators was formed. On the data collection page, 43 registered volunteers and unregistered volunteers made 68, 480 classifications and completed annotation of both similarity measure and word relations for 2,191 subjects (Fig. 1).

Overall, 132,111 people have seen promotion materials from the Facebook page at least once. Most participants (around 60%) are female participants. The main age ranges of the participants are from the mid-20s to mid-30s and from mid-30s to mid-40s.

I am currently analysing the word relations responses for the first stimulus set (pairs of most and least frequent word associations from my doctoral project, 1000 pairs of time and space-related vocabulary). The total number of responses is 9500. The most frequent relations detected in this data set are attribute, part, environment, synonymity, and subordination.

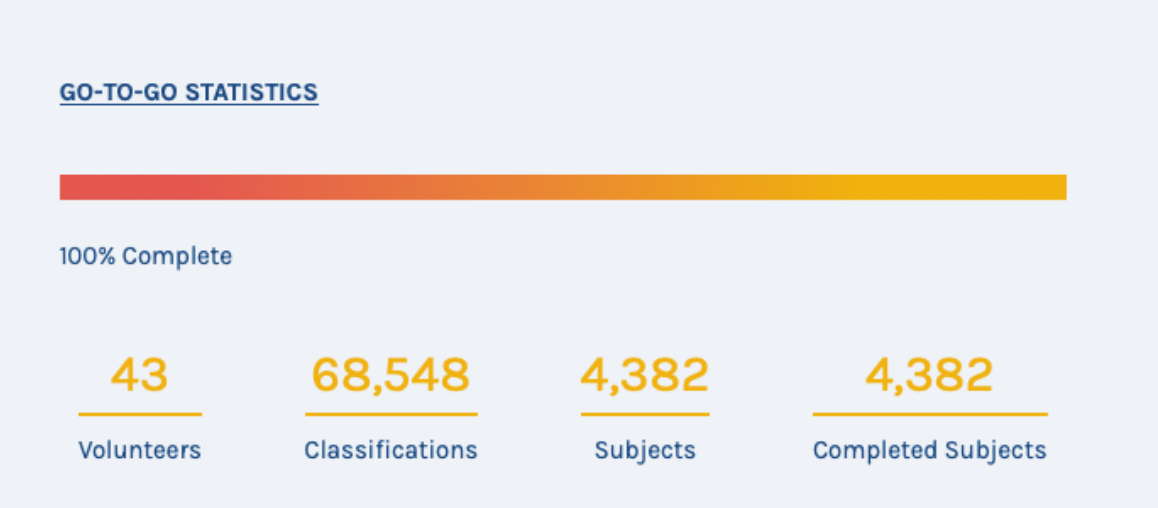

Fig. 1
https://www.zooniverse.org/projects/mariatelegina/go-to-go on Dec 22, 2021