Development and Validation of the Japanese Moral Foundations Dictionary

Akiko MATSUO[1], Kazutoshi SASAHARA[2,4], Yasuhiro TAGUCHI[3], and Minoru KARASAWA[2]
([1]Grad. School of Environmental Studies, Nagoya Univ., [2]Grad. School of Informatics, Nagoya Univ., [3]Aichi Univ., [4]JST PRESTO)
Key Words: Morality, Culture, Language

Abstract

The Moral Foundations Dictionary (MFD) is a useful tool for applying the conceptual framework developed in Moral Foundations Theory and quantifying the moral meanings implicated in the linguistic information people convey. Because the applicability of the MFD is limited because it is available only in English, translated versions of the MFD are therefore needed to study morality across various cultures, including non-Western cultures. Therefore, we developed the first Japanese version of the MFD (referred to as the J-MFD) using a semi-automated method—this serves as a reference when translating the MFD into other languages. We next tested the validity of the J-MFD by analyzing open-ended written texts about the situations that Japanese participants thought followed and violated the five moral foundations. We found that the J-MFD correctly categorized the Japanese participants' descriptions into the corresponding moral foundations, and that the Moral Foundations Questionnaire scores correlated with the frequency of situations, of total words, and of J-MFD words in the participants' descriptions for the Harm and Fairness foundations. Further, we applied the J-MFD by analyzing social media data written in English and Japanese. We found that English users are more likely to talk about immorality-related topics online than Japanese users and that English and Japanese users emphasize different moral categories. The J-MFD helps to explore morality unique to Japanese and can be useful for investigating cultural differences in moral behaviors.

Introduction

Currently, one of the most active research areas in social and behavioral sciences pertains to how and on what grounds ordinary people form moral judgments, focusing on their intuitions. This intuitionist model of moral judgment has produced voluminous empirical research as well as a comprehensive theoretical framework—this is now formulated as the Moral Foundations Theory (MFT; Graham, Haidt, & Nosek, 2009).The central principle of the MFT is that people inherit a limited number of conceptual templates used for their intuitive classification of observed acts that are potentially relevant to morality. Specifically, it is assumed that there are five major moral foundations including:

(1) "Care," which focuses on not harming others and protecting the vulnerable;
(2) "Fairness," which assumes equivalent exchange without cheating to be good;
(3) "Ingroup," which concerns a collective entity instead of individuals, such as family, nation, team, and military;
(4) "Authority," which postulates respect for authority, resulting in maintaining the hierarchy; and
(5) "Purity," which involves a feeling of disgust caused by the impure.

The MFT emphasizes that moral foundations meet not only individuals' adaptive need to fit into their community in the "correct" ways but also a collective need for the community to increase its unity and win against other groups, which allows the MFT to achieve a high level of consensus among the community members. The consensual nature of moral foundations should manifest most visibly in linguistic communication—this can mobilize the community toward solidarity and sanctity. In this group process, moral foundations are assumed to show their political aspects and provide a base for mobilizing members toward their collective goals. To test this hypothesis concerning "political" consensuality, Haidt and colleagues analyzed morality-relevant discourse in daily

contexts (Graham, Haidt, & Nosek, 2009).

A tool developed for this purpose was the Moral Foundations Dictionary (MFD), which quantifies virtues and vices associated with each moral foundation expressed in written texts. The MFD contains a list of words related to one or several moral foundations such as "killing," "justice," and "loyal," which correspond respectively to the Care, Fairness, and Ingroup foundations. Even though the MFD is useful for linguistic analyses of moral foundations, the dictionary is currently available only in English, and thus it is unknown to what extent we can generalize the findings to the linguistic communities outside of the English-speaking world.

This is a problem because it is plausible that the contents—as well as the roles of different moral foundations—would vary across cultures. The use of the MFD translated into different languages might reveal similar differences in communication and discourses. An accumulating body of evidence concerning cultural differences would also be important in the context of criticism about the potential bias in morality research that leans toward the so-called WEIRD (Western, Educated, Industrialized, Rich, and Democratic) cultural samples (Henrich, Heine, Norenzayan, 2010).

Furthermore, social media platforms provide an excellent arena for analyzing human behavior in a natural setting, and some recent research has successfully applied natural language processing (NLP) to social media data to quantify people's moral behaviors (Sagi & Dehghani, 2014). Unfortunately, the attempts to quantify people's online moral behaviors are also limited to English texts, because there is no publicly available dictionary that can be applied to texts written in languages other than English (Fulgoni, Carpenter, Ungar, & Preotiuc-Pietro, 2016).

To overcome this limitation, we describe here how we developed a Japanese version of the MFD (J-MFD) using a semi-automated method. The J-MFD is publicly available online, and hence our methodology can serve as a useful model for further attempts to develop moral dictionaries in other languages.

Methods

Development of J-MFD We translated the original MFD into our J-MFD via two online linguistic resources and two corpora with the aid of computational methods. The original MFD contains 324 English moral terms with 11 categories corresponding to "Virtue" or "Vice" (violates); each is associated with one of the five moral foundations (i.e., Care, Fairness, Ingroup, Authority, and Purity) as well with a more general or abstract category of morality (i.e., Morality General). Care is henceforth denoted as Harm in accordance with the notation of the MFD. The moral terms consisted of 156 words (e.g., impair) and 168 word stems (e.g., justifi*, which covers justification, justifier, etc.)

There were some words associated with multiple categories such as "impair" (Harm Vice and Purity Vice); other words were associated with only a single category, such as "justifi*" (Fairness Virtue alone).

Our development followed five steps (Fig. 1).

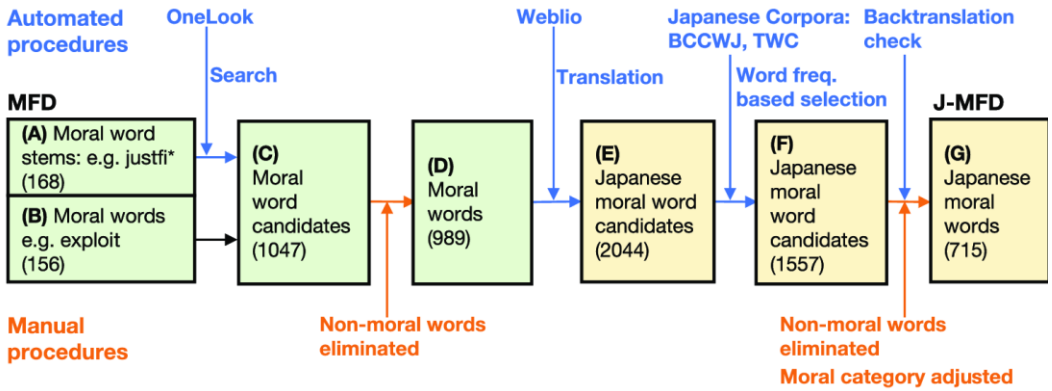

Figure 1. Overview of a semi-automated method for J-MFD development. Automated procedures are indicated by downward arrows, and manual procedures are indicated by upward arrows.

First, our programs automatically collected all words that contained each of the word stems in the MFD by web scraping OneLook—an online dictionary metasearch engine (https://www.onelook.com) (Fig.1A to C). Next, we manually eliminated 58 words that were unrelated to morality (Fig.1C to D). Third, the remaining words were translated into Japanese via Weblio—an online dictionary and encyclopedia designed for Japanese speakers (https://ejje.weblio.jp). This process was also performed by web scraping, which allowed us to cover possible translation equivalents in Japanese (2044 words) (Fig.1D to E). Fourth, we took a frequency-based approach for word selection using two Japanese corpora: Japanese words based on the Balanced Corpus of Contemporary Written Japanese (BCCWJ) and the Tsukuba Web Corpus (TWC). We adopted the top ten most frequent Japanese words for every word stem and the top five for every word in BCCWJ and TWC, thereby filtering out words rarely used (Fig.1E to F). Finally, we adjusted the category assignments for each Japanese word after removing words unrelated to morality and words that failed in back-translation using online dictionaries (Fig.1F to G). More specifically, we merged Japanese words whenever possible, using word stem representation. After the merge procedures, we examined whether Japanese moral word candidates can be back-translated to corresponding English words using online dictionaries. As a result, 23 words that failed this test were removed, and we were left with 741 Japanese moral terms, for which we adjusted the moral categories. This adjustment was necessary because multiple words (or word stems) with different moral categories could be translated into the same single Japanese word (or word stem); hence, a single Japanese word could belong to multiple categories. Among these categories, the central one (or ones) needed to be selected based on native Japanese knowledge and the definition of moral categories.

Validation of the J-MFD To validate our J-MFD, we compared the mean frequencies of the dictionary words for the five moral foundations that were included in the descriptions about moral issues reported by Japanese participants. More specifically, 386 Japanese participants (238 men and 148 women; M_{age} = 35.22, *SD* = 12.30) were recruited online using the Internet crowdsourcing service Macromill. Participants read brief explanations of Haidt's five moral foundations and listed as many situations as possible that they thought followed and violated the five moral foundations.

Relationships between Moral Descriptions and MFQ scores We collected self-reported responses to the Moral Foundations Questionnaire (MFQ; Graham, Nosek, Haidt, Iyer, Koleva, &

Ditto, 2011) from the same Japanese sample to examine relationships between self-reported moral situations and MFQ scores in Japanese. This is important to show the applicability as well as the further validation of the J-MFD.

To measure MFQ scores in Japanese, we used the 30-item version of the Japanese MFQ that was back-translated with the approval of the authors of the original MFQ (available at moralfoundations.org and in Kanai, 2013).

Our assumption here was that people who have a high MFQ score on a certain moral foundation (e.g., Harm) may have a better-organized schema for the corresponding foundation.
Thus, when asked to describe situations about the foundation, they could describe it more easily and appropriately than those who have a low MFQ score. This assumption can be tested by measuring for each of the five moral foundations, (1) how many situations they listed (Virtue and Vice combined), (2) how many words they used to describe the situations, and (3) how many foundation-related words they used to describe the situations. We investigated whether (1), (2), and (3) would correlate with MFQ scores for the corresponding moral foundations.

Cross-linguistic Comparisons using Twitter Data With the original MFD and our newly developed J-MFD, we compared the classification of tweets in English on the basis of the original MFD to that of Japanese tweets on the basis of our J-MFD to show an applicability of the J-MFD. We applied these dictionaries to tweets in respective languages that contained particular words, "immoral" and "immorality," posted on Twitter during the period between March 1 and April 24, 2016. We obtained 770,000 tweets in English and 340,000 tweets in Japanese. From these, we randomly sampled 50,000 Japanese and English tweets five times and compared the tweets in terms of the frequencies of dictionary words in the MFD and J-MFD.

We focused on Vice categories (i.e., Harm Vice, Fairness Vice, Ingroup Vice, Authority Vice, and Purity Vice) because the contents of the tweets in our datasets were related to immorality. For each Vice category, we normalized the occurrence frequency of dictionary words in the sampled tweets by dividing by the number of words in the tweets and the number of dictionary words in a given category in these dictionaries. Next, we averaged the occurrence frequencies based on five sampled datasets for English and Japanese, respectively.

Results

Japanese Moral Foundations Dictionary (J-MFD) Table1 shows the number of words for each category and the total number of words in the J-MFD.

Table 1. The number of words for each category in the J-MFD.

Harm Virtue	Fairness Virtue	Ingroup Virtue	Authority Virtue	Purity Virtue	Morality General
51	42	98	129	89	43
Harm Vice	Fairness Vice	Ingroup Vice	Authority Vice	Purity Vice	Total
93	33	43	53	87	718

Validation of the J-MFD The situations resulted in 16,033 sentences in total after eliminating the responses that were incomprehensible with respect to their meaning or were not related to morality

(e.g., “I can't understand the meaning of the question”). Each sentence in these descriptions of situations was segmented into words, and five pools of morally relevant words (i.e., Harm-related, Fairness-related, and so forth) were constructed. For each of these pools, we computed the frequency ratio of appearances of J-MFD words associated with each moral foundation. To take an example of the word pool produced from the Harm-related context (Virtue and Vice combined), we separately counted the numbers of times that the Harm-related words, Fairness-related words, and so forth contained in the J-MFD appeared in each participant’s descriptions. To obtain ratio scores, we divided those word counts by the size of the pool and by the total number of dictionary words associated with each moral foundation.

Figure 2 shows the mean frequency ratio for each foundation in the J-MFD obtained from the Harm-related word pool.

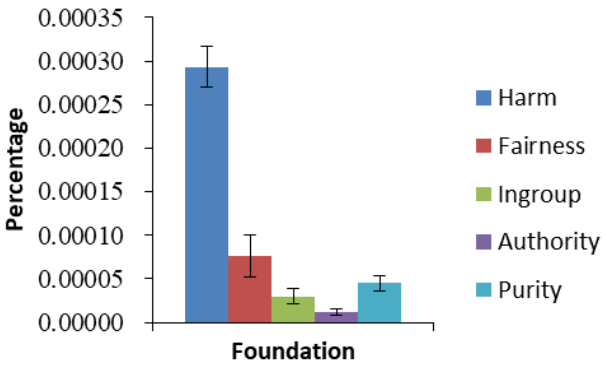

Figure 2. Mean percentages of dictionary words used in open-ended descriptions about the situations that the participants thought followed and violated the Harm foundation.

A one-way ANOVA showed a main effect of moral foundations, indicating a significantly higher frequency of Harm words than that of words from the remaining foundations. We repeated the same analyses for the remaining pools (i.e., Fairness-, Ingroup-, Authority-, and Purity-related), and similarly found the highest frequency ratios in each pool for the corresponding moral foundation, with the main effects of foundation ($F(4,1328) = 52.95$, $p < .001$). These results demonstrate the validity of our J-MFD.

Relationships between Moral Descriptions and MFQ scores As for (1), the correlation of MFQ scores and the number of situations described by participants was significant for most of the foundations (Harm: $r = .25$, $p < .01$; Fairness: $r = .16$, $p = .01$; Authority: $r = .14$, $p < .01$; Purity: $r = .22$, $p < .01$) except the Ingroup foundation ($r = .07$, $p = .19$). As for (2), the correlation of MFQ scores and the number of total words included in participant-made situations was significant for all five foundations (Harm: $r = .29$, $p < .01$; Fairness: $r = .20$, $p < .01$; Ingroup: $r = .12$, $p < .02$; Authority: $r = .15$, $p < .01$; Purity: $r = .22$, $p < .01$). As for (3), the correlation of MFQ scores and the number of J-MFD words included in participant-made situations was significant for the Harm and Fairness foundations ($r = .12$, $p < .02$; and $r = .11$, $p < .04$, respectively), while there was no significant correlation for the other three foundations (Ingroup: $r = -.02$, $p < .67$; Authority: $r = .04$, $p < .49$; Purity: $r = .06$, $p < .21$).

According to the results of (1)−(3), the correlation between self-reported moral descriptions by Japanese people and MFQ scores was consistent for the Harm and Fairness foundations but not for

the other foundations. The implications of this finding are discussed in the Discussion section.

Cross-linguistic Comparisons using Twitter Data An analysis of variance was performed separately for the Japanese and the English data sets. As shown in Figure 3, a highly significant main effect of moral foundation was found for Japanese ($F(4,16) = 2366.48$, $p < 0.001$) and English ($F(4,16) = 1651.53$, $p < 0.001$) tweets, indicating that both Japanese and English users emphasize different categories in online moral conversations.

More specifically, multiple comparison tests yielded the following results: Japanese users are more likely to use immorality-related words for Purity Vice than Harm Vice, Fairness Vice, Ingroup Vice, and Authority Vice (all $p < 0.001$) as well as for Ingroup Vice compared with Harm Vice, Fairness Vice, and Authority Vice (all $p < 0.001$), Harm Vice compared with Fairness Vice and Authority Vice ($ps < .05$), and for Authority Vice compared with Fairness Vice ($p < 0.001$). On the other hand, English users are more likely to use immorality-related words for Authority Vice than Harm Vice, Fairness Vice, Ingroup Vice, and Purity Vice (all $p < 0.001$) as well as Harm Vice compared with Fairness Vice, Ingroup Vice, and Purity Vice (all $p < 0.001$), Fairness Vice compared with Purity Vice ($p < 0.001$), and Ingroup Vice compared with Purity Vice ($p < 0.001$).

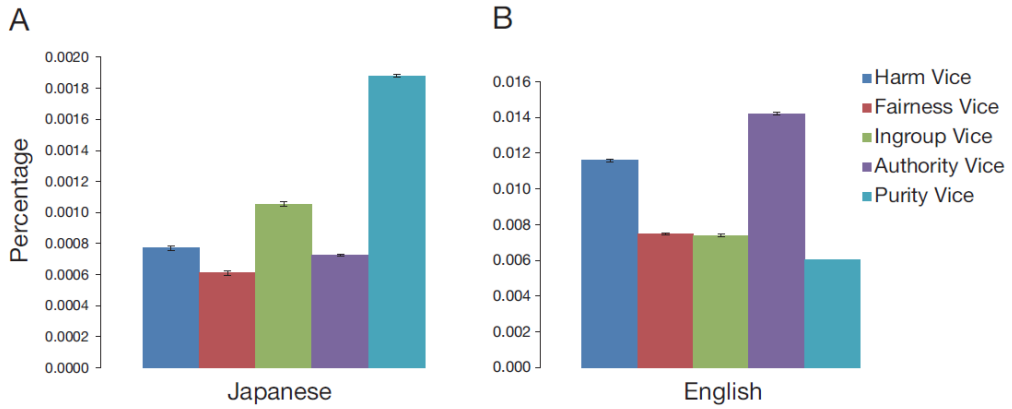

Figure 3. Percentages of all dictionary words for each category contained in Japanese and English tweets.

Discussion

This work proposed a semi-automated method for translating the Moral Foundations Dictionary (MFD) and developed and validated its Japanese version (J-MFD). The J-MFD will be updated with revision via collaborative efforts by its users (https://github.com/soramame0518/j-mfd). Our method is beneficial for developing other language versions of the MFD, which are needed because multilingual versions allow us to test the Moral Foundations Theory (MFT) in different languages and compare diverse cultures using the same basis of the MFT (Haidt, 2001).

We showed that the J-MFD allows us to correctly categorize moral-relevant situations in Japanese-written texts into the corresponding moral foundations, which serves as validation of the J-MFD. Furthermore, our correlation analyses showed that (1) the number of situations, (2) the number of words, and (3) the number of J-MFD words all consistently correlated with the MFQ score in the Harm and Fairness foundations, which implies the existence of a better-developed schema.

People were able to describe Harm- and Fairness-related sentences easily (i.e., (1) and (2)) and

accurately (i.e., (3)). However, such consistent patterns across (1) to (3) were not observed in the Ingroup, Authority, and Purity foundations, which suggests the lack of specific schema for these foundations. A possible explanation for these results is that Harm and Fairness foundations are more fundamental than the other foundations and are better quantified in the MFQ. In contrast, the Ingroup, Authority, and Purity foundations may be more culture-dependent, and the MFQ, as it stands, may inaccurately measure these foundations, and therefore, may need a modification specific to Japanese culture.

These interpretations are consistent with prior research findings, which show that Harm and Fairness are central to moral judgment cross-culturally (Haidt, Graham, & Ditto, 2015) while Ingroup, Authority, and Purity are susceptible to political ideology, ethnicity, culture, and religiosity (Bulbulia, Osborne, & Sibley, 2013). Altogether, while our study shows that the J-MFD is a valid tool for morality research in Japanese culture, it also highlights the need for further research to scrutinize the causal relationship of MFQ scores and the use of morality-relevant words in multilingual settings, not only in Japanese.

In addition, this work is the first step to comparing morality in a multilingual setting. Our cross-linguistic tweet analyses showed different patterns in Vice category usages: English users care more about Authority and Harm, whereas Japanese users care more about Purity and Ingroup. This result is consistent with Graham et al.'s study (Graham, Nosek, Haidt, Iyer, Koleva, & Ditto, 2011) that detected the greater concerns on Ingroup and Purity by Eastern (vs. Western) cultures. These findings suggest that different language users use different communication styles when communicating about morality-relevant topics. Our results stimulate such intriguing questions as the factors related to English users' sensitivity to social hierarchy (i.e., Authority). More importantly, the findings from Japanese tweets imply Japanese users' concerns about cleanliness and sacredness (i.e., Purity), which might suggest the uniqueness of Japanese culture.

This work contributes to interdisciplinary collaborations in morality research across academic fields. First, the MFT can be tested using NLP with the J-MFD by analyzing languages that people express online (Goolsby & Hunt, 1992) In addition to the word-counting approach conducted in this study, NLP allows the J-MFD to be used for word co-occurrence analysis (Dehghani et al., 2017) and latent semantic analysis (Sagi & Dehghani, 2014). Second, the J-MFD can be used in combination with a Japanese emotion dictionary because specific emotions are often associated with moral judgment (matsumoto2005). Third, morality matters in the business world as well; thus, it is important to investigate morality from the perspective of both leaders/employees in organizations and consumers (Goolsby & Hunt, 1992).

Finally, the J-MFD allows future research to scrutinize theoretical frameworks about the standards of moral judgment in various fields of study; this would enrich cross-cultural research by facilitating comparison of morality-related texts in English and Japanese. We are aware that the J-MFD—as well as the original MFD-is not an inclusive dictionary; thus, we had to combine Virtue and Vice categories for words with low occurrence frequency. Both dictionaries might be improved by adding more relevant words (Kaur & Sasahara, 2016). The expansion of moral words in the J-MFD is critical to improving accuracy in measuring morality from written texts—that is one of our key future goals.

References

Bulbulia, J., Osborne, D., Sibley, C. G. (2013). Moral foundations predict religious orientations in New Zealand. *PLoS ONE, 8*, e80224.

Dehghani, M., Johnson, K. M., Garten, J., Boghrati, R., Hoover, J., Balasubramanian, V., Singh, A. Shankar, Y., et al. (2017). TACIT: An open-source text analysis, crawling, and interpretation tool. *Behavior Research Methods, 49*, 538–547.

Fulgoni, D., Carpenter, J., Ungar, L., & Preot¸iuc-Pietro, D. (2016). An empirical exploration of Moral Foundations Theory in partisan news sources. In *LREC*, 3730–3736.

Goolsby, J. R., Hunt, S.D. (1992). Cognitive moral development and marketing. *The Journal of Marketing, 56*, 55–68.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96,* 1029-1046.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology, 101,* 366-385.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834.

Haidt, J., Graham, J., & Ditto, P. (2015). The Volkswagen of moral psychology. *Character & Context Blog*. Retrieved from http://www. spsp. org/news-center/blog/volkswagen-of-morality

Henrich, J., Heine, S.J., Norenzayan, A. (2010). Most people are not WEIRD. Nature, 466, 29–29.

Kanai, R. (2013) *Nou ni kizamareta moraru no kigen: Hito wa naze zen womotomerunoka [The origin of morality engraved in the brain: Why do peoplepursue goodness?]*. Iwanami Publisher.

Kaur, R., & Sasahara, K. (2016). Quantifying moral foundations from various topics on Twitter conversations. *Proceeding of the 2016 IEEE Big Data*, 2505-2512.

Sagi, E., & Dehghani, M. (2014). Measuring moral rhetoric in text. *Social Science Computer Review, 32*, 132–144.

Acknowledgments

This research was supported by JSPS/MEXT KAKENHI Grant Numbers 15H03446 and JP17H06383 in #4903, JST PRESTO Grant Number JPMJPR16D6, and JST CREST Grant Number JPMJCR17A4.