Proceedings of the 1st Workshop on Dataset Creation for Lower-Resourced Languages (DCLRL) @LREC2022, pages 19–26
Marseille, 24 June 2022

Construction and Validation of a Japanese Honorific Corpus Based on Systemic Functional Linguistics

Muxuan Liu, Ichiro Kobayashi
Ochanomizu University
{liu.muxuan, koba}@is.ocha.ac.jp

Abstract

In Japanese, there are different expressions used in speech depending on the speaker's and listener's social status, called honorifics. Unlike other languages, Japanese has many types of honorific expressions, and it is vital for machine translation and dialogue systems to handle the differences in meaning correctly. However, there is still no corpus that deals with honorific expressions based on social status. In this study, we developed an honorific corpus (KeiCO corpus) that includes social status information based on Systemic Functional Linguistics, which expresses language use in situations from the social group's values and common understanding. As a general-purpose language resource, it filled in the Japanese honorific blanks. We expect the KeiCO corpus could be helpful for various tasks, such as improving the accuracy of machine translation, automatic evaluation, correction of Japanese composition and style transformation. We also verified the accuracy of our corpus by a BERT-based classification task. We release our corpus KeiCO for further research: `https://github.com/Liumx2020/KeiCO-corpus/blob/main/keico_corpus.csv`.

Keywords: Japanese corpus, honorific level, systemic functional linguistics

1. Introduction

Japanese honorific or Keigo (敬語) is an expression of respect used in Japanese to indicate social rank, intimacy and other relationships among the speaker, the listener and the person mentioned in the conversation.(Aapakallio, 2021)

In many social situations in Japan, honorifics are necessary to express appropriate social status relationships and politeness. Japanese honorifics are generally divided into three categories: respectful (sonkeigo, 尊敬語), humble (kenjogo, 謙譲語), polite (teineigo, 丁寧語). In addition, depending on the content of the conversation and the listener, the speaker may use honorific prefixes, verb morphing, which forms two particular types of honorific: word beautification (bikago, 美化語), and courteous language (teichogo, 丁重語).

However, there is no corpus that contains detailed information on the language used by social groups, such as the situation of language use, social role relationships among interlocutors, and means of interaction. Therefore, it is not easy to construct a machine learning model that takes social factors into account and uses appropriate honorifics.

In this study, we attempt to construct and validate a Japanese honorific corpus (**KeiCO corpus**) which contains more detailed information on social factors based on systemic functional linguistics, which analyzes language from the viewpoint of language use in social groups. We will also make the constructed KeiCO corpus available as a language resource.

Our work has the following contributions.

- We contributed a corpus of 10,007 Japanese sentences. It is the first corpus about honorific sentences. The corpus is based on systemic functional linguistics and contains detailed information on the honorific level, the social relationship between the speaker and the listener, and conversational situations or topics. They filled in the honorific blanks of machine translation, dialogue system, and semantic analysis.
- On the base of our corpus, we took another step on analysis and we got some characteristics of honorific sentences. Through these characteristics, we can help people to better understand honorific sentences under natural circumstances.

2. Related Work

Because politeness is usually regarded as a style, the level of honorifics in Japanese can be thought of as several different politeness styles.

Recently, there is much research using machine learning to deal with the politeness of sentences. For example, Resmi and Naseer (2019) created a politeness classifier to classify responses as polite, rude or neutral. Niu and Bansal (2018a) proposed three weakly supervised models that could generate different polite (or rude) conversational responses in the absence of parallel data.

In addition, a lot of controllable natural language generation (NLG) research develop generation methods that incorporate various style transformations, such as length, politeness, perspective, descriptiveness, emotion, and so on (Tsai et al., 2021; Liu et al., 2022). Tsai et al. (2021) propose schema-guided NLG focusing on semantic stylistic control, and showed that disentangling context generation and stylistic variations is more effective at achieving semantic correctness and style accuracy. Liu et al. (2022) propose an Edit-Invariant Sequence Loss (EISL), which computes the matching loss of a target n-gram with all n-grams in the generated sequence, and shows the usefulness of EISL applying it to style transferred NLG.

In the task of polite style transformation for English, Madaan et al. (2020) introduce a new task of politeness transfer which involves converting non-polite sentences to polite sentences while preserving the meaning. They also provide a dataset of more than 1.39 million instances from Enron corpus following the same pre-processing by Shetty and Adibi (Klimt and Yang, 2004) , and assign politeness scores to those sentences by using a politeness classifier (Niu and Bansal, 2018b). There are some existing corpora of English that use manual annotation of the politeness and formality of utterances. For example, Rao and Tetreault (2018) has been studied to classify sentences into six levels of formality; and Danescu-Niculescu-Mizil et al. (2013) requires the annotator to indicate how polite she or he considers the request to be using a slider from "very impolite" to "very polite", normalized by a standard Z-score to obtain a definite value. However, there has been no work related to assigning ranks to Japanese honorifics and paraphrasing between the ranks. This is due to two reasons: (1) no corpus exists to support such work, and (2) existing NLP models are not mature enough in their treatment of honorifics. For example, Feely et al. (2019) classifies Japanese sentences into one of three levels of informal, polite or formal speech in parallel texts.

They used the NMT model to learn the difference in the degree of formality in Japanese by identifying honorifics in Japanese in parallel training data and by labelling the source language with additional features. This is a way to control the level of formality of Japanese output in English to Japanese Neural Machine Translation (NMT). However, by simply classifying sentences as informal, polite or formal speech, important linguistic information such as respectful speech and modest speech is lost, which does not help to improve the accuracy of the results after machine learning.

Among other honorific tasks, such as the task of judging the correctness of honorific use, Shirado et al. (2011) constructs a set of rules for evaluating the appropriateness of misused honorific expressions based on subject-verb-object and some grammatical features of honorifics to help judge the appropriateness of honorific use, which can help identify the social relationship between the speaker and the hearer, but still missing information about the different degrees of respect of honorifics. Due to the limitations of automatically extracting linguistic knowledge based on grammatical rules, it is impossible to obtain deep knowledge from a corpus with tags of shallow information or the original corpus. This also leads to challenges in data annotation or data collection due to the subjective nature of language style compared to other NLP tasks such as question answering (Xu, 2017). A possible solution is to assign deep information tags to the corpus depending on the intended use of the linguistic resources. To address this problem, in this study, we present a corpus with the information based on systemic functional linguistics.

SFL	Meaning	Annotation labels in KeiCO corpus
Field	What we want to talk about.	Field
Tenor	The social relationship between the participants in a conversation.	Honorific level Respectful 尊敬語 Humble 謙譲語 Polite 丁寧語

Table 1: Field-Tenor-Mode Framework in SFL

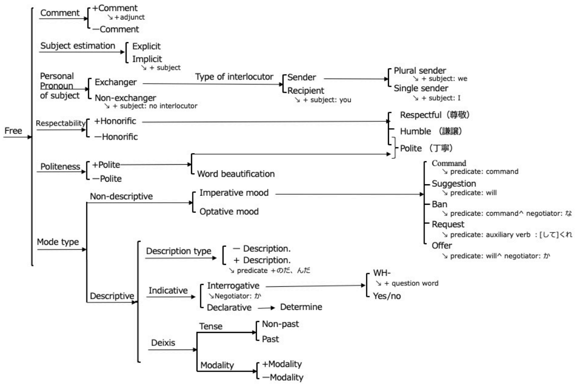

Figure 1: System Network of Mode system

3. Systemic Functional Linguistics

In Systemic Functional Linguistics (SFL, refer to appendix Appendix A for details), a language system is a concentric hierarchy of different types of symbolic systems - semantics, lexico-grammar, and phonology - surrounded by a context. It is a comprehensive model for the representation of language use in situations based on the values and common senses of a social group (Refer to Appendix: Figure 2). The context layer defines situations under three characteristics: the field, which describes the area of language use; the tenor, which describes the social relations between speakers; and the mode, which describes the medium used.

The characteristics are shown in Table 1. There are three meta-functions in the language system corresponding to each of the three properties of the context: ideational, interpersonal, and textual meanings, which constrain the selection of linguistic resources from the selection system network, called "system network", to form utterances appropriate to the situation.

In this study, we annotate each sentence with those mentioned above contextual elements of the three meta-functions to obtain the latent information necessary for language generation.

In particular, the annotation tags of the KeiCO corpus are defined based on the features of the system network of mode system (Refer to Figure 1) in the lexicogrammatical layer reflecting interpersonal meanings.

4. The KeiCO Corpus

In the construction of the KeiCO corpus, we collected the original texts containing honorific expressions from the dictionary (Sakamoto and Nishikata, 2009), the articles on the internet, and crowdsourcing services. Furthermore, we crowdsourced about 40 native Japanese annotators to annotate the level of honorific expressions and other SFL features. Each annotator was assigned about 75 source texts and asked to rewrite them into other honorific levels as much as possible while maintaining the meaning of the source texts. We have allowed annotators to do nothing if they have difficulty in rewriting. After completing the annotation part, we asked another 20 native Japanese speakers to check the annotations and manually correct any errors in the corpus. After all annotating and checking, we got the result: 10,007 sentences in total and 5 annotations per sentence.

More details on annotation are shown in the following Section 4.1, and detailed corpus analysis is in Section 5.

4.1. Structure and Annotation

In the KeiCO corpus, each sentence is annotated with seven annotations: honorific level, respectful (尊敬語), humble (謙譲語), polite (丁寧語) and field. Detailed definitions are given below.

The Table 2 gives an overview of the KeiCO corpus. The first row shows the annotations for the features of the system network in the mode system.

For each corpus sentence in the first column, apart from the honorific level, each annotation is assigned to a value of 0 or 1, where 1 corresponds to the target attribute and 0 indicates the opposite. A detailed definition of each annotation is given in the 4.1 chapter.

4.1.1. Honorific Level

The choice of honorifics primarily reflects role relationships (tenor). Tenor includes social, interpersonal relationships such as hierarchical relationships based on social status (e.g., boss-subordinate, teacher-student, etc.) and relationships (e.g., friend, acquaintance, etc.). In the KeiCO corpus, we set up four honorific levels reflecting the tenor. Each level is defined as follows.

Level 1: The Highest Honorific Level Level 1 is the level of respect most commonly used in the news, very formal speeches, and formal business emails. In sentences at the highest level of respect, it is common for verbs to be transformed into respectful or humble forms according to Japanese grammatical rules and for words that are originally respectful. It could also be a form of honorific linking, combining respectful (尊敬語) and humble (謙譲語) forms.

Level 2: Secondary Honorific Level Level 2 is widely used in business letters, general academic and business speeches, and the service industry. According to grammatical rules, the verb is transformed into respectful or humble, but few honorific linking forms are used.

Level 3: Third Honorific Level Complicated verb inflections are not used, and at most times, only polite (丁寧語) or word beautification (美化語) are used.

Level 4: No Honorifics Used No honorifics are used at all. Level 4 is more informal than Level 3 and may include polite expressions, abbreviations, and internet terminology.

4.1.2. Respectful, Humble and Polite

Based on the features of the system network in the mode system, we use three kinds of honorific expressions as annotations: respectful, humble, and polite. Respectful expressions express the speaker's respect for the subject of the conversation and are used for actions, objects, and names of the respected person. Modest speech indicates the speaker's intention to show respect to the listener by lowering his or her words and actions. Polite speech is mainly used to encourage helping verb endings such as "desu（です）" and "masu（ます）" to beautify the topic and show respect for the language.

4.1.3. Field

The field of activity indicates the area of use of the language, which includes the situation of having a conversation or the topic of the conversation. The use of honorifics is influenced by the specific activity field, such as business documents or lectures. To take this into account, In the KeiCO corpus, annotations are given to indicate specific activity areas. Currently, the KeiCO corpus has 122 different fields as options. (Refer to Appendix: Table 6)

5. KeiCO Corpus Analysis

5.1. Statistics of KeiCO

To analyze the use of vocabulary in KeiCO, we counted the number of sentences, the average sentence length, the average Kanji used in each sentence, the number of word tokens, word types and Yule's characteristic K. The characteristic statistics of KeiCO are shown in Table 3.

The K-characteristic was proposed by Yule (Yule, 1944). The smaller the value, the more diversity the vocabulary has. The K characteristic value assumes that the occurrence of words follows a Poisson distribution. Here, N means the number of word tokens and $V(m,N)$ means the number of word types that occur m times in a dataset. The K characteristic value is defined by the following equation 1.

$$K = 10^4 \times \frac{\sum_{all\,m}\left[m^2 V(m,N)\right] - N}{N^2} \qquad (1)$$

In Table 3, the K characteristic value decreased with the increase of the honorifics level, except for level 2. In fact, this exception is due to the fact that the number of sentences in honorifics level 2 is smaller than in other

Table 2: Overview of the KeiCO corpus

Sentences from KeiCO corpus	Honorific level	Respectful 尊敬語	Humble 謙譲語	Polite 丁寧語	Field
本日は、かねてより相談したいことがあり、参上しました．(I have come here today to discuss something that I have been wanting to discuss for some time.)	1	0	1	0	相談 consult
今日は、折り入ってご相談したいことがあって伺ったのですが．(I came here today because I wanted to ask you about something.)	2	0	1	0	相談 consult
今日は相談したいことがあったため、来ました．(I came here today because I had something I wanted to discuss.)	3	0	0	1	相談 consult
今日はずっと相談したいことがあって来た．(I came here today to consult with you about something.)	4	0	0	0	相談 consult
⋮	⋮	⋮	⋮	⋮	⋮

Table 3: Statistical results of KeiCO

Honorific level	Sentences	Average sentence length	Average Kanji in one sentence	Word tokens	Word types	Yule's characteristic K
Level 1	2584	18.2	2.6	47111	4744	135.70
Level 2	2046	16.4	2.1	33476	3897	136.23
Level 3	2694	15.2	1.8	40980	4448	130.28
Level 4	2683	13.5	1.6	36233	4315	129.80
Total	10007	15.8	2.0	157806	6465	125.54

levels. Regarding vocabulary, we confirmed that the use of Kanji increased with the honorific level.

5.2. KeiCO-based Classification

Since the advent of BERT (Devlin et al., 2019), BERT has achieved excellent performance in many tasks, making it one of the researchers' most commonly used models. In this section, we use BERT to perform a classification task, one of the most common and fundamental tasks, on our corpus KeiCO to examine how our corpus can improve performance in the NLP tasks.

To create a classification model, we use the KeiCO corpus to fine-tune the pre-trained BERT$_{BASE-Japanese}$[1], which was developed by Tohoku University. We divided the KeiCO corpus data into training, validation and evaluation in the ratio $6:2:2$, respectively. The number of epochs was set to 30.

We randomly select sentences from the corpus in the ratio 1%, 10%, 100% (100, 1000, and 10007 sentences) to check the effect of the data quantity on the classification accuracy. Table 4 shows the average of the classification accuracy (10 times) for each extracting ratio and each annotation of the KeoCO corpus.

As a result, respectful (尊敬語), humble (謙譲語), and polite (丁寧語) yielded high classification accuracy, while honorific level yielded relatively low accuracy.

On the other hand, looking at the average increase rate with a tenfold increase in the data, we can find honorific level, respectful (尊敬語) and polite (丁寧語) yielded high, while humble (謙譲語) yielded relatively low on the increase rate.

Respectful (尊敬語), Polite (丁寧語) Grammatical features of respectful (尊敬語), polite (丁寧語) are expressed obviously in sentences, and models can quickly identify those features. Therefore, it is easy to get high classification accuracy and average increase rate.

Humble (謙譲語) We consider the classification model, which caused the low average accuracy increase rate, could not be trained well, because humble (謙譲語) is biased toward one label in the corpus (Refer to Table 5).

Honorific Level As mentioned in Section 4.1.1, the honorific levels are categorised into four levels; therefore, it is natural to assume that the accuracy of the task is lower than other binary classification tasks. The high accuracy increase rate is also due to the balanced number of levels in the corpus, which contributes to the improvement of the accuracy increase rate of the task.

[1] https://huggingface.co/cl-tohoku/bert-base-japanese

Table 4: Classification accuracy of each feature in the KeiCO corpus (10 times average)

Classification accuracy	Honorific level	Respectful 尊敬語	Humble 謙譲語	Polite 丁寧語
data using 1%	0.482	0.646	0.894	0.706
data using 10%	0.653	0.686	0.887	0.810
data using 100%	0.727	0.698	0.906	0.842
Average accuracy increase rate	23.4%	17.3%	0.7%	9.4%

Table 5: Percentage of each annotation in the KeiCO corpus

Honorific level1	Honorific level2	Honorific level3	Honorific level4	Respectful 尊敬語	Humble 謙譲語	Polite 丁寧語
26%	20%	27%	27%	39%	9%	24%

6. Conclusion

Based on the language use taking social roles into account presented in systemic functional linguistics, we have created the KeiCO corpus, a corpus of Japanese honorifics that reflects the social status of speakers and listeners. The KeiCO corpus is annotated to take into account the social roles of dialogue participants in different domains of activity, as well as their modes of communication. As a general-purpose language resource, the corpus is expected to be useful for various tasks, such as improving the accuracy of machine translation, automatic evaluation, correction of Japanese composition and style transformation. We have not yet addressed the following issues: (1) The number of short sentences in each label is not balanced, (2) We need to review how copious the vocabularies are. Because some complex nouns are left intact in the rewritten sentences, which does not reflect the diversity of the vocabulary. In the future, we will increase the number of short sentences in the KeiCO corpus and put our main focus on the rewriting of nouns.

7. Bibliographical References

Aapakallio, N. (2021). Understanding through politeness–translations of japanese honorific speech to finnish and english. Master's thesis, Itä-Suomen yliopisto.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., and Potts, C. (2013). A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria, August. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Feely, W., Hasler, E., and de Gispert, A. (2019). Controlling Japanese honorifics in English-to-Japanese neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 45–53, Hong Kong, China, November. Association for Computational Linguistics.

Jin, D., Jin, Z., Hu, Z., Vechtomova, O., and Mihalcea, R. (2020). Deep learning for text style transfer: A survey. *CoRR*, abs/2011.00416.

Liu, G., Yang, Z., Tao, T., Liang, X., Li, Z., Zhou, B., Cui, S., and Hu, Z. (2022). Don't take it literally: An edit-invariant sequence loss for text generation.

Madaan, A., Setlur, A., Parekh, T., Poczos, B., Neubig, G., Yang, Y., Salakhutdinov, R., Black, A. W., and Prabhumoye, S. (2020). Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online, July. Association for Computational Linguistics.

Niu, T. and Bansal, M. (2018a). Polite dialogue generation without parallel data. *CoRR*, abs/1805.03162.

Niu, T. and Bansal, M. (2018b). Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Rao, S. and Tetreault, J. (2018). Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana, June. Association for Computational Linguistics.

Resmi, P. and Naseer, C. (2019). A deep learning approach for polite dialogue response generation. In *proceedings of the International Conference on Systems, Energy Environment (ICSEE) 2019*.

Sakamoto, T. and Nishikata, K. (2009). *A dictionary for honorific expressions ("Keigo no Ojiten" in Japanese)*. Sanseido.

Shirado, T., Marumoto, S., Murata, M., and Isahara, H. (2011). System for flexibly judging the misuse of honorifics in japanese. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 503–510.

Tsai, A., Oraby, S., Perera, V., Kao, J.-Y., Du, Y., Narayan-Chen, A., Chung, T., and Hakkani-Tur, D. (2021). Style control for schema-guided natural language generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 228–242, Online, November. Association for Computational Linguistics.

Xu, W. (2017). From shakespeare to twitter: What are language styles all about? In *Proceedings of the Workshop on Stylistic Variation*, pages 1–9.

Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge: At the University Press.

8. Language Resource References

Klimt, B. and Yang, Y. (2004). The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pages 217–226. Springer.

Appendix A. Systemic Functional Linguistics

Systemic Functional Linguistics (SFL) is a linguistic theory established by M.A.K. Halliday, who was influenced by the ideas of Malinowski, a cultural anthropologist, and Firth of the London School of Linguistics, who studied under Firth.

The major difference between SFL and other linguistics is that SFL introduces context, including the cultural background of a social group, into its theory and examines the language system from the viewpoint of its function in society, while most linguistics avoid dealing with various meanings comprehensively, limit the treatment of language meanings, and focus on the aspect of grammar. In contrast, SFL introduces a context that includes the cultural background of a social group into its theory, and examines the language system from the perspective of its function in society. The language system represented by SFL is shown in Figure 2.

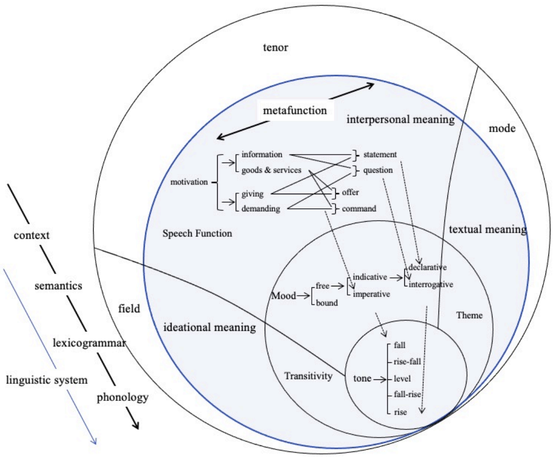

Figure 2: Language systems by systemic functional linguistics

Each layer of the language system expresses constraints on the choice of language resources through a network of choices called a choice network. The layers are organically connected by constraints called "realization statements". The systematization of linguistic resources using SFL and the procedures for their selection were considered to be directly applicable as algorithms for sentence generation, and in the 1980s they were called "systemic grammars" and used as the main linguistic theory for natural language sentence generation.

Table 6: List of fields in KeiCO corpus

Rank	Field	Num.	Rank	Field	Num.	Rank	Field	Num.	Rank	Field	Num.	Rank	Field	Num.
1	email	527	26	enjoy	112	51	calculation	60	76	go back	56	101	self	49
2	food	329	27	control	105	52	seasons	60	77	seek	56	102	creation	49
3	money	326	28	like	103	53	advice	60	78	shop	56	103	aspiration	49
4	guest	234	29	write	101	54	application	60	79	recommend	56	104	ruling	49
5	buy	229	30	work	100	55	ask	59	80	recognize	56	105	beliefs	49
6	attitude	227	31	celebrate	80	56	gather	59	81	ask	56	106	chastisement	48
7	apologize	217	32	anger	74	57	see	59	82	exist	56	107	clothes	47
8	contact	173	33	letter	66	58	ideas	59	83	visit	55	108	review	46
9	gift	162	34	say	64	59	consider	59	84	announcement	55	109	praise	46
10	greeting	159	35	baby	62	60	physique	59	85	body	55	110	appease	46
11	questions	158	36	play	62	61	research	59	86	farewell	55	111	appear	45
12	political speech	158	37	invitation	61	62	sports	58	87	rejection	55	112	walk	43
13	words	156	38	life	60	63	acquire	58	88	experience	55	113	go out	43
14	home	136	39	surprisingly	60	64	hate	58	89	free	54	114	congratulate	40
15	notice	134	40	win	60	65	refute	58	90	help	54	115	confirm	40
16	reception	120	41	plan	60	66	escape	58	91	thanks	54	116	encourage	40
17	relations	120	42	refrain	60	67	Manage	58	92	heart	53	117	anxious	37
18	work	120	43	send	60	68	do	58	93	preparation	53	118	embarrassed	37
19	public	119	44	contract	60	69	socialize	58	94	return	53	119	disagree	36
20	concern	117	45	wear	60	70	wait	57	95	report	52	120	talk	36
21	secret	116	46	physical condition	60	71	take in	57	96	medical condition	52	121	change	31
22	seat	116	47	choose	60	72	entrust	57	97	anxiety	51	122	introduce	24
23	phone	116	48	teaching	60	73	Get in trouble	57	98	humble	51			
24	school	116	49	flattery	60	74	end	57	99	know	50			
25	death	114	50	consultation	60	75	disappointed	57	100	visit	50			

Appendix B. Ethical Considerations and Broader Impact

The corpus is collected through the crowdsourcing platform Lancers`https://www.lancers.jp` and has been stripped of any information in the text that might be specific to the individual, such as gender, sexual orientation, health status, etc. All private information such as the name and address of the person appearing in the text has been anonymised.

Due to the presence of offensive content in the least honorific sentences, we removed uncomfortable content such as sexual topics, excessive swearing, and allegedly discriminatory statements by manually checking the samples. Although differences in language use due to gender and age were not taken into account in the design of this corpus for the time being, we tried to have multiple (three or more) native Japanese speakers annotate the same sentence during the data collection phase, and later calculated the average of each annotation as the final determined value. This was done in order to reproduce, as far as possible, the most common and accepted language expressions in everyday life.

Appendix C. Data Statement

We record information about our dataset following the data statement format proposed by Bender and Friedman (2018).

Data set name: KeiCO Corpus

Data set developer: Muxuan Liu

Dataset license: Creative Commons Attribution- NonCommercial-ShareAlike 4.0 International (CC BY- NC-SA 4.0)

Link to dataset: `https://github.com/Liumx2020/KeiCO-corpus`

Appendix C.1. Curation Rattionale

We provide a corpus of Japanese honorifics with information on social stance and attempt to reflect specific Japanese honorific usage and levels of respect through label. The corpus consists of sentences from the dictionary (Sakamoto and Nishikata, 2009), the articles on the internet, and crowdsourcing services. As the generalizability of the dataset has not been tested for the time being, we didn't actively split the corpus into a training, development and test set, but rather encouraged the data to be split randomly by a certain percentage when performing machine learning tasks.

Appendix C.2. Language Variety

N/A

Appendix C.3. Speaker Demographic

No detailed information was collected regarding the demographics of the authors of the collected sentences. However, we only collected the text or the speech from Japanese native speaker.

Appendix C.4. Annotator Demographic

The annotators are all Japanese native speaker but anonymous from internet, and no restrictions on age, gender or job.

Appendix C.5. Speech Situation

See table 6

Appendix C.6. Text Characteristics

The sentences in this dataset come from the dictionary (Sakamoto and Nishikata, 2009), the articles on the internet, and crowdsourcing services.

Appendix C.7. Recording Quality

N/A

Appendix C.8. Other

N/A

Appendix C.9. Provenance Appendix

The dictionary (Sakamoto and Nishikata, 2009), the articles on the internet, and crowdsourcing services.