Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

Who Is Missing? Characterizing the Participation of Different Demographic Groups in a Korean Nationwide Daily Conversation Corpus

Haewoon Kwak[1], **Jisun An**[1], **Kunwoo Park**[2]
[1]Singapore Management University, Singapore
[2]Soongsil University, Seoul, Korea
haewoon@acm.org, jisun.an@acm.org, kunwoo.park@ssu.ac.kr

Abstract

A conversation corpus is essential to build interactive AI applications. However, the demographic information of the participants in such corpora is largely underexplored mainly due to the lack of individual data in many corpora. In this work, we analyze a Korean nationwide daily conversation corpus constructed by the National Institute of Korean Language (NIKL) to characterize the participation of different demographic (age and sex) groups in the corpus.

1 Introduction

A conversation corpus is an essential resource for various dialogue systems, such as chatbots or virtual agents, to model real-world conversations accurately. There have been many attempts to collect English conversation data from various sources, such as telephone speech (Williams et al. 2013), video recording (Koiso et al. 2018), social media (Shang, Lu, and Li 2015), and chat logs (Lowe et al. 2015). Non-English speaking countries, such as Japan (Koiso et al. 2018) and China (Shang, Lu, and Li 2015), have also put effort in building conversational corpora of their own languages.

In such conversation corpora, the participants' demographic information plays a crucial role when modeling the conversations from two perspectives: individual-level speech modeling and pairwise dialogue modeling. First, the speakers' background information can facilitate generating consistent dialogue by modeling the personas of the speakers (Li et al. 2016). More importantly, an imbalance of demographic groups in a dataset leads to disadvantages toward under- or overrepresented groups in algorithmic systems (Koenecke et al. 2020; Larrazabal et al. 2020). Second, there exist demographic differences in conversations. For example, the way a teenage girl talks to her same-age friend would likely be different from how she talks to her grandmother. The demographic differences in pairwise conversation become more apparent in languages with honorific systems, as often used in non-English speaking countries. An apparent example is the Korean language, which has a rich and complex honorific system, as does the Japanese language. Korean honorifics appear as "subject honorification, object exaltation, and the six various speech levels" (Wikipedia 2021). For example, the short question, "Have you had lunch?", could be strikingly different in character and word levels from "밥 먹었어?" (plain speech) to "식사 하셨습니까?" (honorific speech) or "진지 드셨습니까?" (more honorific speech). Thus, the participants' demographic information in a public dataset should be carefully examined if possible. Unfortunately, many existing datasets fail to provide such individual data, and participants' demographic information in conversation corpora is underexplored.

In this work, we examine a Korean nationwide daily conversation corpus, "Daily Conversation Corpus 2020 (DCC2020 v1.0, 일상 대화 말뭉치 2020 v1.0," constructed by the National Institute of Korean Language (NIKL) (2021). The participants' demographic information is recorded with their explicit consent based on their Korean social security number, which encodes their sex and birth year. Thus, in contrast to other studies based on self-reporting or inference, we ensure a perfect accuracy of the demographic information in our study. We aim to understand the participation of different demographic groups and their impact on the conversations in this corpus.

In Korea, the use of honorific language largely depends on age (Yoon 2004). Thus, we focus on (1) how balanced (or unbalanced) conversations can be collected from pairs made up of different demographic groups, and (2) how other characteristics besides honorifics (e.g., topic choice, participation time, vocabulary choice) can be affected by the inclusion of participants of different ages and sexes.

For this, we pose four research questions that cover different stages of data collection. The first two research questions are about the pre-collection stage, and the remaining two research questions are about the characteristics of the collected conversations. These questions guide us to investigate DCC2020 from multiple perspectives, including participant recruitment and the characteristics of the conversations:

- RQ1: How do demographic features influence the choice of conversation partners? (§3.1)
- RQ2: How are demographic features associated with the participants' topic choices? (§3.2)
- RQ3: How do demographic features affect the participation in conversations? (§3.3)
- RQ4: How do demographic features affect the characteristics of conversations? (§3.4)

2 Daily Conversation Corpus 2020

DCC2020 is a 500-hour-long transcribed corpus constructed by NIKL (2021), a governmental organization aimed at "collecting national language resources and reinforcing the integrated information service." It contains 2,232 conversations (1,079 with strangers and 1,153 with non-strangers) and 2,739 persons (675 males and 2,064 females). Each conversation has metadata, such as age group and sex of the participants, their relationship, and utterances. Personally identifiable information of participants is not included.

All the conversations were recorded. Before the main recordings, the participants had one or two short prerecording sessions of 3 to 5 minutes each to check the recording conditions and to help build rapport between the stranger participants. The recorded conversations are later transcribed with masking all the sensitive information. Each transcriber processed at most four 15-minute recordings per day to ensure quality control.

The participants were required to select the topic of the conversation when applying for the program among 15 generic topics, including sports and family, or among 13 news articles across the business, world, local, and culture sections. In total, 1,818 conversations about generic topics and 414 conversations about news articles were collected.

Compared to conversations extracted from social media, which have a relatively low number of turns (Ritter, Cherry, and Dolan 2010) or that need to be split for topic consistency (Shang, Lu, and Li 2015), DCC2020 has the advantage that a well-trained moderator was always available at the site. Although the moderators usually just monitored the conversations without doing anything, they could ask the participants to return to the topic if the conversation started going off topic.

2.1 Recruiting Strategies

To get a nationwide sample of the entire Korean population, NIKL did proportional sampling across the regions and uniform sampling from (age $\times$ sex) groups. The age groups are the 10s, 20s, 30s, 40s, 50s, and 60s, and sex groups are females and males. Half of the original quota was considered the minimum to be recruited because some groups are challenging to recruit.

The recruiting was done through online channels. According to NIKL, their preference was for two persons, family or friends, to apply together (called non-strangers). When one person applied alone, NIKL matched that person with another applicant who is in a similar age group and who had selected the same conversation topic. Each participant receive 20,000 KRW (around $17 USD) for taking part in the conversation and recording.

3 Result

3.1 A1: Overrepresented Conversation Pairs in the Same Demographic Group

NIKL preferred two persons to apply together for participation, as explained in §2.1. This strategy is widely acceptable for conversation data collection as two *non-stranger* participants can typically have a smoother conversation. Thus, it is crucial to understand potential biases in non-stranger pairs to obtain representative conversations "in the wild."

We use the random rewiring of conversation partners to measure potential biases, which is commonly used in network science (Milo et al. 2002). It randomly picks two pairs (A-B and C-D) and swaps their partners (A-D and C-B). After a substantial number of swaps, we count the number of conversations for each demographic pair. These numbers are the result of one null model. We build 10,000 null models and compute the Z-score of each demographic pair p_i as follows:

$$Z_{p_i} = \frac{N_{p_i}^{actual} - avg(N_{p_i}^{null})}{std(N_{p_i}^{null})} \quad (1)$$

where N_{p_i} is the number of conversations of p_i. A high (low) Z_{p_i} indicates that the pair p_i is over-represented (under-represented) in the actual conversations compared to a "by chance" occurrence.

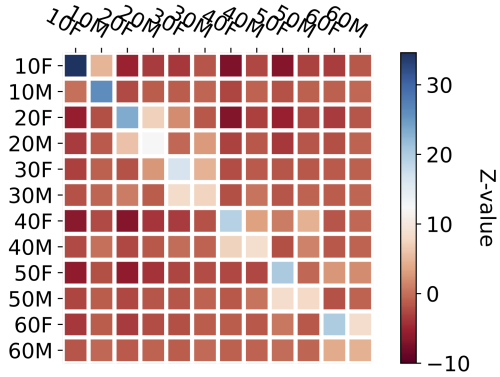

Figure 1: Z-score of non-stranger pairs (label: age+gender).

Figure 1 shows the Z-scores of each demographic pair in the conversations between non-strangers. The larger Z-scores are located along the diagonal, showing the overrepresented conversations between non-strangers of the same demographic group. We note that stranger pairs are not considered in this experiment due to the artificial effect whereby NIKL matched strangers with similar ages (§2.1). Besides the same demographic group, we identify a strong preference toward the same age group with the opposite sex, but weaker than that toward the same demographic group.

In summary, when two persons applied together to participate in dialogue collection (non-strangers), strong preferences exist toward the same demographic group followed by the same age with the opposite sex group. This tendency should be carefully considered to obtain a balanced sample of conversations across different demographic groups. While NIKL matched similar age groups for the stranger pairs, someone might think that considering stranger pairs will be a solution to address the lack of conversations between certain demographic groups. We will investigate the potential of this approach by comparing the characteristics of the conversations between strangers and non-strangers in §3.3 and §3.4.

3.2 A2: Different Topic Preferences of the Varied Demographic Groups

As we explained in §2, participants chose their conversation topic in advance. So how does that decision affect the topic diversity in DCC2020?

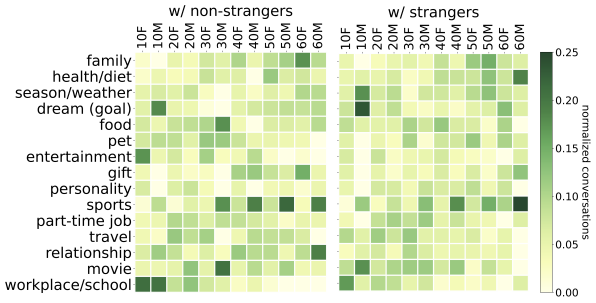

Figure 2: Preferences toward generic topics

Figure 2 shows the normalized topic preferences for each demographic group for when people talk to non-strangers (left) or strangers (right). In other words, the sum of the normalized preferences for each group across all 15 topics is 1. We observe some noticeable trends. For example, males express a preference toward sports, as expected. Preferences toward family and health topics become strong as age increases. Teenage males have a strong preference toward their life dreams, but teenage females tend to prefer to talk about their school (workplace/school). However, the visible differences between the two heatmaps are not statistically significant for all the demographic groups except for the 40s females, according to the chi-square test (p=0.003 for 40s females, $p > 0.05$ for the other demographic groups).

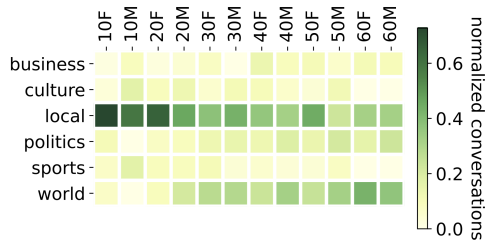

Figure 3: Preferences toward the news sections

Figure 3 shows the normalized preferences toward the news section. As the number of participants who selected a new article as a conversation topic is much lower than those who chose a generic topic (414 vs. 1,818 conversations), we do not divide the result by stranger vs. non-stranger settings to avoid an overemphasis of their differences.

By comparing with the news consumption patterns of Koreans reported in An and Kwak (2017), we can identify some commonalities and differences. In DCC2020, local news as a conversation topic is likely to be chosen by many different demographic groups, while world news is selected by relatively older generations, which are well aligned with the known news consumption patterns. However, a strong preference of 50s males toward politics that is observed in news consumption does not appear in the conversations in DCC2020. Also, females' strong preference towards cultural news in news consumption does not appear in the conversation corpus. Examining this result manually, we find that NIKL selected a news article entitled "immigration issues are not easy to resolve but can strengthen our culture," which might not be considered typical culture news, such as books, music, arts, or films (The Guardian 2021).

In summary, different demographic groups show different topic preferences for the conversations. Thus, in combination with the over-represented conversation pairs with the same demographic, this suggests the topic distribution in a conversation dataset can be biased. Participants' choice of conversation topics can further aggravate this bias.

3.3 A3: Decreasing Participation of the Varied Demographic Groups in Conversations with Strangers

In §3.1, we find that, in a non-stranger setting, a conversation pair is more likely to be formed from the same demographic group, which hinders making balanced conversational data. Then, can 'matching strangers of different demographic groups' be a solution to this issue?

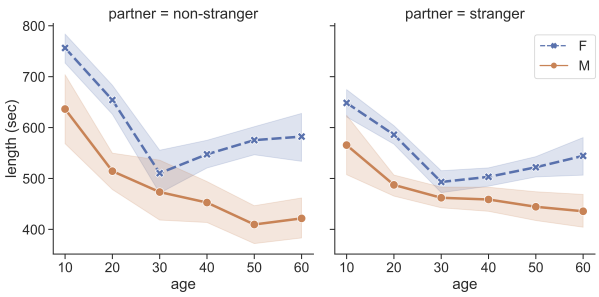

Figure 4: Speaking time of each demographic group

Figure 4 shows the speaking time of a participant of a specific demographic group when they talk to strangers and non-strangers. We compute the speaking time using the start and end timestamps of the utterances of each participant. We find a sex difference in adapting to conversations with strangers by one-sided t-test. Females of all age groups except for the 30s and 60s show significantly decreased speaking time when talking to strangers ($p < 0.005$). By contrast, changes in the males times are not significant ($p > 0.05$).

Another important factor that might influence the conversation, particularly in Korea, is age—"age is the first consideration when Koreans communicate with one another (Park and Kim 1992)". Figure 5 shows how the 30s group adapt to conversations with strangers according to the age of the strangers. We focus on the stranger pairs to understand the effect clearly, and the 30s were selected because they are in the middle of the age groups. Clearly, they demonstrate strong participation in a conversation when they talk to the *same* demographic group. This tendency is observed across all age and sex groups. Moreover, the 20s and 50s females

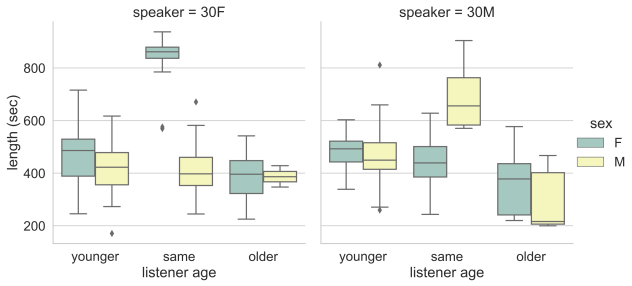

Figure 5: Speaking time of the 30s group according to the listeners (when the listeners are strangers)

and 30s males show statistically significantly longer speaking time when they talk to younger participants than older participants, as confirmed by one-sided t-test ($p < 0.01$), implying that the age of strangers differently affects speakers.

The interactivity, another dimension of participation, can be measured by the number of turns. We compare the number of turns in conversations between strangers and that between non-strangers. We find that conversations between non-strangers are more interactive than conversations between strangers (p=0.012 by t-test), and conversations between the same age group are more interactive than different age groups ($p<0.0001$ by t-test).

In summary, as conversations between non-strangers have a strong bias toward the same demographic or age group, conversations between strangers of different demographic groups are essential to get a balanced sample of conversations. However, strangers of different demographic groups might lead to less active, less interactive conversations.

3.4 A4: Increased Vocabulary When Talking to the Same Demographic Group

In §3.3, we find conversations between the same demographic group are more active than those between different demographic groups. But, how does this affect their vocabularies used in the conversations?

Figure 6 shows the result of the comparisons between the number of unique tokens when people in one demographic group (column) talk to the same demographic group versus when they talk to other demographic groups. We omit the demographic groups that do not have statistically significant results due to the lack of space. Similar to observations in §3.3, we find that each demographic group adapts to conversations with different demographic groups in a different way. Some groups (e.g., 10s males and females) actively adapt, but others (e.g., 30s) do not show statistically significant changes. Also, if some demographic group adapts to conversations with different demographic groups, they always use more unique tokens when they talk to the same demographic group. No opposite effect is observed.

In summary, some demographic groups use more unique tokens in conversations with the same demographic group than those with different demographic groups. In addition to §3.3, one should consider that strangers of different demographic groups might lead to less diverse conversations.

	10F	20F	20M	40F	40M	50F	60F
family		++				++	
health/diet	++	+++					++
season/weather	++	++	++				
dream (goal)	++	++					
food							
pet	+						
entertainment	++			++	+		
gift	+			+++		+++	
personality		+					
sports	+						
part-time job	+++	++					
travel				+			
relationship	++	++					
movie	++						
workplace/school	+++						

Figure 6: Differences in the number of unique tokens when one demographic group (column) talks to the same group (U_s) vs. different groups (U_d). +++: $p < 0.01$, ++: $p < 0.05$, +: $p < 0.1$ when $avg(U_s) > avg(U_d)$ by one-sided t-test. There are no cases of $avg(U_s) < avg(U_d)$.

4 Discussion and Conclusion

We examined a nationwide conversation corpus built by a governmental organization in Korea. We discovered participants' preferences toward specific demographic pairs and topics, and the impacts of these on the conversations.

Many languages use honorifics. Notably, Asian countries share the cultural context regarding social relationships influenced by age. Considering that some of those countries currently lack a large-scale corpus of their own language, there is a potential that government-related agencies may take the lead to build a nationwide corpus (due to their relatively small market). We believe that our work will be a helpful resource to guide such agencies and countries in building a balanced nationwide corpus.

There are some limitations in this work. First, our findings should be carefully interpreted and applied to other countries. Even though we expect that other East Asian countries might show similar trends, such as a strong preference toward the same demographic group or a longer speaking time when speaking to a younger generation, more studies are required to validate the generalizability of our findings beyond South Korea. Second, this work is an observational study. Some interventions could be found as effective that would have elicited more active, lively, and rich conversation even from stranger pairs who did not actively participate in DCC2020. We expect this study will become a stepping stone for future efforts toward understanding biases in conversational corpora and building fair dialogue systems.

Acknowledgments

This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant. K Park was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (No. NRF-2021R1F1A1062691).

References

An, J.; and Kwak, H. 2017. Multidimensional Analysis of the News Consumption of Different Demographic Groups on a Nationwide Scale. In *Proceedings of the Social Informatics*, 124–142.

Koenecke, A.; Nam, A.; Lake, E.; Nudell, J.; Quartey, M.; Mengesha, Z.; Toups, C.; Rickford, J. R.; Jurafsky, D.; and Goel, S. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14): 7684–7689.

Koiso, H.; Den, Y.; Iseki, Y.; Kashino, W.; Kawabata, Y.; Nishikawa, K.; Tanaka, Y.; and Usuda, Y. 2018. Construction of the Corpus of Everyday Japanese Conversation: An Interim Report. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Larrazabal, A. J.; Nieto, N.; Peterson, V.; Milone, D. H.; and Ferrante, E. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23): 12592–12594.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G.; Gao, J.; and Dolan, B. 2016. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 994–1003. Association for Computational Linguistics.

Lowe, R.; Pow, N.; Serban, I.; and Pineau, J. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 285–294. Association for Computational Linguistics.

Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D.; and Alon, U. 2002. Network motifs: simple building blocks of complex networks. *Science*, 298(5594): 824–827.

National Institute of Korean Language. 2021. Daily Conversation Corpus 2020. https://corpus.korean.go.kr/. Accessed: 2021-05-09.

National Institute of Korean Language. 2021. National Institute of Korean Language. https://korean.go.kr/. Accessed: 2021-05-09.

Park, M.; and Kim, M. 1992. Communication practices in Korea. *Communication Quarterly*, 40(4): 398–404.

Ritter, A.; Cherry, C.; and Dolan, B. 2010. Unsupervised Modeling of Twitter Conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 172–180. Association for Computational Linguistics.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural Responding Machine for Short-Text Conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1577–1586. Association for Computational Linguistics.

The Guardian. 2021. The Guardian — Culture. https://www.theguardian.com/uk/culture. Accessed: 2021-05-09.

Wikipedia. 2021. Honorifics. https://en.wikipedia.org/wiki/Honorifics_(linguistics). Accessed: 2021-05-09.

Williams, J.; Raux, A.; Ramachandran, D.; and Black, A. 2013. The Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2013 Conference*, 404–413. Association for Computational Linguistics.

Yoon, K.-J. 2004. Not just words: Korean social models and the use of honorifics. *Intercultural Pragmatics*, 1(2): 189–210.