Investigating the Automatic Translation of Korean Honorifics

Luis Cihlar[▽*] **Minh Duc Bui**[▽*] **Kyung eun Park**[♠]
Manuel Mager[▽] **Walter Bisang**[▽] **Katharina von der Wense**[▽,♣]
[▽] Johannes Gutenberg University Mainz, Germany [♠] University of Mannheim, Germany
[♣] University of Colorado Boulder, USA
minhducbui@uni-mainz.de

Abstract

Honorifics encode social hierarchies and relational nuances, making their correct use a culturally sensitive and challenging aspect of translation. In doing so, they reflect and shape how individuals position themselves and others within a social world. In this work, we investigate how different models handle Korean honorific translation, both in *implicit* scenarios, where only the sentence is given, and *explicit* scenarios. Our findings are as follows: (i) large language models finetuned for translation (`MTLMs`) consistently prefer polite forms more than their instruction-tuned counterparts in both scenarios, (ii) sequence-to-sequence models produce less polite outputs in implicit contexts but shift toward more polite forms when the addressee is explicitly provided; and (iii) both types of LM-based models tend to become more casual when the addressee is known. When compared with human preferences, `MTLMs` diverge more strongly, exhibiting a systematic overuse of polite forms relative to human judgments.

Keywords: machine translation, honorifics, Korean

1. Introduction

Honorifics are a widespread linguistic feature across the world's languages (Helmbrecht, 2013), yet the social distinctions they convey differ greatly between languages and cultures (Shin, 2017). In languages such as Japanese, Hindi, Javanese and Korean, honorifics encode intricate social hierarchies in nearly every utterance, whereas in English they play only a minimal role (Hwang et al., 2021; Song, 2015). In these societies, the correct use of honorifics is taught from an early age and closely monitored by parents and teachers (Yoon, 2004). Misuse of honorifics can result in significant social, economic, or even familial repercussions (Soucova, 2005; Brown, 2010). In addition, grammatical marking of honorific information is obligatory on every verb of a finite clause (Sohn, 1994; Bisang, 2007). Therefore, translating from a language without honorific markers such as English into a language with obligatory expression of honorifics such as Korean requires the relevant information to be enriched in the target language (Feely et al., 2019). If no additional information is provided, a proper translation is exclusively dependent on implicit inference from cues provided in the text. In more explicit cases, translation may be enhanced by providing additional information, as social status or hierarchical relationships.

It is well established that existing models struggle to adequately capture such culturally embedded phenomena (Fernandes et al., 2023; Tenzer et al., 2025; Lee and Wang, 2023). However, it remains unclear whether different model types handle honorifics differently. To shed light on this question, we compare a diverse set of models – including LLMs that are predominantly finetuned on parallel sentences for translation (`MTLMs`) (Cui et al., 2025; Rei et al., 2025; Zheng et al., 2025), instruction-tuned LLMs (`ITLMs`) (Qwen et al., 2025; Gemma et al., 2024), and sequence-to-sequence models `seq2seq` trained on parallel data (Sutskever et al., 2014; Bahdanau et al., 2014) – with regards to their generation of Korean honorifics during English–Korean (`en–ko`) translation, see Figure 1.

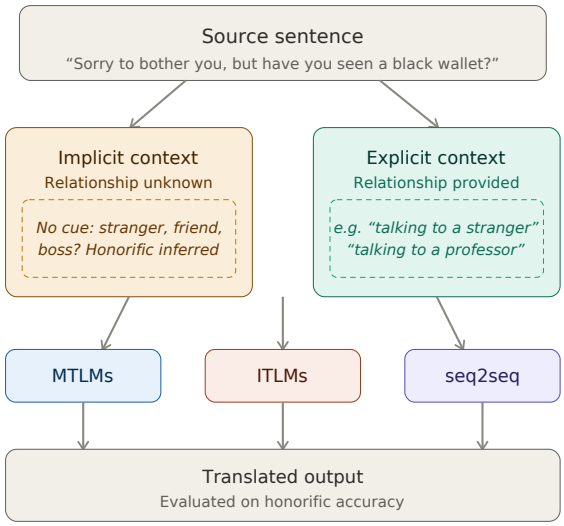


Figure 1: **Experimental Overview.** We assess honorific usage across different model families in translation.

Our research questions are the following:

* These authors contributed equally to this work.

Proceedings of Identity-Aware AI (IAAI) @ LREC 2026, pages 20–31
16 May 2026.

(RQ1) ***Which honorifics do different models generate during en–ko translation when the relationship between speaker and addressee has to be inferred?*** We examine translations without explicit contextual information—where the addressee must be inferred. We find that MTLMs employ polite forms more frequently than their instruction-tuned counterparts, indicating a general preference for higher levels of formality when the addressee is unknown. Interestingly, seq2seq models show a formality level nearly as low as ITLMs.

(RQ2) ***How does the behavior of different models change when the addressee is explicitly specified?*** When explicit addressee information is provided, both MTLMs and ITLMs shift slightly toward more casual forms, suggesting that they adapt their tone once contextual cues clarify the social relationship, while seq2seq become more polite.

(RQ3) ***How do models align with human preferences during en–ko translation?*** Our analysis shows that the consistently higher politeness levels of MTLMs diverge from human preferences, leading to lower accuracy in honorific usage. In contrast, ITLMs exhibit stronger alignment with human judgments, showing more contextually appropriate use of politeness. Seq2seq models align well with humans in the implicit scenario but fail to adapt when the addressee is explicit.

2. Korean Honorifics

Korean people must abide by strict cultural rules, governing every human interaction and every utterance beholding one. While the Korea of today is steadily becoming more egalitarian, the Korean honorific system still holds influence over people's lives (Brown, 2015).

Korean Honorific Typology Korean honorifics encode varying levels of politeness and formality through a rich morphological and lexical system. Two major types are realized as verbal morphemes: addressee honorifics (or speech styles), which signal the hierarchical relationship between speaker and hearer, and referent honorifics, which express respect toward the person being spoken about. In addition, Korean employs lexical substitutions to convey honorification (Brown, 2015).

Speech Styles Our dataset focuses exclusively on Korean addressee honorifics—an information that requires compulsory marking on every finite verb of a sentence. Out of the six speech styles commonly used in the Korean language (§App. A) we only assess the three most critical:[1] *Casual* (해), *Polite* (해요), and *Deferential* (합니다/하십시오).

Related Work: Honorifics in NLP Research on other honorific languages like Javanese (Farhansyah et al., 2025) or Japanese (Feely et al., 2019) sought to improve honorific comprehension and translation by creating datasets and training models on labeled data. Both finding imbalances in the distribution of honorific styles in machine translation, while showing that models trained on language-specific data generally outperform language-agnostic models.

For Korean, Hwang et al. (2021) constructed and annotated a parallel discourse-level corpus from English–Korean movie and TV subtitles, with the intention of providing data that better reflects how humans use honorifics. Other research uses context-aware prompting to add additional information (Lee et al., 2025), analyzing information from previous sentences (Hwang et al., 2021), training with formality classifiers (Kim et al., 2023), or simply letting users choose the intended speech style like the cloud-based Papago (Naver, 2017).

3. Experiments

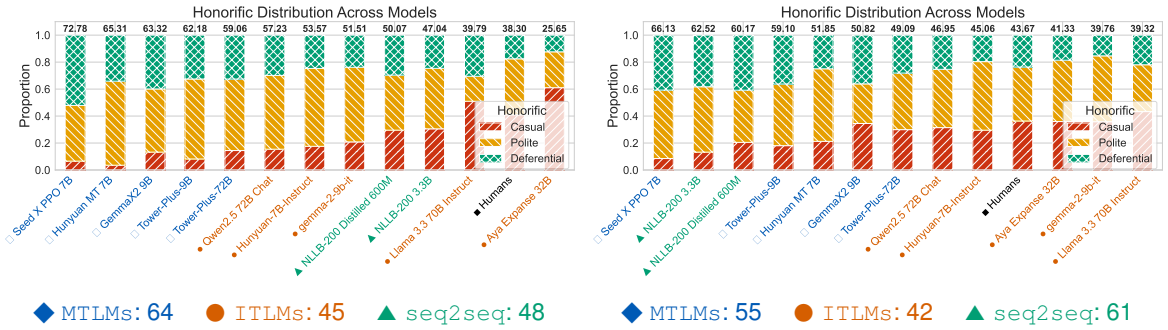
Data Generation We generate 750 source sentences used for translation with GPT-5. The detailed prompting procedure is described in Appendix B.1 and B.2.[2]

Experimental Setup In the implicit scenario, the model receives only the sentence to translate, without any contextual information. In the explicit scenario, we augment the translation input sentence with the following prefix: "I was talking to {addressee}, and I said: {sentence}" By explicitly specifying the addressee, the task provides clearer guidance to infer the honorific. As a result, ITLMs are expected to adapt more effectively than there counterparts.

Automatic Honorific Extraction To automatically identify honorific forms in Korean sentences, we employ GPT-OSS 120B (OpenAI et al., 2025) as an evaluator. As detailed in Appendix B.3, the model achieves 95% accuracy.

[1] The other speech styles are slowly falling out of use in spoken language (Kim, 2023) or are limited to very specific situations (Brown, 2015). Furthermore, these three speech styles are essential in situations that prescribe politeness, as they mark the relative social status in most human interactions (Yoon, 2004).

[2] We publish the code, the dataset, and human-annotated subset (see Section 4) at: https://github.com/MinhDucBui/KoreanHonorifics.



(a) **Implicit Scenario** (b) **Explicit Scenario**

Figure 2: **Normalized Distribution of Honorific Categories.** Blue (◆) denotes MTLMs, Orange (●) denotes ITLMs, and Green (▲) denotes sequence-to-sequence models. Politeness scores are shown above each bar, models are sorted by score, and the mean is reported below each graph.

ITLM	MTLM	Δ
Implicit Scenario		
Qwen2.5 72B Chat	Tower-Plus-72B	+1.8
gemma-2-9b-it	Tower-Plus-9B	+10.7
gemma-2-9b-it	GemmaX2 9B	+11.8
Hunyuan-7B-Instruct	Hunyuan MT 7B	+11.7

Table 1: **Pairwise Comparison of MTLMs and Their ITLM Counterparts in Politeness Scores in Implicit Scenario.** Δ = (MTLM - ITLM).

ITLM	MTLM	Δ
Explicit Scenario		
Qwen2.5 72B Chat	Tower-Plus-72B	+2.1
gemma-2-9b-it	Tower-Plus-9B	+19.3
gemma-2-9b-it	GemmaX2 9B	+11.1
Hunyuan-7B-Instruct	Hunyuan MT 7B	+6.8

Table 2: **Pairwise Comparison of MTLMs and Their ITLM Counterparts in Politeness Scores in the Explicit Scenario.** Δ = (MTLM - ITLM).

Politeness Score To quantify a model's tendency to produce more polite translations, we compute a politeness score using ordinal weights $[0, 0.5, 1]$ for *Casual*, *Polite*, and *Deferential*: The politeness score is computed as $(0.5\,p_{Pol} + p_{Def}) \times 100$, where p_{Pol} and p_{Def} denote the proportions of *Polite* and *Deferential* outputs.

Models We include four LLM-based MTLMs with their instruction-tuned counterparts: GemmaX2 9B (Cui et al., 2025), Tower+ 72B/9B (Rei et al., 2025), and Hunyuan 7B MT (Zheng et al., 2025), derived respectively from Gemma 2 9B (Gemma et al., 2024), Qwen 72B (Qwen et al., 2025)/Gemma 2 9B, and Hunyuan 7B. Each is compared to its instruction-tuned counterpart. We also include SeedX (Cheng et al., 2025), an MTLM without an instruction-tuned counterpart, and Llama-3.3 70B (Grattafiori et al., 2024) and Aya Expanse 32B (Dang et al., 2024) as ITLMs. For Seq2seq models, we evaluate NLLB-200 600M and 3.3B (NLLB et al., 2022). See prompts in Appendix B.5.

3.1. Results for RQ1

We first analyse model behavior in the implicit scenario, i.e., without explicit addressee information.

Overall Results We present the distribution across all models for the implicit scenario in Figure 2a. MTLMs consistently yield higher politeness scores than ITLMs and seq2seq models. On average, MTLMs achieve a politeness score of 64, compared to 45 for ITLMs, while the seq2seq models reach a comparable score of 48. In summary, **MTLMs systematically tend to generate more polite forms**.

Pairwise Comparisons Table 1 presents pairwise comparisons between each MTLM and its instruction-tuned counterpart. Across all pairs, the MTLMs exhibit higher politeness scores. This suggests that **MTLMs systematically produce more polite forms than their instruction-tuned counterpart**.

3.2. Results for RQ2

We now analyze the behavioral shift that occurs when the addressee is made explicit.

MTLMs and ITLMs Become More Casual When explicit addressee information is available, both MTLMs and ITLMs tend to produce more casual forms: their politeness scores *decrease* by 3.07 and 9.13, suggesting that they adjust their tone

Implicit Scenario		Explicit Scenario	
Model Family	**Acc. (%)**	**Model Family**	**Acc. (%)**
● ITLMs	71.78	● ITLMs	**81.11**
◆ MTLMs	60.67	◆ MTLMs	66.67
▲ Seq2Seq	**72.78**	▲ Seq2Seq	60.55

Table 3: **Model-Family Alignment with Human Judgments.** Detailed results for all models are in Figure 8 in the appendix.

once contextual cues clarify the social relationship. In contrast, the politeness score of `seq2seq` models *increases* by 12.79.

Overall, `MTLMs` Remain More Polite Figure 2b presents the overall politeness distribution, and Table 2 shows the pairwise comparison between each `MTLM` and its instruction-tuned counterpart. The pattern from the implicit scenario persists: on average, `MTLMs` continue to produce more polite outputs.

4. RQ3: Human Alignment Study

To answer RQ3, we collect human annotations for a subset of our dataset.

4.1. Annotation and Evaluation

We collect human judgments for 90 sentences, each presented in both its implicit and explicit form. The annotators are randomly assigned to annotate either the implicit or the explicit set, ensuring non-overlapping coverage. Each sentence is annotated in a multiple-choice format by five participants, with a total of 23 participants, with three options corresponding to the three focal speech styles: *Casual* (해), *Polite* (해요), and *Deferential* (합니다/하십시오). We report our survey, demographics and details in App. C.

A speech style is considered correct if selected by at least two participants. This threshold captures natural variation in Korean honorific use (Kwon and Sturt, 2024), while avoiding noisy labels.

4.2. Model Alignment with Humans

Implicit and Explicit Setting As shown in Table 3, `Seq2Seq` models show the highest alignment with human annotations in the implicit scenario, when the addressee is not explicitly stated. This suggests that `Seq2Seq` models may have learned effective structural patterns from their parallel training data and architecture design. `MTLMs` perform the worst in this scenario (61% acc.). When explicit addressee information is provided, the alignment of `Seq2Seq` models with human judgments drops sharply. This indicates that these models struggle to incorporate explicit contextual cues effectively. In contrast, `ITLMs` and `MTLMs` adapt better to the explicit condition, and `ITLMs` achieve the highest overall alignment with human annotations.

`ITLM`	`MTLM`	Δ
Implicit Scenario		
Qwen2.5 72B Chat	Tower-Plus-72B	+4.44
gemma-2-9b-it	Tower-Plus-9B	+7.77
gemma-2-9b-it	GemmaX2 9B	+22.22
Hunyuan-7B-Instruct	Hunyuan MT 7B	+10.00
Explicit Scenario		
Qwen2.5 72B Chat	Tower-Plus-72B	+6.66
gemma-2-9b-it	Tower-Plus-9B	+23.33
gemma-2-9b-it	GemmaX2 9B	+30.00
Hunyuan-7B-Instruct	Hunyuan MT 7B	-11.11

Table 4: **Pairwise Comparison of `MTLMs` and Their `ITLM` Counterparts with regards to Human Alignment**. Differences (Δ) are computed as `ITLM` - `MTLM`.

Pairwise Comparisons As shown in Table 4, across nearly all cases, instruction-tuned models outperform their translation-specialized versions. For instance, in the implicit scenario, `gemma-2-9b-it` achieves a $+22.2$ acc. improvement over `GemmaX2 9B`, and in the explicit scenario, the gap further widens to $+30.0$ points. This suggests that, while `MTLMs` produce more polite translations this does not result in higher alignment. Instead, excessive politeness reflects a **limited ability to adapt to social context**, leading to less appropriate handling of honorific expressions.

5. Conclusion

We investigate how different model families translate Korean honorifics with or without additional context. Our findings reveal the following differences: `MTLMs` tend to produce more polite forms in both implicit and explicit scenarios, their outputs align less closely with human judgments. In contrast, `ITLMs` exhibit greater sensitivity to contextual cues, indicating that instruction-tuning more effectively captures the social nuance necessary for appropriate honorific usage. `Seq2Seq` models are less polite and closer to human preferences in implicit settings but become overly polite when the addressee is given, diverging from human preferences.

6. Limitations

In this work, we focus exclusively on Korean, one of the most systematically developed honorific languages in the world (Sohn, 1994). While our findings offer indicative evidence of model deviations in handling honorific translation, they should not be generalized to all languages that employ honorific systems. Additionally, language is not static but evolves in response to sociocultural change, as well as divergences across speaker demographics (Blount and Sanches, 2014; Trudgill, 2000). Although our current study does not capture these dynamics, future research could investigate how model behavior shifts over time and to what extent such changes mirror real-world language use.

We use a politeness score to assess differences in honorific translation behavior. This metric is designed to capture and aggregate the behavioral distinction between deferential and polite forms for easier comparison. However, it is not an established measure and may therefore introduce potential ambiguity.

Furthermore, while we examine differences across models, we do not control for the training datasets used. Future work could adopt a more controlled experimental setup in which models are fine-tuned from the same base LLM and dataset using different objectives, allowing a more precise analysis of how training schemes influence honorific translation.

Future studies should include more expansive and sophisticated human surveys, since the participants in our survey represented a specific subset of the Korean population.

7. Ethics Statement

All participants voluntarily took part in our human survey. They were recruited through personal networks, including friends, family, and acquaintances, and did not receive any form of payment. The collected data contain no personally identifying information. All participants were informed about the purpose of the study prior to their participation.

We use AI assistants, specifically GPT-5, to help edit sentences in our paper writing.

8. Acknowledgment

This work was supported by the Carl Zeiss Foundation through the TOPML project, grant number P2021-02-014.

9. Bibliographical References

Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Walter Bisang. 2007. Categories that make finiteness: discreteness from a functional perspective and some of its repercussions. *Finiteness: Theoretical and empirical foundations*, pages 115–137.

Ben G Blount and Mary Sanches. 2014. *Sociocultural dimensions of language change*. Elsevier.

Lucien Brown. 2010. Politeness and second language learning: The case of korean speech styles. *Journal of Politeness Research-language Behaviour Culture - J POLITENESS RES-LANG BEH CUL*, 6:243–269.

Lucien Brown. 2015. *Honorifics and Politeness*, chapter 17. John Wiley & Sons, Ltd.

Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang et al. 2025. Seed-x: Building strong multilingual translation llm with 7b parameters.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier.

Mohammad Rifqi Farhansyah, Iwan Darmawan, Adryan Kusumawardhana, Genta Indra Winata, Alham Fikri Aji and Derry Tanti Wijaya. 2025. Do language models understand honorific systems in javanese?

Weston Feely, Eva Hasler and Adrià de Gispert. 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. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Team Gemma, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé et al. 2024. Gemma 2: Improving open language models at a practical size.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan et al. 2024. The llama 3 herd of models.

Johannes Helmbrecht. 2013. Politeness distinctions in pronouns. In: Dryer, Matthew S. & Haspelmath, Martin (eds.), *WALS Online* (v2020.4) [Data set]. Zenodo. Available online at `http://wals.info/chapter/45`, Accessed on 2025-10-06.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Yongkeun Hwang, Yanghoon Kim and Kyomin Jung. 2021. Context-aware neural machine translation for korean honorific expressions. *Electronics*, 10(13).

Dohee Kim, Yujin Baek, Soyoung Yang and Jaegul Choo. 2023. Towards formality-aware neural machine translation by leveraging context information. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7384–7392, Singapore. Association for Computational Linguistics.

Minju Kim. 2023. Between honorifics and non-honorifics: A study of the korean semi-honorific style and a comparison with japanese. *Discourse Studies*, 25(5):664–691.

Nayoung Kwon and Patrick Sturt. 2024. When social hierarchy matters grammatically: Investigation of the processing of honorifics in korean. *Cognition*, 251:105912.

Minjae Lee, Youngbin Noh and Seung Jin Lee. 2025. A testset for context-aware LLM translation in Korean-to-English discourse level translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1632–1646, Abu Dhabi, UAE. Association for Computational Linguistics.

Soo-Hwan Lee and Shaonan Wang. 2023. Do language models know how to be polite? In *Proceedings of the Society for Computation in Linguistics (SCiL)*, volume 6, pages 375–378. University of Massachusetts Amherst Libraries.

Corporation Naver. 2017. Naver papago.

Team NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht et al. 2022. No language left behind: Scaling human-centered machine translation.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai et al. 2025. gpt-oss-120b & gpt-oss-20b model card.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu et al. 2025. Qwen2.5 technical report.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms.

Moun Kyoung Shin. 2017. A comparative study of honorific systems in north and south korea: Shifts since 1950. Unpublished.

Ho-Min Sohn. 1994. Korean. *London & New York: Routledge*.

Sanghoun Song. 2015. Representing honorifics via individual constraints. In *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 Workshop*, pages 57–64, Beijing, China. Association for Computational Linguistics.

J. Soucova. 2005. The japanese honorific language: Its past, present and future. Conference paper.

Ilya Sutskever, Oriol Vinyals and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Helene Tenzer, Oumnia Abidi and Stefan Feuerriegel. 2025. Designing llms for cultural sensitivity: Evidence from english-japanese translation.

Peter Trudgill. 2000. *Sociolinguistics: An introduction to language and society*. Penguin UK.

Kyung-Joo Yoon. 2004. Not just words: Korean social models and the use of honorifics. *Intercultural Pragmatics - INTERCULT PRAGMAT*, 1:189–210.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun and Di Wang. 2025. Hunyuan-mt technical report.

A. Korean Honorific Details

A.1. Speech Style Details

Table 5 and Table 6 present a quick overview of all Korean speech styles.

A.2. Unexplored Honorifics

The Korean honorific system is more nuanced and far more complex than what our paper might suggest. Referent honorifics and lexical substitutions are almost as common and definitely as important as the aforementioned speech styles.

Speech styles, referent honorifics and lexical substitutions are also not exclusively shaped by social dynamics. As many spoken interactions feature a mixture of formal and informal styles (Brown, 2015).

B. Methodology Details

B.1. Data Generation Details

We define ten addressee scenarios for each of the three Korean honorific levels (§Sec. 2); the complete list of addressees per level is provided in Table 7 in the appendix. For each scenario, we use GPT-5 (gpt-5-2025-08-07) to generate 25 "typical" utterances that one might naturally say to the respective addressee. In total, this results in 750 base sentences.

B.2. Sentence Generation Prompt

To generate base sentences for each addressee. The prompt is as follows:

Sentence Generation Prompt.

```
Write exactly 25 unique English
sentences that you could say to
target. Each sentence must naturally
reflect the {honorific} level of
Korean honorifics when translated to
Korean and must only contain 1
sentence. Do not mention {target}
explicitly in the sentence. Output
only the sentence, one per line, with
no numbering, lists, or extra
commentary. Separate each sentence
with a newline.
```

B.3. Honorific Extractor

To automatically extract honorifics, we use GPT-OSS 120B. To validate our extractor, we sample a small subset of translations consisting of five implicit and five explicit sentences from each model, resulting in a total of 130 labeled sentences. An expert manually identified the honorific used in each sentence. We use the following prompt for our evaluator:

Honorific Extractor.

```
Given the following Korean segment:
'{response}'
Extract first the Korean utterance by
removing the Korean prefix of 'I was
talking to {addresse}, and I said:'.
Then, identify the honorific speech
style in the provided Korean
utterance by ONLY looking at the verb
endings. The possible choices are:
Casual (해), Polite (해요), Deferential
(합니다/하십시오) or Others (If nothing
fits). Be very short and give your
final answer at the end in the format
'Honorific:
<Casual/Polite/Deferential/Others>'"
```

Note that in the implicit scenario, we remove the instruction *"Extract first the Korean utterance"*. Furthermore, we include an "Other" option, but since it was selected in only about 5% of cases on average, we exclude it from our analysis.

Comparing our evaluator with the expert annotations yields an accuracy of 95%, which we consider sufficient for our purposes.

B.4. Hardware Details

All experiments are run on three H100 GPUs. Inference on the largest models takes about 15 minutes with a batch size of 64, while evaluation with GPT-OSS 120B takes roughly 1 hour 30 minutes with a batch size of 16. We use greedy decoding for all generations, limiting output to 128 tokens, except for the evaluator, which we allow up to 512 tokens.

Style and Name	Politeness	Formality	Example
합쇼체 (Hapsio-che; Deferential)	High	High	날씨가 춥습니다. nal-ssi-ga chub-seub-ni-da
해요체 (Haeyo-che; Polite)	High	Low	날씨가 추워요. nal-ssi-ga chu-wo-yo
하오체 (Hao-che; Semiformal)	Neutral	High	날씨가 춥소. nal-ssi-ga chub-so
하게체 (Hagae-che; Familiar)	Neutral	Low	날씨가 춥네. nal-ssi-ga chub-ne
반말체 (Banmal-che; Intimate)	Low	High	날씨가 추워. nal-ssi-ga chu-wo
해라체 (Haela-che; Plain)	Low	Low	날씨가 춥다. nal-ssi-ga chub-da

Table 5: **All Korean Speech Styles by Formality and Politeness.** The example phrase *"the weather is cold"* as realized in every Korean speech style (Hwang et al., 2021).

English name	*Korean Name*	*Declarative ending*[†]	*Formal/ Informal*	*Honorific category*
"deferential" style	*hapsyo-chey*	*-supnita*	Formal	Honorific
"polite" style	*hayyo-chey*	*-eyo*	Informal	
"semiformal" style	*hao-chey*	*-(s)o*	Formal	Authoritative
"familiar" style	*hakey-chey*	*-ney*	Formal	
"intimate" style"	*hay-chey*	*-e*	Informal	Non-honorific
"plain" style"	*hayla-chey*	*-ta*	Formal	

Table 6: **All Korean Speech Styles by Honorific Category.** The authoritative speech styles are rarely used by the younger generations of Korean speakers (Brown, 2015).

B.5. Prompts

For the `ITLM`, we use the following prompt:

`ITLM` Translation Prompt.

```
Translate the following {src_lang}
source segment into {tgt_lang}.
Return only the translation, without
any additional explanations or
commentary.
{src_lang}: {source_sentence}
{tgt_lang}:
```

For the `MTLM`, we use the recommended prompts from the authors.

C. Survey

C.1. Pilot Study

Before conducting the human survey, a pilot study was conducted. Two participants were given a translation based questionnaire, and two additional participants received a multiple-choice based questionnaire. Both pilot studies contained 18 sentences. The multiple-choice pilot study initially provided four possible answers: *Casual* (해), *Polite* (해요), *Deferential* (합니다/하십시오) and *unclear*. The results from the two pilot studies clearly showed that only looking at three speech styles was the right choice. Since only one out of twenty sentences was translated into a speech style other than the three expected styles and in the multiple-choice pilot study, no participant answered with *unclear*. Additionally, translating takes more time and requires better English skills than reading. Considering this, a multiple-choice version with only three possible answers was decided to be the most fitting for our purposes. A second multiple choice pilot study was conducted and timed, after which the final questionnaires were additionally adjusted to contain 45 questions and a more elaborate introduction was added. The introductions to the surveys can be found in Figure 3 and Figure 4.

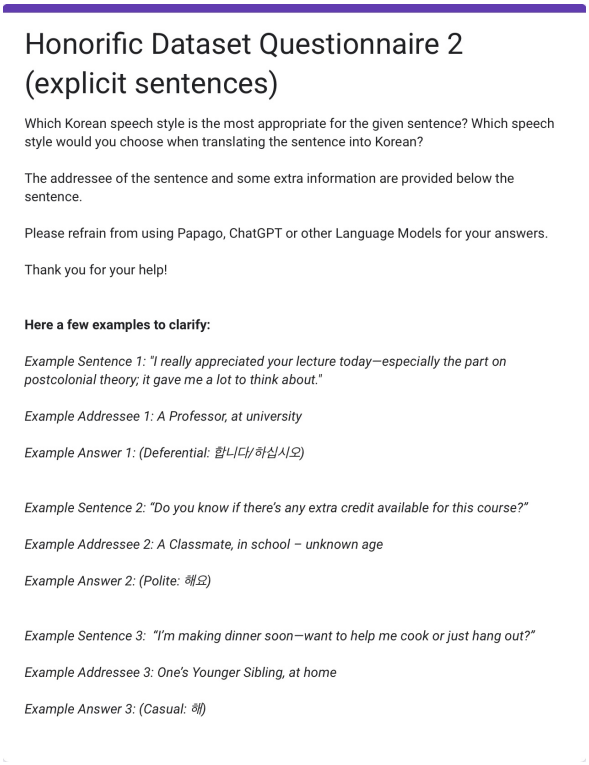


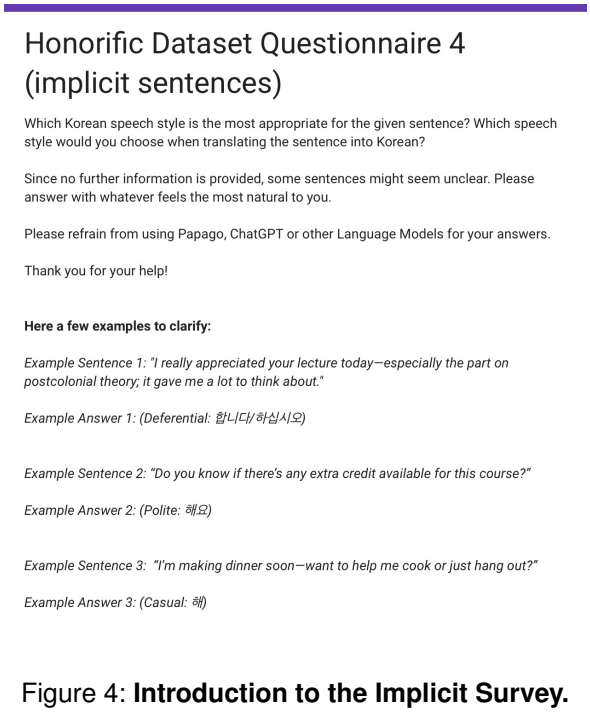
Figure 3: **Introduction to the Explicit Survey.**

C.2. Survey Design

There were 4 questionnaires covering 90 sentences in total, two questionnaires containing explicit sentences and two questionnaires containing implicit sentences. The differences in the two types of surveys can be seen in Figure 5 and Figure 6. Each questionnaire contained 4 questions about the age, gender and highest completed level of education of the participants. As seen in Figure 7, participants were asked about their age, gender and highest form of completed education, as those are clear, quantifiable metrics that have direct influence over speech style choice. Answering the final multiple-choice questionnaire took approximately 10 minutes.

Honorific	Addressee	Example Sentence
Deferential	One's professor	I've read the article you recommended, and I'd love to hear your thoughts on how it connects to the themes we discussed in class.
	A stranger	Hey, do you need a hand with that?
	A clerk, in a store	Excuse me, do you know where I can find the phone chargers?
Polite	A waiter	Hi, could I get a menu, please?
	A taxi driver	Is it okay if I roll the window down a bit?
	A classmate	Hey, did you understand the homework for today's class?
Casual	One's younger sibling	Don't forget to clean your room before Mom gets back.
	One's best friend	Man, this weather's perfect — we should do something spontaneous today.
	One's romantic partner	Come here, I need a proper hug.

Table 7: **Addressees and Example Sentences.** We present three representative addressees for each honorific level, each accompanied by one example sentence. The remaining addressees are as follows: **Deferential**—One's boss, one's in-laws (first meeting), a police officer, a government official, a group of students when giving a presentation, a job interviewer, a customer at one's company; **Polite**—One's teacher, a nurse, one's mother, one's in laws (already acquainted), a member of one's church, a co-worker, writing on an online forum; **Casual**—One's younger cousin, one's roommate, a classmate (well acquainted), a strange child, one's pet, chatting with Chat-GPT, talking with oneself


Honorific Dataset Questionnaire 4 (implicit sentences)

Which Korean speech style is the most appropriate for the given sentence? Which speech style would you choose when translating the sentence into Korean?

Since no further information is provided, some sentences might seem unclear. Please answer with whatever feels the most natural to you.

Please refrain from using Papago, ChatGPT or other Language Models for your answers.

Thank you for your help!

Here a few examples to clarify:

Example Sentence 1: "I really appreciated your lecture today—especially the part on postcolonial theory; it gave me a lot to think about."

Example Answer 1: (Deferential: 합니다/하십시오)

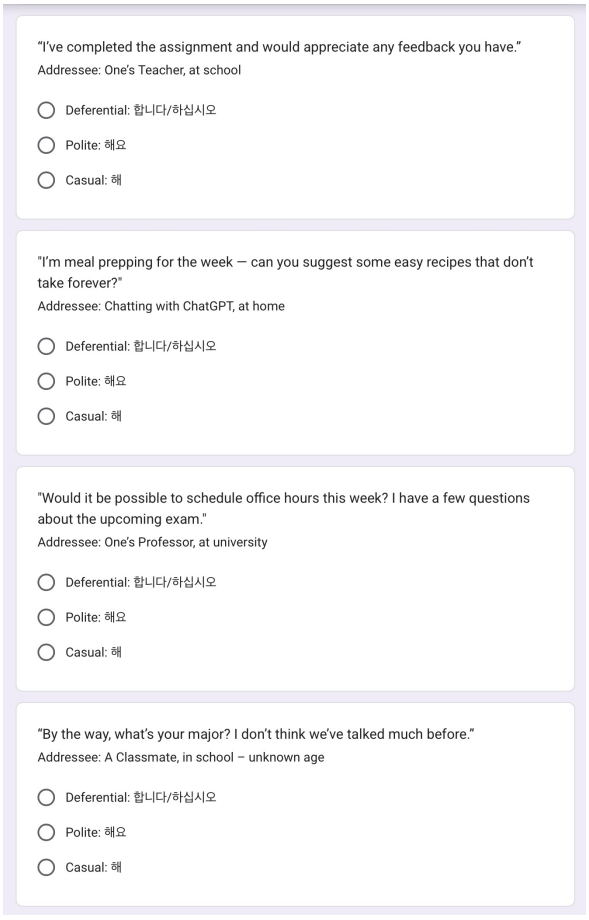
Example Sentence 2: "Do you know if there's any extra credit available for this course?"

Example Answer 2: (Polite: 해요)

Example Sentence 3: "I'm making dinner soon—want to help me cook or just hang out?"

Example Answer 3: (Casual: 해)

Figure 4: **Introduction to the Implicit Survey.**


"I've completed the assignment and would appreciate any feedback you have."
Addressee: One's Teacher, at school
Deferential: 합니다/하십시오
Polite: 해요
Casual: 해

"I'm meal prepping for the week — can you suggest some easy recipes that don't take forever?"
Addressee: Chatting with ChatGPT, at home
Deferential: 합니다/하십시오
Polite: 해요
Casual: 해

"Would it be possible to schedule office hours this week? I have a few questions about the upcoming exam."
Addressee: One's Professor, at university
Deferential: 합니다/하십시오
Polite: 해요
Casual: 해

"By the way, what's your major? I don't think we've talked much before."
Addressee: A Classmate, in school – unknown age
Deferential: 합니다/하십시오
Polite: 해요
Casual: 해

Figure 5: **Questions from the Explicit Survey.**

C.3. Survey Annotation/Evaluation

Of the 24 participants, one participant's results were declared invalid. To assess who seriously participated in a survey, it is common to use metrics like Cohen's kappa-score (CKS) (Hovy et al., 2013). But since there is no prescribed speech style for any sentence, using such a metric is problematic. The human results should also reflect the cultural and individual differences of the participants, so more than one speech style might be appropriate for any given sentence. As such, CKS was not used for eliminating unserious participants. Instead, the approach deemed most fitting for this survey was weighing each participant against the total speech style distribution. A participant was disqualified if the percentage of any chosen speech style digressed from the average speech style dis-

Model	Accuracy (%)
Implicit Scenario	
● gemma-2-9b-it	80.00
● Qwen2.5 72B Chat	77.78
▲ NLLB-200 3.3B	74.44
◆ Tower-Plus-72B	73.33
◆ Tower-Plus-9B	72.22
● Aya Expanse 32B	70.00
▲ NLLB-200 Distilled 600M	70.00
● Hunyuan-7B-Instruct	66.67
● Llama 3.3 70B Instruct	64.44
◆ GemmaX2 9B	57.78
◆ Hunyuan MT 7B	56.67
◆ Seed X PPO 7B	43.33
Explicit Scenario	
● Aya Expanse 32B	90.00
● gemma-2-9b-it	88.89
● Qwen2.5 72B Chat	86.67
◆ Hunyuan MT 7B	80.00
◆ Tower-Plus-72B	80.00
● Llama 3.3 70B Instruct	71.11
● Hunyuan-7B-Instruct	68.89
◆ Tower-Plus-9B	65.56
▲ NLLB-200 Distilled 600M	62.22
◆ GemmaX2 9B	58.89
▲ NLLB-200 3.3B	58.89
◆ Seed X PPO 7B	48.89

Table 8: **Model-Level Alignment with Human Judgments.** Reported values indicate the percentage of cases in which each model's output matches at least two human annotators' labels under implicit and explicit addressee scenarios. Blue (◆) denotes `MTLMs`, Orange (●) denotes `ITLMs`, and Green (▲) denotes `Seq2Seq` models.

tribution by 25%.

C.4. Survey Demographics

78.3% of participants identified as female and 21.7% identified as male. Based on the results in Figure 9, there was no significant difference in what speech style was chosen by which gender.

Age differences on the other hand had slightly more influence over speech styles, but not to a degree were any correlations between age and speech style choice can be made. At least not, with such a low number of participants. Figure 10 shows that the youngest participants belonged to the 18-24 years age group and the oldest participant to the 39-45 years age group.

As seen in Figure 11, the most influential demographic trait, according to the survey's results, was a participant's level of education. 39.1% of participants finished at least high school, 52.2% of participants had a bachelor's degree and 8.7% of participants had a master's degree or higher.

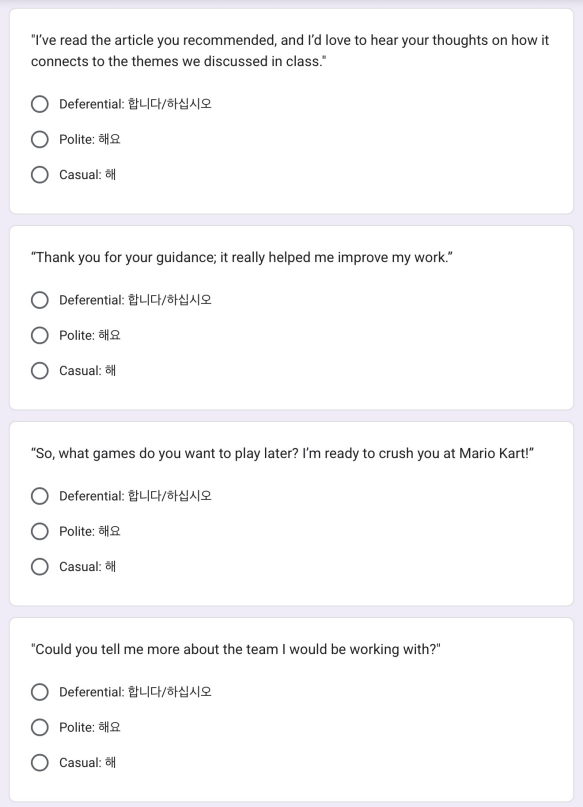


Figure 6: **Questions from the Implicit Survey.**

Less educated people were more likely to use the deferential speech style. An expected result considering that the deferential speech style is used to acknowledge another person's social, economic or academic status.

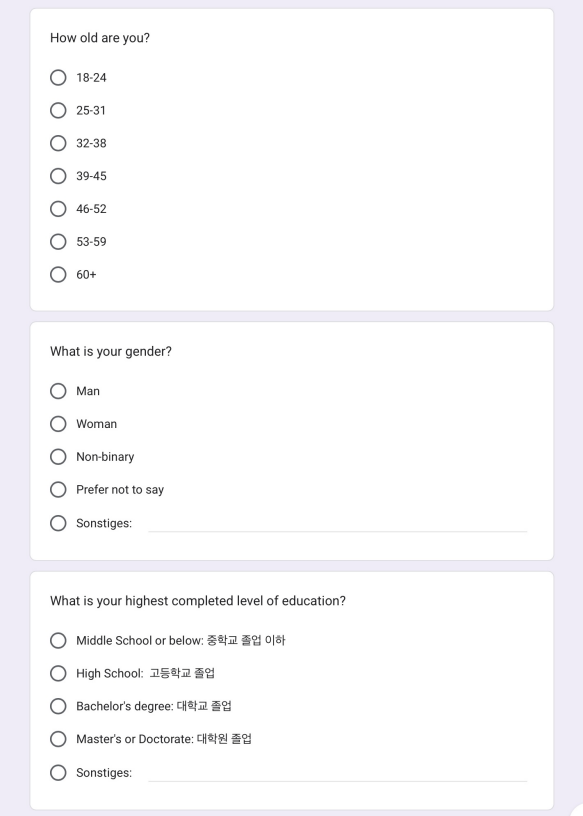


Figure 7: **Demographic Questions.**

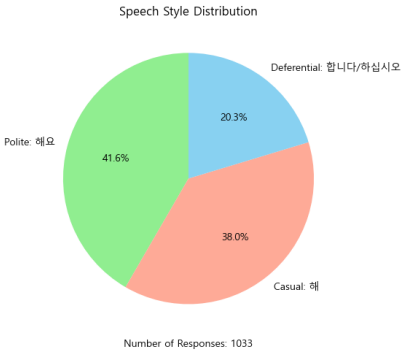


Figure 8: **Speech Style Distribution in the Human Survey.**

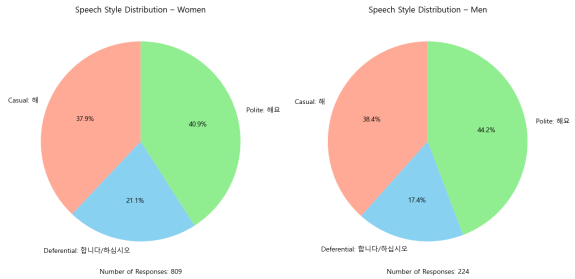


Figure 9: **Human Speech Style Distribution by Gender.**

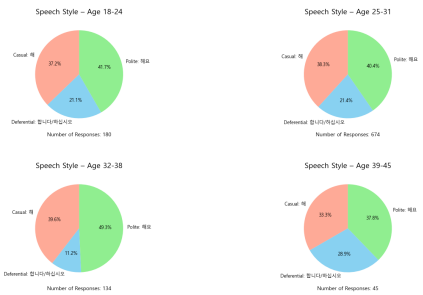


Figure 10: **Human Speech Style Distribution by Age.**

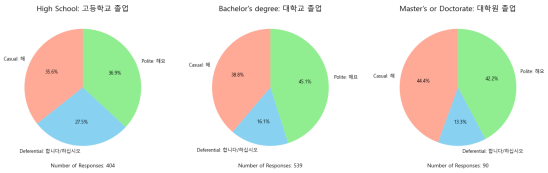


Figure 11: **Human Speech Style Distribution by Level of Education.**

C.5. License

Under OpenAI's Terms of Use, you own the outputs generated by GPT-5, see https://openai.com/policies/row-terms-of-use/. We therefore release the dataset (see Section 3) under the Creative Commons Attribution 4.0 International License (CC BY 4.0) with the human annotation acquired in Section 4.