A Mandarin-Cantonese Parallel Corpus with Formality Ranking

John S. Y. Lee[1], Qiong Wang[1]

[1]Department of Linguistics and Translation, City University of Hong Kong

Abstract

Formality-controlled machine translation allows users to specify the formality level of the target sentence, so that it would be suitable for the intended audience. While formality-annotated datasets have been constructed for some major languages, no such resource is currently available for Cantonese. This paper presents a Mandarin-Cantonese parallel corpus with 300 Mandarin sentences, each of which is aligned to a list of five or more Cantonese sentences ranked according to their level of formality. To our knowledge, this is the first parallel translation corpus with manual formality ranking, which provides more nuanced judgment than the formal/informal dichotomy in most current formality-annotated datasets. This corpus can support future research towards more fine-grained notions of formality in terminology, translation and text style transfer.

Keywords

formality ranking, formality-controlled machine translation, parallel corpus, Large Language Models, Cantonese

1. Introduction

It is important for a translated text to have a level of formality that is appropriate for the target audience. Consider the following sentences, ordered from the most formal to the least:

1. According to the staff, the lavatory was defective.
2. We were informed that the lavatory was not functioning.
3. They told us that the toilet wasn't working.
4. The loo's broken, that's what those guys said.

Sentence (1) would be appropriate for formal communication, for example reports and public speeches, while sentence (4) may be suitable for casual communication such as everyday conversations. Between these two extremes of the continuum of formality [1], sentence (2) or (3) could be the preferred choice in other contexts, such as polite conversations or social media posts.

Formality-controlled machine translation (FCMT) allows users to specify the formality level of the translation output [2, 3]. Since it is expensive to collect parallel bilingual data that include both formal and informal target sentences, FCMT can be challenging for low-resource languages. Recent initiatives, such as the Special Task on Formality Control for Spoken Language Translation, have provided formality-annotated datasets for some major languages [4]. No such resource has yet been developed for Cantonese, a variety of Chinese that has 85 million speakers worldwide [5]. With its lack of standard written forms, Cantonese expresses nuanced differences in formality using a considerable number of linguistic devices, including newly-coined Chinese characters, code-switching with English, and an elaborate system of sentence-final particles (SFPs). Table 1 shows paraphrases of the same sentence across the spectrum of formality through vocabulary choices and SFP usage.

This paper presents a Mandarin-Cantonese parallel corpus in which each Mandarin sentence is aligned to multiple formal and informal Cantonese sentences. Notably, these Cantonese sentences are manually ranked according to their level of formality. Most current language resources for FCMT and Formality Style Transfer (FST) [6] adopt the formal/informal dichotomy, which can hardly reflect the diversity of linguistic contexts in the continuum of formality [1]. To our knowledge, this is the first

4th International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2025, June 19-20, 2025, Thessaloniki, Greece.

jsylee@cityu.edu.hk (J. S. Y. Lee); wang.qiong@cityu.edu.hk (Q. Wang)

Language	Example sentence	Rank	Lexical differences			
Mandarin	不需要害怕癌症 '(You) do not need to be afraid of cancer'	n/a	不需要 *buxuyao* 'do not need'	害怕 *haipa* 'afraid'	癌症 *aizheng* 'cancer'	(No SFP)
↑ More ↑ formal	唔需要害怕癌症 '(You) don't need to be afraid of cancer'	(1)	唔需要 *m4 seoi1 jiu3*	害怕 *hoi6 paa3*	癌症 *ngaam4 zing3*	(No SFP)
	唔洗怕癌症 'Don't worry about cancer'	(2)	唔洗 *m4 sai2*	怕 *paa3*	癌症 *ngaam4 zing3*	(No SFP)
Cantonese	唔洗驚癌症 'Don't get scared by cancer'	(3)	唔洗 *m4 sai2*	驚 *geng1*	癌症 *ngaam4 zing3*	(No SFP)
	癌症使乜驚 'Cancer? Why get scared?'	(4)	使乜 *sai2 mat1*	驚 *geng1*	癌症 *ngaam4 zing3*	(No SFP)
↓ More ↓ informal	Cancer 咋嘛使乜驚呀 'Cancer my foot! Why get scared?'	(5)	使乜 *sai2 mat1*	驚 *geng1*	Cancer *ngaam4 zing3*	咋嘛; 呀 *zaa3 maa3*; *aa3* (SFP)

Table 1
Example Mandarin sentence and its Cantonese equivalents, ranked from formal (1) to informal (5). Lexical differences include terms at different formality levels and usage of sentence-final particles (SFPs)

attempt to annotate a parallel corpus with formality ranking, which can support research towards more nuanced gradations of formality.

The contribution of this paper is two-fold. First, we have constructed a parallel corpus with 300 Mandarin sentences and over 2000 Cantonese paraphrases with formality ranking. This corpus can facilitate more fine-grained FCMT and FST, as well as the development of formality-annotated lexica. Second, we describe and evaluate the use of Large Language Models (LLMs) for semi-automatic corpus construction, which can inform future efforts in creating similar corpora for other low-resource languages.

2. Research Background

2.1. Linguistic resources

Parallel corpora with formality-annotated target sentences have been constructed for a number of major languages. For example, the CoCoA-MT dataset [4], adopted by the Special Task on Formality Control for Spoken Language Translation, provides formal and informal translations from English into German, Spanish, Hindi, and Japanese. With the exception of the Japanese subset, which makes a three-way distinction ("formal", "polite", "informal"), all other languages have a binary annotation of formality only. Since the formal/informal dichotomy can hardly reflect the continuum of formality [1], there may be many linguistic contexts that call for translations between these two options.

Monolingual corpora of formal-informal sentence pairs are available for English, Brazilian Portuguese, French, and Italian [6, 7]. A more fine-grained, 7-point Likert scale on formality has been applied in annotating English [8] and German sentences of a variety of genres [9]. Since these datasets do not contain paraphrases of the same sentence at different formality levels, they could not be used directly in evaluating formality ranking of candidate translations.

As the target language in this paper, Cantonese is the "most widely known and influential variety of Chinese other than Mandarin" [10], which is generally considered to be standard Chinese. Though Cantonese and Mandarin share similar writing systems and many cognates, they are mutually unintelligible in their spoken form. Despite its 85 million speakers worldwide, Cantonese is a low-resource language [5]. With its lack of standard written forms, Cantonese expresses nuanced differences in

formality using a considerable number of linguistic devices, including newly-coined Chinese characters, code-switching with English, and an elaborate system of sentence-final particles (SFPs). Current Cantonese resources include monolingual corpora (e.g., [11]), English-Cantonese parallel corpora (e.g., [12]) and Mandarin-Cantonese parallel corpora (e.g., [13, 14, 15]). Since none has been annotated according to formality, they cannot support development of formality-controlled machine translation or formality style transfer. Table 1 shows paraphrases of the same sentence across the spectrum of formality through lexical differences (e.g., from formal to informal, 唔需要 *m4 seoi1 jiu3*, 唔洗 *m4 sai2*, 使乜 *sai2 mat1*) and SFP usage (e.g., 咋嘛 *zaa3 maa3*).

2.2. Formality-controlled machine translation (FCMT)

In a study on Mandarin-to-Cantonese FCMT, Wong and Lee [14] proposed a rule-based system to generate low-register and high-register output based on dictionary look-up and syntactic transformation. In a human evaluation, 62% of the output sentences were judged as satisfactory. Other FCMT approaches include the use of a "side constraint" to facilitate generation at different levels of formality [16, 17], for example by placing tags in front of the input sentence [2]. Pre-trained language models, such as mT5-large and mBART-large, have been fine-tuned to translate English into six languages, in both informal style and formal style [3]. Recent advances in AI have led to Large Language Models (LLMs) that offer superior performance in many NLP tasks. Zero-shot and one-shot prompting of LLMs have been demonstrated to produce high-quality translations with appropriate levels of formality [18]. Due to the lack of formality-annotated corpus, however, FCMT studies typically rely on manual evaluation or predictions of formality classifiers or regression models [19].

3. Dataset

Our Mandarin-Cantonese parallel corpus was constructed out of 300 Mandarin-Cantonese sentence pairs. These sentence pairs were drawn from a parallel corpus of Mandarin and Cantonese [13], which contains the Mandarin subtitles and Cantonese speech transcriptions of television programs broadcast in Hong Kong; and from a parallel treebank of Mandarin[1] [20] and Cantonese[2] [21], annotated with Universal Dependencies, based on Mandarin subtitles and Cantonese speech transcriptions of short films produced by undergraduate students in Hong Kong. The average sentence length is 13.7 characters in Mandarin sentences and 14.5 in Cantonese sentences.

4. Corpus Construction

We adopted a two-stage process to expedite corpus construction. In the first stage (Section 4.1), translation drafts were automatically generated with Large Language Models (LLMs). Subsequently, these drafts were annotated and edited by human judges (Section 4.2).

4.1. Generation of translation drafts

In order to generate translations with more diverse levels of formality, we implemented two FCMT approaches. The "Direct" method prompts an LLM to directly generate formal and informal translations. In contrast, the "Pipeline" method first performs formality-agnostic MT, and then applies Formality Style Transfer (FST) [6] to modify the level of formality of the MT output.

Direct method This method uses few-shot prompting to directly generate formal and informal target sentences with an LLM. The FCMT prompt (Table 2) incorporates ten example Mandarin-Cantonese sentence pairs, consisting of five formal and five informal samples. To generate formal

[1] https://universaldependencies.org/treebanks/zh_hk/index.html
[2] https://universaldependencies.org/treebanks/yue_hk/index.html

Task	Prompt (in original Chinese)	Prompt (English translation)
Formality-controlled machine translation (FCMT)	我要你擔任香港學校中文老師。你的工作就是將下列中文句子翻譯成<formality>，并且不需要提供拼音。 這裏是供參考的一些示例： 示例1: 普通話：<Mandarin> <formality>: <Cantonese> 示例2:	I would like you to be a Chinese teacher at a school in Hong Kong. Your task is to translate the following Mandarin sentence into <formality>, without any romanization. Here are some examples: Example 1: Mandarin: <Mandarin> <formality>: <Cantonese> Example 2: ... ...
Formality Style Transfer (FST)	我要你擔任香港學校中文老師。你的任務是將下列中文句子改寫成<formality>。 你需要解釋原文是如何改寫。我們一步一步來。你可以從如下兩個層面思考：1. 翻譯後的粵語句子應該屬於粵語口語還是粵語書面語。2. 翻譯後的句子中有哪些索表明它是<formality>？ <CoT sample> 請針對如下文本提供簡單的解釋，以便改寫成<formality>，並提供改寫後的結果。 輸入句子:<Cantonese>	I would like you to be a Chinese teacher at a school in Hong Kong. Your task is to revise the sentence into <formality>. You need to explain how the text is rewritten. Let us think step by step. You may think of the following two issues. (1) Is the rewritten sentence formal or colloquial? (2) What characteristics of the rewritten text make it <formality>? <CoT sample> Please give an explanation on how to revise the following sentence to <formality>, and provide the output. Input sentence: <Cantonese>

Table 2
Prompts used for formality-controlled translation and formality style transfer. <Mandarin> and <Cantonese> are example sentences in Mandarin and Cantonese, respectively; <formality> is replaced with either "formal Cantonese" (粵語書面語) or "informal Cantonese" (粵語口語); <CoT sample> refers to the demonstrations given for the chain-of-thought explanation.

outputs, the <formality> in the instruction is substituted with "formal Cantonese" (粵語書面語); to generate informal outputs, it is substituted with "informal Cantonese" (粵語口語).

Pipeline method This method performs formality-agnostic MT, and then revises the MT output to the desired formality level via FST. For formality-agnostic MT, the <formality> in the FCMT prompt (Table 2) is substituted with "Cantonese" (粵語), i.e., without any formality specification. For FST, the <formality> in the prompt (Table 2) is substituted with "formal Cantonese" to generate formal output, and "informal Cantonese" to generate informal output. The FST prompt uses the chain-of-thought strategy, which has been shown to produce higher-quality output than a number of competitive baselines [22].

We used GPT-4o, accessed through the Azure OpenAI Library, to implement the two methods above. For each of the 300 Mandarin sentences in our dataset (Section 3), the Direct method generated a formal output (henceforth, `Direct-formal`) and an informal (`Direct-informal`) output; the Pipeline method also generated a formal output (`Pipeline-formal`) and an informal output (`Pipeline-informal`). The sentence lengths were similar across these four output categories, with average length ranging from 14.1 to 15.0 characters. Through this procedure, each Mandarin sentence was aligned with five Cantonese paraphrases: four produced by GPT-4o, along with the Cantonese speech transcript (henceforth, `Transcript`) harvested from the original dataset (Section 3).

Method	Formality rank	% revised
`Transcript`	4.44	24.8%
`Direct-informal`	3.63	23.7%
`Pipeline-informal`	3.30	26.8%
`Direct-formal`	2.35	29.3%
`Pipeline-formal`	1.20	79.7%

Table 3
Average formality ranking (out of 5) of the Cantonese sentences, and the proportion of sentences that were revised. The larger the value of the rank, the more informal the sentence.

4.2. Annotation

We recruited 9 undergraduate students, all native speakers of Cantonese, to judge the degree of formality and quality of these Cantonese sentences. They were shown the Mandarin sentence, followed by the five Cantonese sentences in random order. Each output was independently annotated by two of the judges. They were instructed to perform the following two tasks:

Formality ranking Rank the five candidate translations from 1 (most formal) to 5 (most informal), without ties. An example is provided in Table 1.

Revision Determine if the candidate translation is acceptable. If not, edit the translation while preserving its level of formality.

5. Results

5.1. Inter-annotator agreement

In terms of formality ranking, the annotators attained a Kendall's τ of 0.6579. This level of correlation in ranking compares favorably with those in previous studies [7]. In terms of revision, we measured how often two annotators agreed on the need (or lack thereof) to edit the translation. Depending on one's strictness, it could be subjective to draw the line between acceptable and unacceptable translations. The annotators achieved a Kappa of 0.4116, which corresponds to 'moderate' level of agreement [23].

5.2. Analysis

Table 3 reports the average formality ranking of sentences obtained with different methods, and the proportion of Cantonese sentences that were deemed incorrect and therefore required revision. The larger the value of the average rank, the more informal the sentence.

Informal output. The Cantonese speech transcripts were considered to be most informal (average rank 4.44 out of 5). A substantial proportion (24.8%) of these transcripts were revised, mostly due to discrepancies in content with the Mandarin subtitles. The subtitles did not always include all details in the speech, perhaps due to constraints over screen size.

The Direct and Pipeline methods were both successful in producing informal Cantonese. The output of the Direct method was judged to be slightly more informal (rank 3.63) than that of the Pipeline (rank 3.30). The fact that only 23.7% to 26.8% of the output was edited suggests that these automatic methods do not necessarily require more manual revision effort than the transcripts. However, we found that LLM-generated content was less likely than the transcripts to include English code-switching and appropriate sentence-final particles. Their presence, which leads to more colloquial and engaging text, was often favored by judges for the informal style.

Formal output. The Pipeline method produced more formal output (rank 1.20) than the Direct method (rank 2.35). However, it required significantly more intervention, with 79.7% of the output edited by the judges. Our post-hoc analysis showed that this was mainly due to an overuse of Mandarin terms, making the sentence too formal to sound natural in Cantonese. Overall, these results suggest that LLMs could be effective in producing high-quality drafts for semi-automatic corpus construction.

6. Conclusions

We have presented the first Mandarin-Cantonese parallel corpus with formality ranking. For each Mandarin sentence, at least five Cantonese paraphrases are manually ranked according to their degree of formality. Most current approaches in formality-controlled machine translation (FCMT) and formality style transfer (FST) adopt the formal/informal dichotomy. This corpus can support future FCMT and FST research in exploring more fine-grained notions of formality in terminology and translation. Further, we have reported our experience with an LLM-based, semi-automatic approach for corpus construction. Evaluation results suggest that this approach is feasible, and may be considered for future development of formality-ranked corpora and lexica for other low-resource languages.

Acknowledgments

This work is partially supported by a Strategic Research Grant (project number 70006037) from City University of Hong Kong.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] F. Heylighen, J.-M. Dewaele, Formality of language: Definition, measurement and behavioral determinants, in: Technical report, Center "Leo Apostel", Free University of Brussels, Brussels, Belgium, 1999.

[2] X. Niu, S. Rao, M. Carpuat, Multi-Task Neural Models for Translating Between Styles Within and Across Languages, in: Proc. 27th International Conference on Computational Linguistics (COLING), 2018, pp. 1008–1021.

[3] E. Rippeth, S. Agrawal, M. Carpuat, Controlling Translation Formality Using Pre-trained Multilingual Language Models, in: 19th International Conference on Spoken Language Translation (IWSLT 2022), 2022.

[4] M. Nădejde, A. Currey, B. Hsu, X. Niu, M. Federico, G. Dinu, CoCoA-MT: A Dataset and Benchmark for Contrastive Controlled MT with Application to Formality, in: Findings of the Association for Computational Linguistics: NAACL 2022, 2022.

[5] R. Xiang, E. Chersoni, Y. Li, J. Li, C. Huang, Y. Pan, Y. Li, Cantonese natural language processing in the transformers era: a survey and current challenges, Language Resources and Evaluation (2024).

[6] S. Rao, J. Tetreault, Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer, in: Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2018, p. 129–140.

[7] E. Briakou, D. Lu, K. Zhang, J. Tetreault, Olá, Bonjour, Salve! XFORMAL: A Benchmark for Multilingual Formality Style Transfer, in: Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2021, p. 3199–3216.

[8] E. Pavlick, J. Tetreault, An Empirical Analysis of Formality in Online Communication, Transactions of the Association for Computational Linguistics 4 (2016) 61–74.

[9] E. Eder, U. Krieg-Holz, M. Wiegand, A Question of Style: A Dataset for Analyzing Formality on Different Levels, in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, p. 580–593.

[10] S. Matthews, V. Yip, Cantonese: A Comprehensive Grammar, Routledge, New York, 2011.

[11] O. Y. Kwong, Probing a Two-Way Parallel T&I Corpus for the Lexical Choices of Translators and Interpreters, New Perspectives on Corpus Translation Studies, New Frontiers in Translation Studies (2021) 101–132.
[12] V. Yip, S. Matthews, The bilingual child: Early development and language contact, Cambridge University Press, 2007.
[13] J. S. Y. Lee, Toward a Parallel Corpus of Spoken Cantonese and Written Chinese, in: Proc. 5th International Joint Conference on Natural Language Processing (IJCNLP), 2011, p. 1462–1466.
[14] T.-S. Wong, J. S. Y. Lee, Register-sensitive Translation: A Case Study of Mandarin and Cantonese, in: Proc. Association for Machine Translation in the Americas (AMTA), 2018.
[15] H. Y. Mak, T. Lee, Low-Resource NMT: A Case Study on the Written and Spoken Languages in Hong Kong, in: Proc. 5th International Conference on Natural Language Processing and Information Retrieval (NLPIR), 2021.
[16] Y. Wang, J. Zhang, F. Zhai, J. Xu, C. Zong, Three Strategies to Improve One-to-Many Multilingual Translation, in: Proc. EMNLP, 2018.
[17] G. Lample, S. Subramanian, E. Smith, L. Denoyer, M. Ranzato, Y. L. Boureau, Multiple-attribute Text Rewriting, in: Proc. ICLR, 2019.
[18] E. Marrese-Taylor, P. C. Wang, Y. Matsuo, Towards Better Evaluation for Formality-Controlled English-Japanese Machine Translation, in: Proc. 8th Conference on Machine Translation, 2023.
[19] E. Briakou, S. Agrawal, J. Tetreault, M. Carpuat, Evaluating the Evaluation Metrics for Style Transfer: A Case Study in Multilingual Formality Transfer, in: Proc. 2021 Conference on Empirical Methods in Natural Language Processing, 2021.
[20] H. Leung, R. Poiret, T. sum Wong, X. Chen, K. Gerdes, J. Lee, Developing Universal Dependencies for Mandarin Chinese, in: Proc. Workshop on Asian Language Resources, 2016.
[21] T.-S. Wong, K. Gerdes, H. Leung, J. S. Y. Lee, Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank, in: Proc. 4th International Conference on Dependency Linguistics (Depling), 2017, pp. 266–275.
[22] C. Zhang, H. Cai, Y. Li, Y. Wu, L. Hou, M. Abdul-Mageed, Distilling Text Style Transfer With Self-Explanation From LLMs, in: Proc. 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), 2024.
[23] J. R. Landis, G. G. Koch, The Measurement of Observer Agreement for Categorical Data, Biometrics 33 (1977) 159–174.