Emoji as an Elements of Politeness in ESL CMCs

Wanitchaya Poonpatanapricha
University of Chicago
wanitchaya@uchicago.edu

Abstract

This paper investigates whether a speaker whose primary language has higher degree of politeness (than English) is more likely to use emoji as elements of politeness when the speaker has to use computer-mediated communications (CMCs) to communicate politely in English. The results suggest that this hypothesis is true for speakers whose primary languages have very high degree of politeness (e.g., Japanese, Thai). These speakers are more likely to use emoji as well as use higher number of emoji than speakers whose primary languages are not as highly polite. It is still unclear whether this hypothesis is also true for speakers whose primary languages have moderate degree of politeness.

1 Introduction

Emoji, a form of ideograms, support computer-mediated communications (CMCs) in the same way nonverbal cues–such as facial expression, gesture, and tone of voice (Tang and Hew, 2019)–support face-to-face (FTF) communications. One way that emoji can support CMCs is to convey politeness. Some languages have more linguistic elements for politeness (e.g., honorifics in Japanese, sentence-final particles in Thai) than the others (e.g., English). It is therefore possible that when a speaker whose primary language has a higher degree of politeness has to communicate politely through texts in a less polite foreign language, that speaker might resort to emoji due to the lack of linguistic elements for politeness and nonverbal cues to convey politeness as in FTF. This paper investigates the possibility of emoji as an element of politeness in ESL[1] CMCs.

[1]English as a secondary language

2 Related Works

A number of research has linked the use of emoticons—the predecessor of emoji—with politeness strategies.

Sampietro (2016) proposed that one of the pragmatic function of emoticons was to mitigate possible face-threats. For example, emoticons were used with requests and orders to soften these speech acts (Dresner and Herring, 2010; Darics, 2012; Skovholt et al., 2014). Specifically, Darics (2012) found that emoticons were mainly used to mitigate or to clarify the message, usually to reach a successful cooperation. Here are two example sentences mitigated by emoticons from Dresner and Herring (2010):

> I would like a noncircumventing solution ;->
>
> I wonder if you could recommend me some good readings related to conversational data. We just collected some IM data and are about to conduct some analysis on it. Since I've never worked on this kind of data before, I am writing for some suggestions.:)

Furthermore, Sampietro (2019) proposed that emoticons may contribute to politeness in CMCs. Specifically, Skovholt et al. (2014) found emoticons to be positive politeness markers and rapport building devices, and Vandergriff (2013) found emoticons to be mostly used in the service of politeness and to mitigate disagreement.

In terms of cross-language analysis, prior works have studied emoticon usages for politeness across different languages. Komrsková (2015) studied emoticons in Czech and English, and found that phrases of greeting and thanks were very often accompanied vy emoticons in both languages. Kavanagh (2016) studied emoticon as a medium

for channeling politeness within American and Japanese online blogging communities and found that Japanese used emoticons significantly more than Americans. In contrast to these works which investigated usage of emoticons for politeness in each speaker's primary language, this paper investigates the usage of emoji for politeness in ESL CMCs.

3 Method

The goal of this paper is to investigate whether a speaker whose primary language has higher degree of politeness (than English) is more likely to use emoji as elements of politeness when the speaker has to communicate politely in English.

To do so, an ideal dataset would be a CMC which requires politeness, is in English, has emoji, and is a product of speakers with different primary languages. In addition, those languages should have varied degrees of politeness relative to English, and each language's degree of politeness could be clearly classified. If the hypothesis of this paper is correct, ESL speakers whose primary languages have high degrees of politeness should use emoji more than ESL speakers whose primary languages have lower degrees of politeness.

In this paper, the next best ideal data is used because it is not feasible to conduct a carefully conduct a controlled experiment. In particular, he ideal dataset was approximated from a CMC in English with emoji which some speakers' primary languages were known. The *politeness distinctions in pronouns* feature from WALS (Dryer and Haspelmath, 2013) was used as a proxy for each language's degree of politeness. Lastly, dialogue act (DA), which was automatically classified by a classifier trained on a corpus with known DA tags, was used as a proxy for politeness.

In terms of statistical test, the problem was set up as a classification task. The usage of emoji was measured in two different ways: 1) proportion of utterances with emoji and 2) number of emoji per utterance (that has emoji). An utterance was defined as a turn of a speaker in a chat type of CMC. An utterance did not need to be a complete sentence and could contain more than one sentence. The values from each measure were used as the labels of the response variable for the classification tasks respective to each measure. The classification models were trained using conditional inference trees (Hothorn et al., 2015) with the speaker's primary language's degree of politeness and the utterance's DA as input features to be partitioned.

3.1 Data

The data of a CMC in English with emoji was obtained from the official Discord server of Tsuki Adventure, a free-to-play mobile game. Discord is a proprietary freeware VoIP application and digital distribution platform designed for video gaming communities that specializes in text, image, video and audio communication between users in a chat channel.

What is special about this data is that each user in this community is asked to tag one's user profile with the language other than English that the user uses. Although the language tagging is not mandatory, there are significant number of the users in this community did the tagging. Hence, there is sufficient data to investigate the paper's question. Nevertheless, using this data required a strong assumption that the language tagged was the primary language of that user. If this assumption has been violated, it would be harder to get significant partitions by the degree of politeness feature in the models. Hence, this assumption did not compromise the statistical test.

The data was scraped from the official discord server of Tsuki Adventure on 02/21/2020 using a Python Discord scraper from Dracovian (2019). The total number of scraped utterances was 4,477: 448 ($\sim$ 10%) utterances contained at least one emoji while 4,039 ($\sim$ 90%) utterances contained no emoji. 2,628 utterances came from users with language tags. There were 12 different languages being tagged: Bahasa, Chinese, French, German, Japanese, Korean, Portuguese, Russian, Spanish, Tagalog, Thai, and Vietnamese.

3.2 Degree of Politeness

Politeness distinctions in pronouns feature from WALS (Dryer and Haspelmath, 2013) was used as a proxy for each language's degree of politeness. The scope of this feature is restricted to politeness distinctions in second person pronouns. Below are the 4 values for this feature with corresponding descriptions from WALS:

1. No politeness distinctions - Languages that were assigned this value have no personal pronouns in their paradigms which are used to express different degrees of respect or intimacy toward the addressee.

2. Binary politeness distinctions - Languages that were assigned this value have a paradigmatic opposition between one intimate or familiar pronoun of address and another one expressing respectful address.

3. Multiple politeness distinctions - Languages that were assigned this value have two or more degrees of politeness within a pronominal paradigm. Note that these systems are rare cross-linguistically.

4. Pronoun avoidance - The term "pronoun avoidance" describes a strategy of pronoun usage which has an effect on the overall shape of the paradigm. Languages of East and Southeast Asia such as Japanese, Burmese and Thai have a strong sensitivity to politeness in language usage and within their grammars. Speakers have to account for a variety of social distinctions linguistically. Social distinctions between speaker and hearer may reflect relative age, kinship, social ranking, intimacy, and other social features. From a linguistic point of view, one of the most important strategies of being polite is to avoid of addressing people directly.

Each utterance in the data was tagged with one value acording to the speaker's language tag. In addition, to allow comparisons between speakers whose primary languages have different degree of politeness (which will be interchagably called as *polite level* from this point onward for succincity), this paper assumed WALS' values for this feature to be ordinal, which is reasonable given how these values were linguistically classified. Table 1 summarises the polite level classification for the languages present in the data.

Note that English was also included in the classification. This is because there were many utterances without the language tags in the data. To make use of these utterances, the paper assumed the speakers of these utterances to have English as their primary languages. If this assumption has been violated, it would be harder to get significant partitions by the polite level feature in the model. Hence, this assumption did not compromise the statistical test.

3.3 DA Classifier

The DA classifier used in this paper was trained on the NPS chat corpus with `MaxentClassifier`, both from `NLTK`. The number of iterations was 100, the training accuracy was 0.977, and the testing accuracy was 0.971.

The NPS chat corpus was created by Forsythand and Martell (2007). It consists of 10,567 posts gathered from various online chat services and has dialogue-act tagged. The dialogue-act tags and corresponding examples from Forsythand and Martell (2007) are shown in Table 2.

When using this DA classifier on the paper's data, emoji were removed from each utterance before the utterance was input into the DA classifier.

3.4 Conditional Inference Trees

Conditional inference trees (CTree) is a framework for recursive partitioning which embeds tree-structured regression models into a well defined theory of conditional inference procedures (Hothorn et al., 2006). It is a regression model describing the conditional distribution of a response variable Y given the status of m covariates. In addition, a statistically motivated and intuitive stopping criterion is implemented in CTree: the algorithm stops when the global null hypothesis of independence between the response and any of the m covariates cannot be rejected at a pre-specified nominal level α.

The models were trained through `partykit::ctree` (Hothorn et al., 2015) in `R` with $\alpha = 0.05$. Polite level and utterance's DA were used as input features. Two CTrees had proportion of utterances with emoji as the response variable: one trained on all utterances while the other excluded utterances without the language tags. The other two CTrees had number of emoji per utterance (that has emoji) as the response variable: one with only polite level as the input feature while the other included both polite level and utterance's DA.

4 Results

As stated in 3, the usage of emoji was measured in two different ways: 1) proportion of utterances with emoji and 2) number of emoji per utterance (that has emoji).

4.1 Proportion of Utterances with Emoji

Figure 1 shows two CTrees with proportion of utterances with emoji as the response variable. CTree 1A included utterances without the language tags, which, as stated in 3.2, are assumed to

Polite Level	Politeness Distinctions in Pronouns	Languages
High	Pronoun avoidance	Japanese, Korean, Thai, Vietnamese
Medium High	Multiple politeness distinctions	Tagalog
Medium Low	Binary politeness distinctions	Bahasa, Chinese, French, German, Portuguese, Russian, Spanish
Low	No politeness distinctions	English

Table 1: The polite level classification for the languages present in the data

Classification	Example
Accept	yeah it does, they all do
Bye	night ya'all.
Clarify	i meant to write the word may.....
Continuer	and thought I'd share
Emotion	lol
Emphasis	Ok I'm gonna put it up ONE MORE TIME 10-19-30sUser37
Greet	hiya 10-19-40sUser43 hug
No Answer	no I had a roomate who did though Other 0
Reject	u r not on meds
Statement	Yay...democrats have taken the house!
System	JOIN
Wh-Question	11-08-20sUser70 why do you feel that way?
Yes Answer	why yes I do 10-19-40sUser24, lol
Yes/No Question	cant we all just get along

Table 2: NPS chat corpus' dialogue-act tags and corresponding examples

be English and classified as *Low* for polite level. On the other hand, CTree 1B excluded utterances without the language tags. In either case, the partition of polite level was the first partition and was significant ($p < 0.001$), with *High* polite level has higher estimated proportion of utterances with emoji ($\hat{y} = 0.204$) from the rest of polite levels ($0.005 \leq \hat{y} \leq 0.121$) . After the first partition on polite level, in both CTrees, polite levels lower than *High* were further significantly partitioned by DA.

4.2 Number of Emoji per Utterance

Figure 2 shows two CTrees with number of emoji per utterance (that has emoji) as the response variable. CTree 2A included only polite level as the feature while CTree 2B included both polite level and utterance's DA as the input features. Note that both CTrees excluded utterances without the language tags and hence had no *Low* polite level. From CTree 2A, the partition by polite level was significant ($p = 0.045$) with *High* polite level having higher estimated number of emoji per utterance ($\hat{y} = 1.588$) than *Medium high* and *Medium low* polite levels ($\hat{y} = 1.215$).

When DA was also added into the model as another input feature, there were significant partitions by DA, and the model represented the data better as CTree 2B had lower MSE than CTree 2A by approximately 14.5%. From CTree 2B, the first partition was by DA ($p < 0.001$) with *Emphasis* and *Other* having the highest estimated number of emoji per utterance ($\hat{y} = 2.316$). The second partition was by polite level ($p = 0.01$) with estimated number of emoji per utterance on *High* polite level ($\hat{y} = 1.484$) higher than the average estimate on the rest of polite levels ($\hat{y}_{avg} = 1.143$) . The last partition (on *Medium high* and *Medium low* polite levels) was by DA ($p = 0.016$) where *Accept, Emotion,* and *ynQuestion* had higher estimated number of emoji per utterance that the rest of DA.

5 Discussion

The CTrees in 4 suggest it is possible that a speaker whose primary language has higher degree of politeness (than English) may use emoji as elements of politeness when the speaker has to

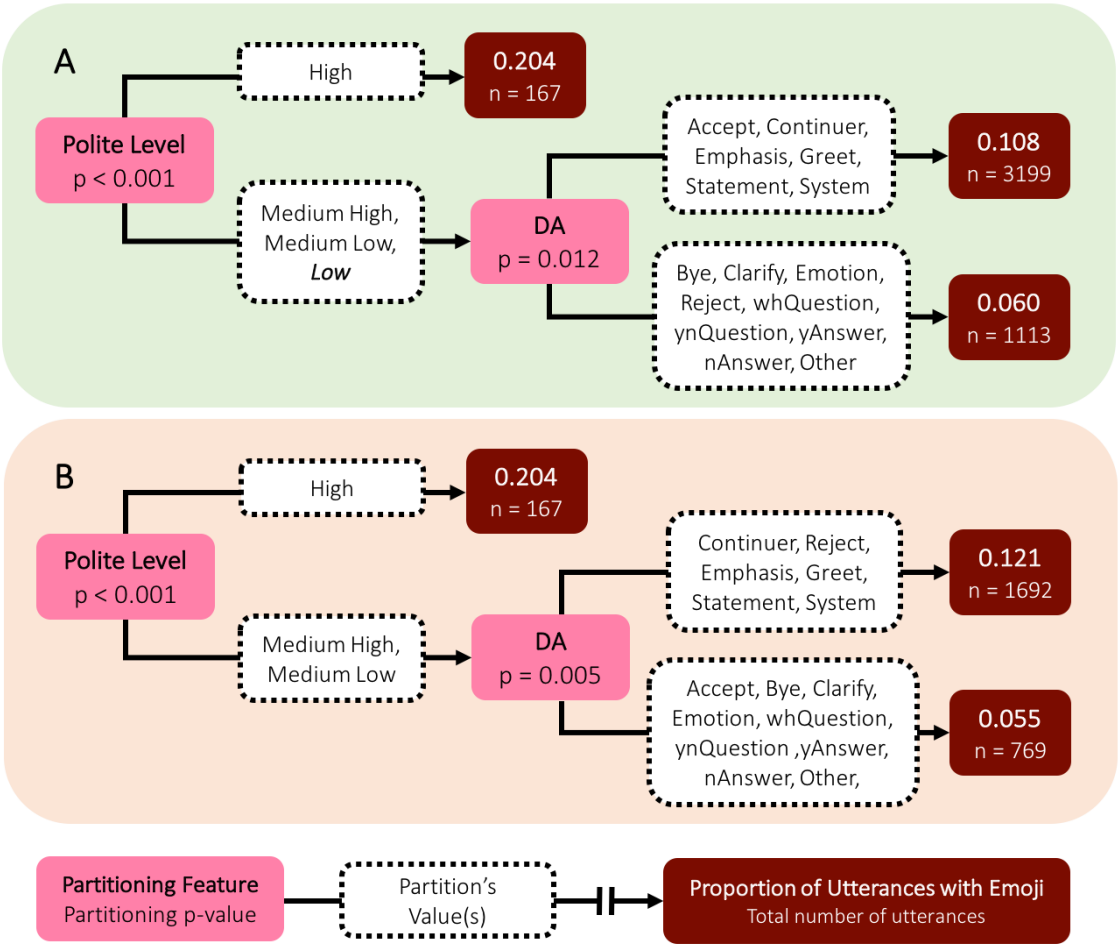

Figure 1: CTrees with proportion of utterances with emoji as the response variable. (A) included utterances without the language tags (assumed to be English and hence were classified as *Low* for polite level). (B) excluded utterances without the language tags.

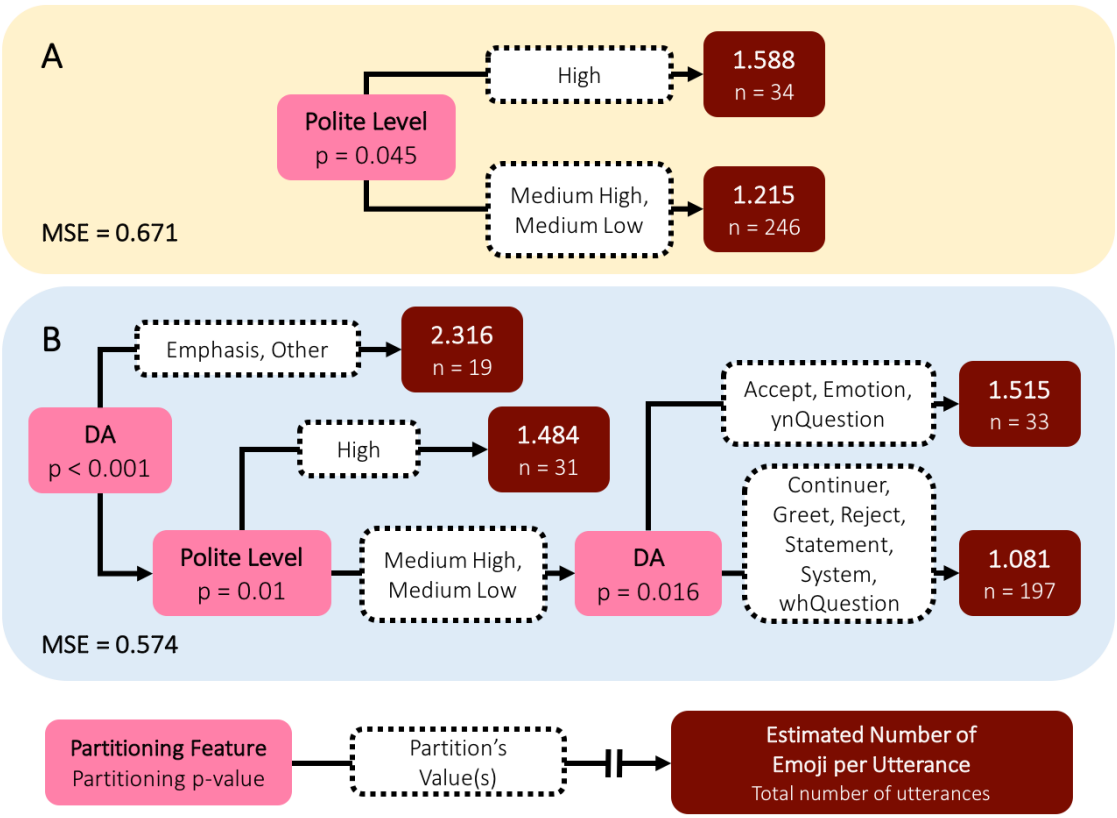

Figure 2: CTrees with number of emoji per utterance (that has emoji) as the response variable. (A) included only polite level as the input feature. (B) included both polite level and DA as the input features.

communicate politely in English.

Particularly, this hypothesis is significant for speakers whose primary languages are classified as *Pronoun avoidance* by WALS (*High* polite level in this paper). All CTrees estimate an utterance with *High* polite level to be more likely to have emoji as well as to have more emoji than other utterances with lower polite levels.

On the other hand, it is still unclear whether a speaker whose primary language is classified as *Binary* or *Multiple politeness distinctions* by WALS (respectively *Medium low* and *Medium high* polite levels in this paper) uses emoji as elements of politeness when the speaker has to communicate politely in English. The lack of significant partition separating *Low* polite level from *Medium low* and *Medium high* polite levels in CTree 1A could result from 1) the assumption made in 3.1 and 2) the lack of data from ESL speakers in *Low* polite level. First, the users without the language tags were assumed to use English as primary languages and hence were labeled as *Low* polite level. It is, however, likely that significant number of users without the language tags use other languages as their primary languages because the tagging is not mandatory. Second, out of the 12 tagged languages in this data, only English is classified as *No politeness distinctions* by WALS (and hence *Low* polite level). If there were other tagged languages classified as *Low* polite level, it would have been possible to make sense of the result of *Low* polite level versus *Medium low* and *Medium high* polite levels.

In addition, it is still unclear whether there is any difference between *Medium low* and *Medium high* polite levels in using emoji as elements of politeness in ESL CMCs. The lack of significant partition between *Medium low* and *Medium high* polite levels could result from 1) the lack of data and 2) the nature of classification in Politeness distinctions in pronouns feature. First, as stated in WALS, languages with *Multiple politeness distinctions* are rare. Indeed, there was only Tagalog in this class in the current data. Insufficient number of utterances with *Medium high* polite level may obscure possible differences in emoji usage from utterances with *Medium low* polite level. Second, how Politeness distinctions in pronouns feature is linguistically classified leads to imbalanced ordinal classes. The gap in degree of politeness between *Pronoun avoidance* and *Multiple politeness distinctions* is much bigger than the one between *Multiple politeness distinctions* and *Binary politeness distinctions.* Hence, in this paper's context, it might be more appropriate to combine *Medium low* and *Medium high* polite levels into one *Medium* polite level.

In terms of DA, there were significant partitions by DA, and using DA as an additional input feature improved the models' estimations. DA being important fit prior works' findings and this paper's idea of emoji having specific functions in certain types of speech acts (specifically for politeness), not just any random purposes. How DA was partitioned in 4, however, does not entirely resonate with prior findings and is hard to interpret. This is likely because the DA labels from NPS chat corpus were not specifically designed to capture or classify characteristics related to politeness. In addition, the DA classifier used was trained and tested on NPS chat corpus, not on the actual data in this paper. Hence, there likely were cascading errors on the actual data. It would be better to have a more parsimonious set of DA labels specifically designed for politeness as well as to hand-label, train, and test the DA classifier on the actual data.

6 Conclusion and Future Works

The goal of this paper is to investigate whether a speaker whose primary language has higher degree of politeness (than English) is more likely to use emoji as elements of politeness when the speaker has to communicate politely in English. The results suggest that this hypothesis is true for speakers whose primary languages have very high degree of politeness. These speakers are more likely to use emoji as well as use higher number of emoji than speakers whose primary languages are not as highly polite. It is however still unclear whether this is also true for speakers whose primary languages have moderate degree of politeness (still higher than English's). Future works should:

1. use a better and larger data where all speakers' primary languages are known and there are sufficient number of different languages,

2. create a more parsimonious set of DA labels specifically designed for politeness as well as hand-label, train, and test the DA classifier on the actual data, and

3. consider controlled experiments so that causations can be drawn.

References

Erika Darics. 2012. *Instant messaging in work-based virtual teams: the analysis of non-verbal communication used for the contextualisation of transactional and relational communicative goals.* Ph.D. thesis, Loughborough University.

Dracovian. 2019. Dracovian/discord-scraper.

Eli Dresner and Susan C Herring. 2010. Functions of the nonverbal in cmc: Emoticons and illocutionary force. *Communication theory*, 20(3):249–268.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Eric N Forsythand and Craig H Martell. 2007. Lexical and discourse analysis of online chat dialog. In *International Conference on Semantic Computing (ICSC 2007)*, pages 19–26. IEEE.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674.

Torsten Hothorn, Kurt Hornik, and Achim Zeileis. 2015. ctree: Conditional inference trees. *The Comprehensive R Archive Network*, pages 1–34.

Barry Kavanagh. 2016. Emoticons as a medium for channeling politeness within american and japanese online blogging communities. *Language & Communication*, 48:53–65.

Zuzana Komrsková. 2015. The use of emoticons in polite phrases of greetings and thanks. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, 9(4):1309–1312.

Agnese Sampietro. 2016. Exploring the punctuating effect of emoji in spanish whatsapp chats. *Lenguas modernas*, (47).

Agnese Sampietro. 2019. Emoji and rapport management in spanish whatsapp chats. *Journal of Pragmatics*, 143:109–120.

Karianne Skovholt, Anette Grønning, and Anne Kankaanranta. 2014. The communicative functions of emoticons in workplace e-mails::-. *Journal of Computer-Mediated Communication*, 19(4):780–797.

Ying Tang and Khe Foon Hew. 2019. Emoticon, emoji, and sticker use in computer-mediated communication: A review of theories and research findings. *International Journal of Communication*, 13:27.

Ilona Vandergriff. 2013. Emotive communication online: A contextual analysis of computer-mediated communication (cmc) cues. *Journal of Pragmatics*, 51:1–12.