Transformations in Chinese Greeting Expressions: A Diachronic Analysis of Question and Non-Question Forms

Zhen Li
zhenli.craig@gatech.edu

1 Introduction

Greetings are a type of speech act that serves to establish and maintain social relationships. They are a fundamental part of everyday communication, and their usage and form are influenced by social and cultural factors. This project investigates the evolving nature of Chinese greetings, focusing on the use of question forms like "你吃了吗?" (*Have you eaten?*) and "你好吗?" (*How are you?*), compared to non-question forms, "你好!" (*Hello!*) and "早上好!" (*Good morning!*). Contrary to the common comparison with "How are you doing?" in English, such question forms are less prevalent in Chinese greetings, especially among strangers.

By analyzing a conversational corpus alongside movie subtitles crawled from a Chinese movie resource website, we aim to understand how these greetings have changed over time and what these changes reveal about Chinese society and culture. This examination can be part of a broader inquiry into the nature of language as an evolving entity that mirrors social dynamics. In addition to a statistical approach, we also attempt to delve into the pragmatics and sociocultural significance of these greetings, contributing to our understanding of language's role in reflecting and influencing social interactions.

2 Literature Review

2.1 Greetings

(Duranti, 1997) proposed a operational definition for analyzing greetings across languages and the relationships between conventional expressions. Duranti's six criteria for alternative uses of daily expressions like greetings are instrumental in understanding the substantive and social purposes of these linguistic interactions.

(Gumperz, 2015) and (Boxer, 2002) delve into the broader field of sociolinguistics and interactional studies, providing foundational concepts and methodologies such as interactional sociolinguistics and conversation analysis that are essential for analyzing greeting behaviors in different social settings.

2.2 Previous Research on Chinese Greetings

(Xia et al., 2023)'s study provides a historical context, tracing the evolution of greeting culture in China from the 17th to the 20th century. It reveals a shift towards impersonalization in greetings, increased semantic informativeness, and a departure from traditional politeness norms of self-denigration and other-elevation. This historical perspective is crucial for understanding the current state of Chinese greeting practices.

(Liu, 2016) and (House et al., 2022) focus specifically on the comparison between English and Chinese greetings. Liu underscores the significance of greetings in social identity and cross-cultural communication, while House addresses the challenges Chinese learners of English face due to pragmatic differences in greeting conventions. House's inclusion of empirical studies adds depth to our understanding of these challenges.

(Bobgan, 2000) and (Qu & Chen, 2001) contribute to the understanding of demographic and linguistic specifics. Bobgan explores the influence of age and gender on responses to greetings in English, while 曲卫国 offers a detailed analysis of the linguistic form, topic, and pragmatic constraints of Chinese greetings, highlighting their openness, convertibility, and diversity.

Overall, these works collectively offer a comprehensive view of the sociolinguistic and pragmatic aspects of greetings in Chinese and English, emphasizing historical evolution, comparative analysis, and the influence of social and cultural factors on these everyday linguistic practices.

3 Data and Methodology

3.1 Research questions

In this project, we analyze 2 representative dataset for Chinese greetings. The first dataset is MAGICDATA Mandarin Chinese Conversational Speech Corpus (Yang et al., 2022). The second dataset is a set of movie subtitles for Chinese movies, ranging from 1960s to 2020s. They are obtained from Srtku, which is a website for downloading movie and TV show subtitles. These two datasets are chosen because they contains a representative set of daily conversations. The first dataset consists of 219,325 lines of speach and the second dataset consists of about 50 movies for each decade, enabling us to analyze the diachronic changes of Chinese greetings.

By comparing the usage of the greetings in the 2 datasets, we aim to answer the following three questions:

1. Are there any diachronic changes of greetings used in Chinese conversations?
2. If yes, what are the differences?
3. What are the socio-historical motivations underlying these differences?

3.2 Data Collection

In this study, we intend to extract the greetings from the coversation a corpus and movie subtitles and count their frequency.

The corpus MAGICDATA Mandarin Chinese Conversational Speech Corpus includes 180 hours of Mandarin Chinese speech from 633 speakers. All the transcripts are combined into a 219,325 line file for further analysis.

The movie subtitles are crawled from Srtku. We use the movie lists from Douban, an online movie database that provides rich filters for movie search. This website provides recommendations for movies from different decades, which enabled us to perform a diachronic analysis over a representative set of movies. For each decade from 1960s to 2020s, we select the recommended 200~300 movies and download their subtitles from Srtku. After cleaning, the final dataset contains 828 readable movies subtitles in total:

1960s	1970s	1980s	1990s	2000s	2010s	2020s
40	102	202	173	91	143	77

Table 1: Number of movies for each decade

3.3 Greeting Extraction

We use a naïve regular expression approach to extract the greetings from the corpus and movie subtitles. For each category of greetings (non-question forms and question forms), 2 regular expressions are designed to match the most frequently used greetings in daily conversations. The regular expressions are listed in the table below.

Category	Python syntax for the regular expression
Non-question forms	`r'((?<!\p{Han})(你|您)好[啊]?(?!\p{Han}))'`
	`r'((?<!\p{Han})(上午|下午|晚上|中午|早上)好[啊]?(?!\p{Han}))'`
Question forms	`r'((?<!\p{Han})(你|您)(最近)?好[吗么没嘛啊](?!\p{Han}))'`
	`r'(吃[过]?[饭]?了[吗么没嘛啊](?!\p{Han}))'`

Listing 1: Regular expressions for matching greetings

Take the first regular expression as an example, there are 5 major parts:

- `(?<!\p{Han})` : matches the position where the previous character is not a Hanzi character.
- `(你|您)` : matches either `你` or `您` .
- `好` : matches `好` .
- `[啊]?` : matches `啊` or nothing.
- `(?!\p{Han})` : matches the position where the next character is not a Hanzi character.

The leading and trailing `(?<!\p{Han})` and `(?!\p{Han})` are used to avoid disambiguation from the sentences like "你好棒" (*You are so amazing*). We don't force the last one `吃[过]?[饭]?了[吗么没嘛啊]` to have a leading non-Hanzi character because there exist several variations with different leading texts, such as "你已经吃过饭了吗?" (*Have you already eaten?*), and it's probably sufficient to match the similar semantics with the trailing constraint only.

Full width puctuations, such as `。` and `？`, are not matched by `\p{Han}` .

4 Result and Discussion

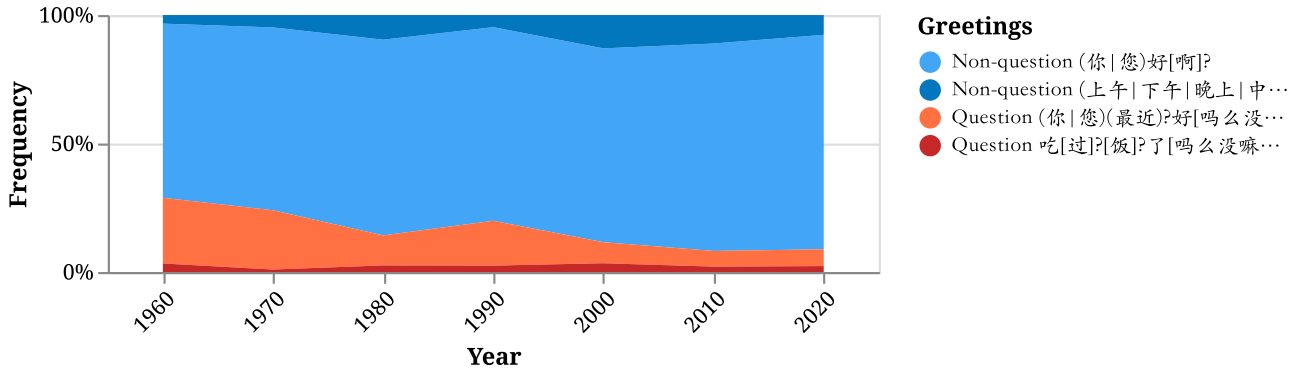

Figure 1: Changes of the frequency of greetings in the movie subtitles from 1960s to 2020s

Greeting	Matches
你好	
你好吗 or 吃了吗	

Table 2: Regular expressions for matching greetings

5 Conclusion

Bibliography

Bobgan, J. E. (2000). *A sociolinguistic investigation of the response variable to the question "How are you doing?"* (p. 86). https://www.proquest.com/dissertations-theses/sociolinguistic-investigation-response-variable/docview/231356778/se-2

Boxer, D. (2002). *Applying sociolinguistics: Domains and face-to-face interaction* (Vol. 15). John Benjamins Publishing.

Duranti, A. (1997). Universal and Culture-Specific Properties of Greetings. *Journal of Linguistic Anthropology*, *7*(1), 63–97. https://doi.org/https://doi.org/10.1525/jlin.1997.7.1.63

Gumperz, J. J. (2015). Interactional Sociolinguistics A personal perspective. *The Handbook of Discourse Analysis*, 309–323.

House, J., Kádár, D. Z., Liu, F., & Liu, S. (2022). Greeting in English as a Foreign Language: A Problem for Speakers of Chinese. *Applied Linguistics*, *44*(2), 189–216. https://doi.org/10.1093/applin/amac031

Liu, L. (2016). Different Cultures and Social Patterns Matter in English and Chinese Greetings. *Literacy Information and Computer Education Journal*, *7*. https://doi.org/10.20533/licej.2040.2589.2016.0310

Qu, W., & Chen, L. (2001). 汉语招呼分析 *(An Analysis of Chinese Greetings)*.

Xia, D., Huang, L., & Yang, T. (2023). A diachronic study of Chinese greetings between new acquaintances. *Journal of Pragmatics*, *217*, 156–171. https://doi.org/https://doi.org/10.1016/j.pragma.2023.08.007

Yang, Z., Chen, Y., Luo, L., Yang, R., Ye, L., Cheng, G., Xu, J., Jin, Y., Zhang, Q., Zhang, P., & others. (2022). Open Source MagicData-RAMC: A Rich Annotated Mandarin Conversational (RAMC) Speech Dataset. *Arxiv Preprint Arxiv:2203.16844*.

Appendix

All the code and data used in this project can be found at the repository sociolinguistics-greeting-analysis.

Obtaining and Analyzing Movie Subtitle Data

1. Collect the movie names from Douban Explore to filter the movies. Paste the scripts in the browser console and it will automatically click the expand button for 10 times and print 200 ~ 300 movie names in the console.
2. Run `movie_caption_crawl.ipynb` to crawl the movie subtitles. Switch IP if blocked by the website.
   - The movie names are only fuzzy matched in the search box of zimuku.net. Some manual work is needed to remove the wrong matches.
3. Run `clean.ipynb` to clean the downloaded subtitles.
4. Run `encoding.ipynb` to convert the subtitles from `UTF-8 with BOM` / `GB2312` to `UTF-8` encoding.
5. Run `stat.ipynb` to count the frequency of greetings by regex matches in the subtitles.
6. Run `eda.ipynb` to aggregate the results.
7. Copy the generated `regex_sum.json` to `visualization.html` and open it in a browser to visualize the results.

Data Result