# Improving Menu Localization with Culture-Aware Language Models

*Zhonghe Zhang*

4th Year Project Report
Artificial Intelligence and Computer Science
School of Informatics
University of Edinburgh

2024

# Abstract

Cross-cultural communication heavily relies on translating culturally specific texts, posing significant challenges for many machine translation (MT) systems in handling Culture Specific Items (CSI). This difficulty becomes particularly acute in the context of translating Chinese culinary terminology, where MT systems would produce nonsensical translations, exemplified by phrases such as ”Ants Climbing Tree” for ”Sauteed Vermicelli with minced Pork” or ”Squirrel Fish” for ”Sweet and Sour Mandarin Fish”. In this research, we introduce the most comprehensive Bilingual Chinese-English Menu (BiMenu) Dataset to date, along with an innovative automatic menu layout parser pipeline engineered to extract textual content from photographs of menus. Furthermore, this study investigates the impact of CSI on translation processes, introducing a three-tiered novel categorisation to examine its complexity in greater depth. We utilised our dataset to assess the performance of the strongest commercial MT system (Google Translate) and Large Language Model (LLMs)-based MT systems(GPT3.5 and GPT 4). To identify CSIs within texts, we devised an innovative automatic CSI identification pipeline equipped with three metrics, which has proven effective in identifying CSIs. Moreover, we introduced a recipe-based translation strategy integrating two prompts to incorporate external recipe knowledge into the translation process. Our results indicate that CSIs substantially influence translation tasks, and integrating recipe information can markedly improve the translation accuracy of CSIs, especially for dishes with abstract names.

# Research Ethics Approval

This project obtained approval from the Informatics Research Ethics committee.
Ethics application number: 2024/125449
Date when approval was obtained: 2024-03-04
The participants' information sheet and a consent form are included in the appendix.

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Zhonghe Zhang*)

# Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Alexandra Birch-Mayne, for her continuous guidance and support throughout the academic year. I would also like to thank Vivek Iyer for listening and giving important advice for my research.

Last but not least, I would like to thank my girlfriend, Yu Meng, for listing my ever-emerging ideas and giving feedback, and my friends who helped me evaluate the datasets. More importantly, my family for their constant support throughout the year.

# Table of Contents

# Chapter 1

# Introduction

## 1.1 Motivation

Cultures, whether understood as tangible materials or abstract concepts, imbue languages with a depth that extends beyond mere words, complicating the task of translation. This complexity has been further exacerbated in the context of Culture-Specific Items (CSIs), introduced by Álvarez and Vidal [1996], defined **CSI as "concepts that are specific to a specific language or group"**. Further work on CSI has been a long-standing part of translation studies, including Persson [2015], Sveučilište u Zadru [2024]. However, existing methodologies and datasets for terminology translation have predominantly focused on general domains such as medicine and finance, as highlighted in Dinu et al. [2019a], Ghazvininejad et al. [2023]. Unlike general terms, the nuanced characteristics of CSIs can render their literal translations challenging for individuals from different cultures to understand. These approaches often overlook the intricate details of cultural-specific aspects, leaving a significant gap in translation studies.

Moreover, identifying and adapting to cultural differences in language usage is crucial and challenging, as discussed in Hershcovich et al. [2022]. For example, many CSIs

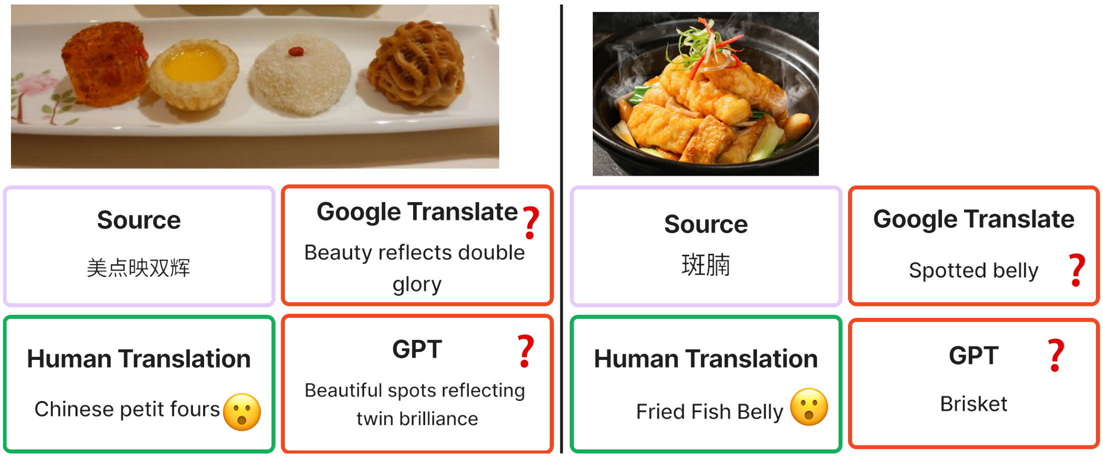


Figure 1.1: Cultural translation errors made by Google Translate and ChatGPT systems

have no direct translations into other languages, resulting in literal translations often failing to accurately convey the intended meaning due to these cultural differences, as discussed in works by Akinade et al. [2023a], Liebling et al. [2022]. Extensive examples can be found in Chinese cuisine, where the names of dishes may derive from their appearance, scent, flavour, background stories, or historical significance, requiring readers with cultural background knowledge to understand, as discussed in Huo et al. [2020], Amenador and Wang [2022a]. Literal translations are often opaque or odd: a Chinese dish (literally, ‘Three Freshnesses of the Earth’) can be adapted in translation to ‘Stir Fried Eggplant, Potato and Pepper’ in English. Other well-known examples include “Ants Climbing a Tree” (Sauteed Vermicelli with minced Pork) or “Lion’s Head Meatballs” (Braised Pork Meatballs), where the metaphorical essence and cultural context are lost. This scenario underlines the need for translations that do more than merely convert words; they must also convey cultural meanings in the target language and culture.

Recent advancements in MT have introduced innovative methods to incorporate cultural contexts into machine translations in light of the complex relationship between culture and language. Among these, Large Language Models (LLMs) have been leveraged to incorporate cultural knowledge into translation tasks using prompting techniques. This approach aims to bridge the cultural gaps that often arise in translation, especially in text with CSIs. Despite the potential for enhanced cultural sensitivity, the accuracy of LLM-based translations is highly dependent on the effectiveness of the prompting strategies employed. Furthermore, there is a tendency for these models to generate incorrect or “hallucinated” outputs, as evidenced in Ji et al. [2023]. This issue is exemplified by an incorrect translation of the CSI ”Fried Fish Belly” into ”Brisket (Cow Belly)”, as shown in Figure 1.1, illustrating a misalignment in cultural and contextual understanding.

In this study, we took a step closer to the challenges of translating texts with CSIs, drawing on a dataset we collected to explore how to identify texts that are CSI in the culinary domain and produce translations that are both linguistically accurate and also culturally. We aim to bridge the gap in understanding and translating CSIs, thereby contributing to developing more effective and culturally aware machine translation systems.

## 1.2 Objectives

This research aims to analyse and investigate how to produce a more culturally accurate translation of the Chinese culinary domain text. Therefore, the primary research question is:

> How can we translate Chinese culinary texts that are culturally accurate and understandable to the target culture?

To comprehensively address this overarching question, the study is partitioned into subsidiary objectives. These objectives examine the various methods, each a standalone approach, that together contribute to the enhancement of culturally nuanced translations.

**Research Question 1:** In what ways does the incorporation of CSI influence the performance and outcomes of machine translations in NMT systems and LLMs?

**Research Question 2:** Which categories of CSI are most problematic for identification and translation, and what does this reveal about the intrinsic characteristics of these categories?

**Research Question 3:** What are the most effective methods to translate CSIs into another language while being culturally accurate and understandable to the target culture?

## 1.3 Contributions

The main contributions of this research are outlined below:

### 1.3.1 BiMenu Dataset

We have collected the largest parallel Bilingual Chinese-English (BiMenu) Restaurant Menu dataset to date, surpassing the scale of existing datasets in this domain, with 4,562 human-verified dish entries, comparing with Lim [2018] with 3,606 entries, Chinese official approved translation Zhao [2008] with 2,860 entries, Zhao [2010] with 2,882 entries. This dataset is distinguished not only by its size but also by the depth and variety of information it encompasses. It includes essential details such as pricing, allergen information, vegetarian options, and, for a subset of entries, detailed descriptions. A unique aspect of our dataset is its incorporation of structural information about each dish in menu page photos, including images of the menu pages and bounding box coordinates for each dish name. This feature is important to advance research in the Document Layout Analysis field by providing a rich resource for studying the visual and textual organization of menu information.

Additionally, we developed an automation pipeline capable of parsing restaurant menu photos to extract aligned dish names in Chinese and English with near-perfect accuracy. This pipeline represents a significant advancement, as it can be applied to menus across languages, offering a universal menu analysis tool. We provide it open source at this link: github.com/Henry8772/BiMenu.

### 1.3.2 CSI Analysis and Translation

We further performed a series of research and investigations on the BiMenu dataset we collected to address the challenge of identifying and translating CSI within the context of Chinese cuisine. Below are the contributions:

- **CSI Annotation:** We have devised and implemented a human annotation guideline for identifying CSIs within the names of Chinese dishes. This methodology leverages four metrics: back-translation fidelity, cultural uniqueness, historical story presence, and cultural context relevance. Utilizing this framework, we annotated 400 dish names within our dataset, employing a team of three native Chinese speakers to ensure the annotations' authenticity and cultural sensitivity.

- **CSI Categories:** We have developed three distinct categorizations for CSI to analyse the impact of CSI on translation. These categories are designed to reflect varying levels of difficulty and complexity in translating CSIs, delineated as follows: (1) CSI with Common Words, (2) CSIs with Figuratively Combined Meanings, and (3) CSIs with Abstract Combined Meanings. To facilitate a thorough analysis, we classified 150 dishes into these specified categories from unanimous consensus among all three annotators.

- **CSI Identification:** We have introduced an innovative automatic pipeline for CSI identification, demonstrating superior performance compared with LLMs, including GPT 3.5 and GPT 4, across most CSI categories. This pipeline was tested against our annotated dataset, providing a thorough analysis of each metric used for CSI identification.

- **CSI Translation:** We have proposed a novel method, Recipe-based Translation, to improve the translation accuracy of these items. Utilizing recipes as external knowledge, we empower LLMs to infer the meanings of CSIs within the context of a recipe. This method leads to a substantial improvement in the quality of translation, particularly for dish names with abstract meanings.

- **CSI Evaluation:** We conducted evaluations of both automatic metrics (such as COMET, BLEU, ChrF, and ROUGE-L) and human judgment. Ten native Chinese speakers fluent in English participated in reviewing the quality of our recipe-based translations, confirming the effectiveness of our approaches.

# Chapter 2

# Literature Review

This chapter discusses the background knowledge that will help the reader understand Culture-Specific Items (CSI), Large Language Models (LLMs) for translations, and cross-cultural translations. It also includes literature reviews of related work in cross-cultural translation via large language models.

## 2.1 Chinese Culture and Cross-Cultural Translation

Chinese cuisine, with its vast and illustrious history, stands as a testament to the depth and diversity of culinary arts. As Mu (2010) highlights, the complexity of Chinese gastronomy is unmatched, largely due to the variety of preparation methods and the wide range of natural ingredients used. This richness is not just culinary but also cultural, with many dishes bearing names linked to historical events and figures. For example, a Chinese dish ”Ants Climbing a Tree”, is a dish with a name that whimsically evokes the image of ants ascending the branches of a tree, which is reflective of the ground meat that clings to the strands of glass noodles, resembling ants. This playful imagery is a hallmark of the inventiveness and symbolism often found in Chinese dishes. The dish itself is a stir-fried masterpiece, typically consisting of mung bean vermicelli seasoned and cooked with minced pork and various condiments, giving it a distinctive flavour that is both spicy and savoury. The direct translation of the dish's name might be puzzling to those unfamiliar with Chinese cuisine, but the English name provided, "Sautéed Vermicelli with minced Pork," describes the main ingredients and cooking method, making it more understandable to a global audience while losing the poetic and visual elements of the original name. This illustrates the need to balance between maintaining

| Chinese Names | Direct Translation | English Name |
|---|---|---|
| 蚂蚁上树 | Ants Climbing a Tree | Sauteed Vermicelli with minced Pork |
| 美点映双辉 | Beauty reflects double glory | Chinese petit fours |

Figure 2.1: Chinese dish name with direct translation and actual translation

cultural significance and ensuring comprehensibility in cross-cultural translation

This culinary diversity highlights the challenge of translating Culture Specific Items (CSI), where words or phrases are unique to a specific culture only , discussed in Álvarez and Vidal [1996], Yao et al. [2024]. This underscores the complex relationship between language and culture, as discussed by Gee [2017] and Kramsch [2014]. Direct translations of Chinese culinary names frequently result in opaque and odd; for example, "Beauty reflects double glory", which is a direct translation from its Chinese name, whereas its English localised translation, "Chinese petit fours," lacks a direct correlation with the original Chinese name. This accurate translation depends on the cultural background knowledge and linguistic proficiency of native speakers, requiring translators to study multiple dictionaries and connect with the cultural context to interpret allusive meaning. The complexities of Chinese culture extend this challenge to other domains, including idioms, proverbs, poetry, and literary allusions, each imbued with meanings that demand a deep understanding of the cultural and contextual background knowledge for effective translation.

## 2.2 Culture Specific Translation

The emergence of Neural Machine Translation (NMT), as introduced by Sutskever et al. [2014a], Bahdanau et al. [2014], represented a notable advancement in machine translation through the utilization of deep learning for language representation. This was achieved through the introduction of sequence-to-sequence learning (proposed in Sutskever et al. [2014b]) and attention mechanisms (proposed in Vaswani et al. [2023]). Nevertheless, translating texts embedded with cultural nuances remains a significant challenge. This challenge primarily stems from the inherent discrepancies in cultural contexts associated with different languages, as examined by Liebling et al. [2022]. They found the biases stemming from users' diverse cultural backgrounds, underlining the importance of bridging the cultural gap in translation efforts. Moreover, Akinade et al. [2023b] illustrates the limitations of NMT models, such as NLLB (Costa-jussà et al. [2022]) and Google Translate, in translating culturally-specific greetings in low-resource African languages.

The discrepancy between the cultural norms inherent in different languages constitutes significant challenges, often resulting in translations that fail to convey original meaning or cultural essence, as discussed in Das [2020]. There has been considerable progress in NMT research focused on enhancing translation quality within specific domains, such as medicine and finance. Several enhancements have been made in the field of domain-specific translation. For instance, terminology constraints for domain-specific languages are created in Dinu et al. [2019b]. Additionally, in Khandelwal et al. [2020], the utilization of cached data stores has been employed to enhance word selection in domain-specific tasks. Furthermore, domain adaptation techniques have been developed for out-of-domain corpora, as discussed in Hu et al. [2019]. Lastly, in Hu et al. [2022], which utilized pre-training NMT models to improve the translation of named entities using monolingual data. Unlike general terms in domain-specific, CSIs often lack direct equivalents in other languages, complicating their translation and making them difficult for people from different cultural backgrounds, as discussed in Yao et al. [2024].

Additionally, many Chinese cuisines have figurative or abstract names, which exceed the limits of NMT's literal translation. This requires a profound understanding of the source culture and the ability to infer meanings or metaphors beyond the literal sense.

## 2.3 Cultural Awareness in Large Language Models

Large Language Models (LLMs), including BERT in Devlin et al. [2018] and GPT in Brown et al. [2020], have brought tremendous improvements in machine translation and other NLP tasks, as discussed in Min et al. [2022], Kim et al. [2021], Costa-jussà et al. [2022]. Although GPT and NMT are based on the Transformer architecture, they differ in many ways. First, the GPT model only contains a decoder, using the same parameters for context and source input, while the NMT model uses an encoder-decoder architecture to translate sentences. GPT is trained on monolingual data, mainly in English, in contrast to NMT's reliance on curated paired bilingual data. Lastly, GPT requires a large number of parameters for in-context multilingual capability. According to Hendy et al. [2023], GPT models have achieved competitive translation accuracy in high-resource language. However, a significant gap exists in the translation quality for languages with limited resources, where they do not yet match the performance of commercial Neural Machine Translation (NMT) like Google Translate, as found in Zhu et al. [2023].

Moreover, a novel approach used prompts to guide LLMs in MT with zero-shot or few-shot methods. This approach has been shown to improve task-specific performance without the requirement for fine-tuning, as discussed in the reference Brown et al. [2020]. The research conducted by Zhang et al. [2023] demonstrated that this technique has enhanced the effectiveness of LLMs in MT tasks. Furthermore, it indicates that GPT models can outperform NMT systems in certain tasks when given carefully crafted prompts. The effectiveness of these prompts demonstrates the adaptability of the models and the level of authority they provide in generating output, as stated in Yao et al. [2024]. The success of prompts highlighted the models' flexibility and the degree of control they allow over the generated output, as discussed in Yao et al. [2024]. However, it is also important to note that LLM is sensitive to prompting and may lead to hallucination, as found in Ji et al. [2023].

In order to translate culture-specific text, it's essential for LLMs to possess cultural awareness. **Cultural awareness**, as explained in Huang and Yang [2023], is understanding the language on three levels: (1) being familiar with specific cultural norms; (2) identifying linguistic contexts that reflect these norms; and (3) adapting translations to include culture-specific interpretations. Research, including the studies by Huang and Yang [2023] and others like Yao et al. [2024], suggests that models with cultural awareness can significantly improve the precision and reliability of translations by accurately conveying the complex cultural nuances present in texts. However, the cultural sensitivity in LLMs remains unclear, as discussed in Yao et al. [2024]. Given the scarcity of datasets rich in CSI, further research into cultural awareness in both LLMs and NMT is crucial.

## 2.4 Explicitation in Cross-Cultural Translation

| Text | Implicit | Explanation |
|---|---|---|
| Sushi | Not Implicit | Globally well-know Japanese dish |
| Mooncake | Implicit | Chinese pastry consumed during Mid-Autumn Festival. |

Table 2.1: Examples of Implicit Tokens

In cross-cultural translation, explicitation plays a crucial role by classifying specific words or phrases as “implicit”—understood in one language but not immediately clear in another. This process involves introducing additional information to make these implicit tokens explicit, aiding their comprehension across different languages. This approach is highlighted in Han et al. [2023], where they detail how to bridge the gap between languages by providing clear explanations for these tokens.

According to the definition of implicit, CSI is a sub-category of implicit tokens in most cases. These are items or concepts recognized and only understood within one culture. The paper suggests that understanding CSI requires not just a direct translation but an explanation that encompasses cultural understanding. For instance, the term “Baozi” (transliteration of original Chinese pinyin) is CSI in Chinese and has no direct translation in English. However, since those original terms are widely used in English, English speakers do not need a cultural background to understand them. This recognition implies that **not all CSIs are inherently implicit**; their understanding can vary based on the reader’s familiarity with the cultural concept.

This paper will focus on CSI only since classifying implicit tokens requires considering cross-cultural understanding, which is beyond the scope of this project. However, in translating the CSI token, we will adopt the idea of explicitly translating it as it provides a clearer understanding of the target language and cultural context.

## 2.5 Related Work

In the study by Yao et al. [2024], who is also investigating how LLMs can translate CSI with prompt. In their work, they evaluated the CSI translation from multiple languages to Chinese, where we investigated translating Chinese cuisine names into English. Furthermore, their approach to translating CSI tokens broadly across various categories—from ideas and food to concepts—reveals an overview of the existing results of CSI that have been translated into different categories. We focused only on Chinese cuisine to investigate the CSI more deeply. We created three categories to investigate how each CSI presents different translation problems and how to identify CSI with automatic metrics. Moreover, the culinary domain requires a deeper understanding of its cooking methods, appearance, and flavours to translate culinary-related CSIs precisely. Also, we addressed some limitations in their study. Firstly, they primarily relied on data sourced from Wikipedia. Although Wikipedia offers a vast and varied pool of information, its content must often capture the nuanced and sophisticated quality characteristic of real-world translations. Some of Wikipedia’s material may be generated through MT, compromising its reliability, especially in CSI translations. In professional contexts, translations must achieve accuracy, resonate culturally and explain explicitly. In our BiMenu dataset, we collected a large amount of parallel Chinese cuisine that was well-localised in English. This data comes from high-quality restaurants that curated their translation to attract more customers.

Another related work is in recipe adaptation, one of the latest literature that adapted recipes from Chinese into English, by Cao et al. [2024], where they focus on how to accurately adapt the Chinese recipe into an English recipe, using monolingual Chinese recipe and English datasets., their work are also interesting as they are many different cultural norms in recipe need to adapt for, the measurement, the way of description, the equivalent ingredients. In our work, we also used recipes, but we used them as external knowledge to give LLMs external knowledge to explain CSI and product more explicit translation.

In the research from Lim [2018]), they primarily investigated the use of the term "braised" in Chinese dish names and their English translations. To perform the analysis, they collected menus from 27 Michelin-starred restaurants, using web crawling to gather a total of 3606 menu items. Our study takes a broader approach, collecting Chinese restaurant menus across the United Kingdom with three stars and above. We compiled a more extensive dataset of 4562 menu items from 314 restaurants. This wider selection allows us to examine a more diverse range of Chinese culinary styles and their representation in English translations.

Lastly, in the study Amenador and Wang [2022b], authors focused on the methods for translating culturally specific items (CSI) in the names of Chinese dishes within the field of linguistics. They examined the various translation methods and their frequency of use and suggested reasons for choosing strategies such as transference alone, transference with explicitation, transference with explanation, and using expressions from the target language. Regarding the dataset, they utilized a dataset published by Chinese authorities containing 2,816 entries, published in Zhao [2010] and combined it with data from Hubei Province, which included 4,618 dishes, published in University. Due to significant overlaps and many drinks, the number of unique dishes analyzed was reduced to 4,000. One major limitation identified was that the dataset, which was released in 2010, uses outdated terminology. For instance, it relies on transliteration for several terms, such as ’vanilla’, even though more accurate Chinese equivalents have become available. This points to a broader problem where the dataset fails to reflect the evolution of language, undermining its relevance for current linguistic studies and its effectiveness in translating CSI. Additionally, the translations in the dataset might not be easily understood by English speakers, often seeming odd. This issue likely stems from the translations done directly by native Chinese speakers without being localised in English.

# Chapter 3

# Data Collection and Composition

In this chapter, the rationale behind collecting the BiMenu dataset is elucidated, followed by a detailed explanation of the collection methodology. Furthermore, various methods employed during the dataset assembly are examined, highlighting those that proved to be most effective. Finally, a comprehensive overview of the dataset is provided, offering insights into its composition and potential applications.

## 3.1 The Dataset

### 3.1.1 Limited Publicly Available Parallel Chinese-English Culinary Datasets

The availability of parallel Chinese-English culinary datasets is markedly limited. In Lim [2018], they made an effort to compile a dataset encompassing 3,606 dish names from Michelin-starred restaurants. However, the scope of this dataset is confined, largely due to the exclusive nature of Michelin restaurants which tend to offer a limited selection of dishes. Moreover, these establishments frequently present dishes under innovative, albeit non-traditional, names that do not reflect the authentic nomenclature of Chinese cuisine. This limitation significantly hinders comprehensive research in CSI. Additionally, the dataset compiled by Lim [2018] is not publicly accessible, further constraining research efforts.

In a separate study, Yao et al. [2024] created a parallel dataset from Wikipedia, a total of 794 CSI in English to Chinese, covering 18 categories relevant to CSI across six language pairs, with Chinese cuisine representing only a fraction of the dataset. The limitation of the dataset on Wikipedia is discussed in Section 2.5. Notably, their dataset was not shared with the public until March 23, 2024, almost a year after their paper was published on May 23, 2023. By then, we had already compiled a more extensive dataset, totalling 4,562 entries.

Furthermore, the Chinese government has published a parallel Chinese-English culinary dataset in Zhao [2010], which, in its latest iteration released in 2010, contains 2,886 entries. A significant issue with this dataset is the inadequacy of its translations for

English-speaking audiences. Two limitations in this dataset are outdated and not localised in English, as discussed in Section 2.5.

## 3.2 Overview of BiMenu Dataset

Our dataset, named BiMenu (Bilingual Chinese-English Restaurant Menu Dataset), comprises more than 4,500 pairs of menu items, presented in both Chinese and English, derived from over 300 restaurant menus. The collection of these restaurant menus was facilitated by our automated crawler and an automatic menu parser, which were instrumental in extracting aligned text. This dataset underwent a verification process conducted by an annotator, who confirmed the accuracy of 4,562 dish names. Additionally, three annotators also identified and annotated a test set of 400 dish names, which are further discussed in the Methodology chapter.

## 3.3 The collection pipeline

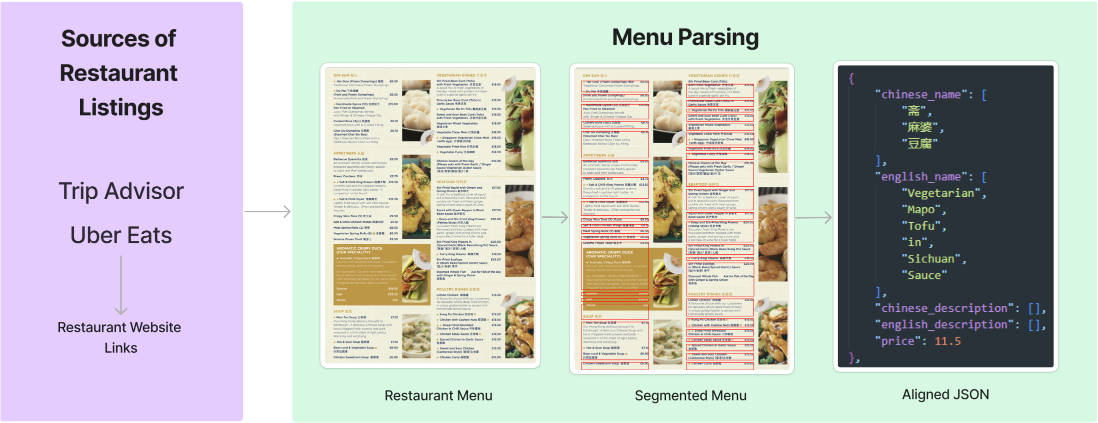
### 3.3.1 Sources of Restaurant Listings

Bilingual menus in English and Chinese within the restaurant industry can be classified into two principal categories, according to their origins and cultural contexts. The initial category is attributed to international dining establishments located within China, which predominantly offer cuisines of French, Italian, and American origin. Conversely, the second category is prevalent outside of China, mainly in establishments founded by Cantonese immigrants and Chineses. This study concentrates on the latter category, with a specific focus on the evolution and refinement of menu translations tailored to English-speaking customers. The central objective is to explore the translation of Chinese cuisine into English, requiring a translation direction where Chinese serves as the source language and English as the target language.

To facilitate this investigation, we utilized two widely recognized and publicly available platforms, TripAdvisor and Uber Eats. These platforms are among the largest websites that hosted restaurant information and were chosen due to their extensive database and the reliability of their user-generated ratings. Such ratings are considered indicative of the quality of the restaurants and, by extension, the quality of their menus. A key criterion for selection within this study was the price range of the dining establishments. Only restaurants classified above the mid-range category, out of the three categories - inexpensive, mid-range, and fine dining - were considered. This selection criterion was imposed to ensure that the quality of the restaurant menu translations collected met a high standard.

### 3.3.2 Automated Menu Collection

As there are no publicly available datasets for restaurant menus, we developed an automated crawler to extract menus directly from restaurant websites, as identified through TripAdvisor and Uber Eats. Menus are primarily in PDF format, though some



```
{
    "chinese_name": [
        "斋",
        "麻婆",
        "豆腐"
    ],
    "english_name": [
        "Vegetarian",
        "Mapo",
        "Tofu",
        "in",
        "Sichuan",
        "Sauce"
    ],
    "chinese_description": [],
    "english_description": [],
    "price": 11.5
},
```

Figure 3.1: Enter Caption

are images. The collection process involved navigating each website to detect and download menu content, overcoming challenges like the absence of direct menu download links. Collecting menus presented challenges due to the absence of straightforward links or buttons for direct downloads on many restaurant websites. To address this, we developed a method that involves recursively navigating through each website to classify any content resembling a menu. Once detected, our system automatically downloads the menu, whether it is in PDF or image format.

### 3.3.3 Menu Segmentation

Segmenting the menus was the most complex phase. Restaurant menus come in a wide variety of designs, formats, and structures. They can range from simple text lists to complex layouts with multiple columns, decorative fonts, images, and varying sizes of text. Some menus might use unique visual elements to differentiate sections, while others might rely on subtle spacing or font size changes. The variability in menu design significantly complicates the task of menu segmentation. To overcome this problem, we first converted PDFs to individual page images. Recognizing the diverse structures of restaurant menus, which vary significantly. We experimented with several strategies and verified them by manually checking a small selection of menus. Ultimately we chose the final strategy **"Segmentation Based on Price Tag"** which worked almost perfectly, but we described our other attempts and their shortcomings.

#### 3.3.3.1 Pure Text Alignment

We first experimented with text extracted from Optical Character Recognition (OCR) and performed an alignment of Chinese and English text. However, we found several limitations in this approach, and we **did not use this method in the end**.

The initial approach was to align the English and Chinese extracted sentences from the menu through OCR into the aligned parallel dataset. We encountered an array of complexities not initially anticipated. This alignment task is inherently challenging for several reasons, particularly due to the nature of language translation and the OCR's limitations in capturing text flawlessly. One significant challenge is the assumption required for alignment: the existence of a direct one-to-one correspondence between English and Chinese sentences. In reality, however, the alignment may often be one-to-many or many-to-one, where a single sentence in Chinese could correspond to multiple sentences in English (dish name, description and ingredients) or vice versa. This introduces complexity to the alignment process, deviating from the straightforward one-to-one mapping.

Moreover, the precision of OCR is not absolute; it occasionally fails to capture all text precisely, leading to gaps in texts that make matching English and Chinese phrases challenging due to the absence of corresponding counterparts. Moreover, this approach struggles to utilize structural layout information, which can offer significant clues for text alignment. Menus often use visual cues like spacing and typography to convey information about dish categories or importance, but these cues are lost in the OCR process, which focuses only on the text. This loss of structural information further complicated the alignment process.

#### 3.3.3.2 Neural Network-Based Document Layout Parsing

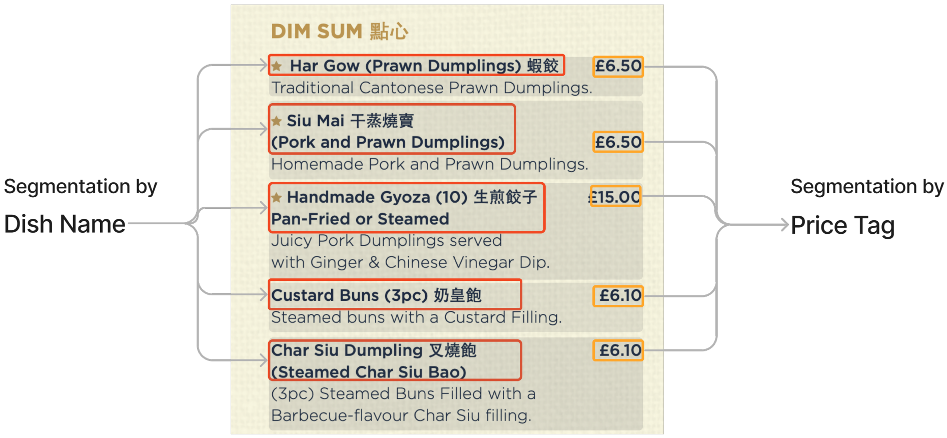
We experimented using a deep learning model to predict document layouts. However, we discovered that the model's generalization capabilities were insufficient for the diverse and unique layouts of restaurant menus. These menus frequently include a range of design components, such as watermarks and dish photographs, that make menu segmentation more complex. Moreover, the arbitrary placement of text and sudden alterations in layout, or even the total lack of uniform formatting, provide considerable obstacles for automatic parsing. **We did not use this approach in the end.**

The Document Layout Parsing model, by Shen et al. [2021], utilizes deep learning techniques for document image analysis to predict the layout of documents. Although we fine-tuned the mask R-CNN model with 50 self-labelled menus, its ability to generalize across different layouts was limited. This model showed proficiency with layouts it was originally trained on, such as academic papers, which adhere to certain stylistic conventions with only minor variations in text placement. However, restaurant menus introduced complexities that the model was ill-equipped to handle, particularly due to the inclusion of various design elements and the lack of distinct spacing between each dish item. This often requires human inference to understand the structure based on the meaning of words. Moreover, the efficacy of this model is constrained by the requirement for extensive training data and the difficulties in processing menus that do not have distinct segmentation boundaries.

#### 3.3.3.3 Segmentation Based on Potential Dish Names

After experimenting with the previously mentioned methods, we opted for a hybrid approach that integrates computer vision with text classification techniques. This method showed more promise in accuracy but still encountered several challenges; thus, **we did not use this approach in the end**.

We first recognized each text-bounding box through OCR and examined its content to classify whether it was a dish name. This step is used to determine where each



DIM SUM 點心

★ Har Gow (Prawn Dumplings) 蝦餃 £6.50
Traditional Cantonese Prawn Dumplings.

★ Siu Mai 干蒸燒賣
(Pork and Prawn Dumplings) £6.50
Homemade Pork and Prawn Dumplings.

★ Handmade Gyoza (10) 生煎餃子 £15.00
Pan-Fried or Steamed
Juicy Pork Dumplings served
with Ginger & Chinese Vinegar Dip.

Custard Buns (3pc) 奶皇飽 £6.10
Steamed buns with a Custard Filling.

Char Siu Dumpling 叉燒飽 £6.10
(Steamed Char Siu Bao)
(3pc) Steamed Buns Filled with a
Barbecue-flavour Char Siu filling.

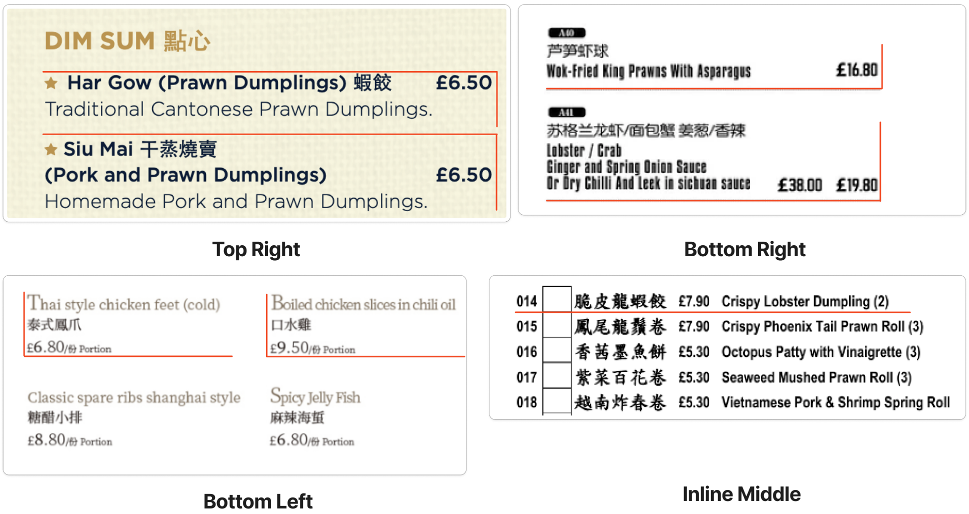
Figure 3.2: Segmentation by Dish Name and Price Tag Methods

dish starts and ends, enabling us to draw clear boundaries for each dish. Initially, we utilized a Support Vector Machine (SVM) to classify whether a given piece of Chinese phrase represents a dish name. Classifying dish names in English proved to be more challenging due to the lack of distinctive features within most English dish names and their descriptions. Moreover, since OCR may frequently produce fragments of text in one line, resulting in a complete dish name into fragments of phrases, they have been classified as non-dish names. The SVM model demonstrated accurate performance on both the training and test datasets. However, its capacity to generalize to new, unseen Chinese characters was somewhat limited. Despite these limitations, this method marked an improvement in segmentation accuracy. Nonetheless, it encountered difficulties in accurately distinguishing between dish names and their descriptions. Ultimately, we did not adopt this method due to limited generalization ability.

#### 3.3.3.4 Segmentation Based on Price Tag

After testing various methods without achieving the ideal results, we discovered an innovative approach that utilizes price tags for segmentation, which **worked almost perfectly**. Since each dish name is accompanied by a price tag, segmenting the menu content to the next price tag allowed for effective segmentation of almost all restaurant menus. This technique proved highly efficient, eliminating the need for training data and achieving near-perfect accuracy, except in cases where price tags were incorrectly placed on the menus. Moreover, this method works for any lingual type of restaurant menu in varied layouts and formats.

Despite the clear advantage of using price tags as segmentation markers, some adjustments were necessary to accommodate different menu layouts. For instance, some menus align text to the left, while others align to the right. This means that the price tag can only suggest where the segment of the dish happens, but it can not distinguish where the main dish content is relative to the price tag. If we simply assume that price



DIM SUM 點心

★ Har Gow (Prawn Dumplings) 蝦餃 £6.50
Traditional Cantonese Prawn Dumplings.

★ Siu Mai 干蒸燒賣
(Pork and Prawn Dumplings) £6.50
Homemade Pork and Prawn Dumplings.

Top Right

A40
芦笋虾球
Wok-Fried King Prawns With Asparagus £16.80

A41
苏格兰龙虾/面包蟹 姜葱/香辣
Lobster / Crab
Ginger and Spring Onion Sauce
Or Dry Chilli And Leek in sichuan sauce £38.00 £19.80

Bottom Right

Thai style chicken feet (cold)
泰式鳳爪
£6.80/份 Portion

Boiled chicken slices in chili oil
口水雞
£9.50/份 Portion

Classic spare ribs shanghai style
糖醋小排
£8.80/份 Portion

Spicy Jelly Fish
麻辣海蜇
£6.80/份 Portion

Bottom Left

| | | | | |
|---|---|---|---|---|
| 014 | □ | 脆皮龍蝦餃 | £7.90 | Crispy Lobster Dumpling (2) |
| 015 | □ | 鳳尾龍鬚卷 | £7.90 | Crispy Phoenix Tail Prawn Roll (3) |
| 016 | □ | 香茜墨魚餅 | £5.30 | Octopus Patty with Vinaigrette (3) |
| 017 | □ | 紫菜百花卷 | £5.30 | Seaweed Mushed Prawn Roll (3) |
| 018 | □ | 越南炸春卷 | £5.30 | Vietnamese Pork & Shrimp Spring Roll |

Inline Middle

Figure 3.3: Price Tag Relative Position in Dish Content

tags are always on the right, some restaurants have price tags on the bottom left and even in the middle, as shown in Figure 3.3, leading to an entirely failed recognition.

To overcome this variation, we considered the semantic meaning of the word we found; we performed four types of price tag relative position alignment, calculated each in-bounding box relevance score of its Chinese and English name, and returned the top matching relative position alignment.

## 3.4 Pre-processing the Segmented Dish Text

To prepare the text for analysis and translation, we applied standard pre-processing techniques, including tokenization and removing irrelevant information like Value Added Tax (VAT) details and allergen warnings. To protect privacy, the dataset has been anonymized by removing names and any identifiable details linked to the restaurants. This step ensured the dataset was clean and ready for further use.

## 3.5 Details of Dataset

The BiMenu dataset comprises several key components characterized by their statistical properties and annotations. It features 4,562 dish names that have been verified by an annotator. In addition, the dataset contains structural information extracted from menus, including a parallel dataset of each aligned text bounding box vertices and corresponding menu photo.

Furthermore, the dataset includes 15,370 unverified menu item pairs that are in both Chinese and English, sourced from over 312 Chinese restaurants in United Kingdom.

The average length of dish item names is 4.82 tokens, with the average length of English tokens slightly longer at 5.02 characters. Out of the total dishes, 266 are accompanied by English descriptions, 14 include Chinese descriptions, and 88 have indicators signifying whether the dish is vegetarian. It's important to note that price information is available for all dishes within the dataset.

The structural information in the dataset refers to bounding boxes for each segment of aligned text in Chinese and English. These bounding boxes correspond to specific sections within the menu photographs, providing a direct visual link to the textual data. Each bounding box's vertices are matched with an equivalent menu photo to ensure accurate representation and alignment.

# Chapter 4

# Methodology

This chapter outlines the approach to improving the translation of culturally specific items (CSIs) in dish names to produce culturally accurate translations. The process begins with the identification of CSIs, which represent the elements within dish names that pose significant challenges for translation due to their cultural uniqueness. By pinpointing these CSIs, we can apply targeted external knowledge to enhance the LLMs understanding of the specific cultural context, thereby improving translation accuracy.

## 4.1 CSI Definition

Since culture is an abstract concept, it is hard to directly capture fine-level cultural characteristics from texts. With this consideration in mind, we referred to an existing CSI classification framework, proposed in Newmark [1988], which has been popularly used in the study of human translations of cultural concepts. CSI is defined as terms without direct meaning in another language or culture and divided into five categories: *1) ecology; 2) material culture; 3) social culture; 4) organizations, customs, and ideas; 5) gestures and habits.* As CSI is a broad term that categorizes everything we can see, hear, feel and understand, while in the culinary domain, food-related items are often culture-specific items and present unique challenges in translation, as suggested by Marco [2019].

There are a lot of classifications for translating CSI, but no uniform and standard guidelines for how to classify CSI. To address this, we suggest a standardized set of **guidelines for human annotators to identify CSI**. This aims to minimize subjective interpretation as much as possible, recognizing that individuals may have varying perceptions of culture.

The methodology for identifying CSI in terms of their relevance to culture is detailed through four key metrics. A token qualifies as a CSI if it **meets at least two out of these four metrics**: Back-Translation Fidelity (BTF), Cultural Uniqueness (CU), Historical Story Presence (HSP), and Cultural Context Relevance (CCR).

- **Back-Translation Fidelity (BTF):** This evaluates the differences in meaning between the original text and its back-translation; we follow the idea that if a term can not be accurately back-translated, we will assume it is a CSI given in quote by Newmark [1988], CSI are terms or concept without direct meaning in another language.

- **Cultural Uniqueness (CU):** measures how unique a term is within its cultural setting, especially in terms of rarity or exclusivity when describing cultural practices, ingredients, or culinary items.
- **Historical Significance (HS):** This criterion determines whether a term possesses a historical background, categorizing its presence as true or false.
- **Cultural Context Relevance (CCR):** measures how closely a term is linked to the food name. Specifically, it examines whether removing the term significantly alters the meaning of the dish name.

Identifying CSI can be challenging because many terms that are well-recognized in English also refer to CSI. For instance, "dumplings" are a common term in English, yet they can also be culture-specific. In our identification, a term is considered a CSI only if it does not have a well-known direct translation in another language. Take "Wonton" as an example: its direct translation from Chinese could be "Fried hand" or "Copying hands," which categorizes it as a CSI. Conversely, "dumplings" translate directly to "dumplings" in English and, therefore, would not be classified as CSI.

## 4.2 Categories of CSI

In this section, we dived deeper into the intricacies of CSI, as they are interesting problems, and some present harder translation tasks compared with the rest. In this section, we proposed three novel categories to rank and classify each type of CSI, and each posed unique challenges in its field. These categories are designed to capture the varying levels of explicitness and figurativeness in the linguistic construction of dish names. The following definitions and examples aim to clarify the criteria for each category.

#### 4.2.0.1 Category 1: CSI with Common Words

**Definition:** This category includes dish names that contain a mix of CSI and common words. CSI words are those that suggest certain qualities, origins, or culinary techniques without stating them directly, requiring some level of cultural or culinary knowledge to interpret. Common words are straightforward and widely understood, requiring no specialized knowledge. While the main ingredients or the general idea of dish remain clear even without CSI.

**Example:** A dish named *Bistro Chicken*. "Bistro" implies a certain casual, French-inspired dining experience or preparation style, while "chicken" is a common word that directly refers to the main ingredient.

#### 4.2.0.2 Category 2: Figurative Combined Meaning

**Definition:** In this category, each word within the dish name is explicit and directly understandable (not CSI). However, when combined, these words convey a figurative or metaphorical meaning that transcends their literal interpretation. This category highlights the creative use of language to evoke certain imagery, emotions, or associations.

**Example:** A dish named *Sunset Salmon*. Individually, the words describe a time of day and a type of fish. Collectively, they suggest a dish with a visual appeal or flavour profile that mirrors the beauty and colours of a sunset.

#### 4.2.0.3 Category 3: Abstract Combined Meaning

**Definition:** This category encompasses dish names that exist beyond the realm of direct interpretation, understandable only with a certain depth of knowledge or cultural context. These names are crafted from the rich tapestry of metaphor, allegory, and symbolism, severing ties with the literal descriptions of ingredients, methods, or presentation. Instead, they lean into the art of storytelling through language, aiming to encapsulate a broader narrative, evoke specific emotions, or allude to deeper themes of mood, inspiration, or heritage.

**Example:** A dish named *Dragon's Breath Chili*. The name does not describe the ingredients or preparation method but instead evokes a powerful image of heat and intensity, suggesting a very spicy dish.

## 4.3 CSI Identification Pipeline

In this section, we explored methodologies for identifying CSI terms in dish names in automatic pipeline. Our approach starts by segmenting the dish name into smaller pieces, known as tokens. Next, we apply our novel automatic CSI identification pipeline. We also compare its effectiveness against using large language models (LLMs) like GPT-3.5 and GPT-4, tailored with prompts for the best results.

### 4.3.1 CSI Tokens Segmentation

We first segmented the Chinese dish name using a Xiachufang recipe, by Liu et al. [2022], as an additional dictionary to generate better token segmentation in jieba, in fxsjy [2024]. This allowed us to precisely segment phrases into tokens before further analysis. This step was crucial to circumvent the variable results produced when GPT attempts direct segmentation and identification, potentially leading to inaccuracies. This pre-processing ensures a stable foundation for comparing the effectiveness of GPT identification and CSI automatic identification pipeline in capturing the subtle complexities of Chinese culinary terms.

### 4.3.2 CSI Automatic Identification

In this approach, we followed the three guidelines for human annotation, as discussed in Section 4.1 and introduced an automated pipeline for each: Back-Translation, Cultural Uniqueness, and Historical Significance. However, measuring Cultural Context Relevance (CCR) presents challenges, and we acknowledge this as a limitation of our approach.

We also combined the three automatic identification pipelines above into one called **Combined CSI Identification**, which requires at least two metrics satisfying to be considered as CSI.

#### 4.3.2.1 Back Translation

The concept of CSI suggests that "certain terms or expressions do not have direct equivalents in other languages or cultures." To identify these CSIs, we employed the back translation method using Google Translate. This method helps evaluate how effectively the original meaning is maintained.

This approach involves two main steps:

1. **Initial Translation:** Translate the Chinese dish name to English.
2. **Back Translation:** Translate the English version back to Chinese in a new instance to avoid original term cache.

**Identification of CSI:** By comparing the original Chinese dish name with the back-translated version, discrepancies are noted. Terms or segments that were not accurately translated back to Chinese are classified as CSI. These discrepancies indicate a potential lack of direct equivalence in English, highlighting cultural or linguistic nuances.

**Limitations:** Back-translation presents two significant limitations. Firstly, it fails to recognize instances where dish names are overly abstract, classifying the entire name as a CSI. This results in a back-translation that is identical to the original text. For instance, the term "wonton" directly translates to "copying hands." Subsequent back-translation perfectly matches to its original Chinese form, erroneously indicating an absence of CSI. This issue predominantly occurs in categories 2 and 3 of CSIs, which involve figuratively combined meanings and abstract combined meanings, respectively. Secondly, the method's reliance on token match evaluation leads to the identification of minor character differences as CSIs, even when these differences do not alter the overall meaning. This approach inadvertently increases the incidence of false positives.

#### 4.3.2.2 Cultural Uniqueness (CU)

In this approach, we focused on identifying the Cultural Uniqueness (CU) within Chinese dish names from our BiMenu datasets. Our process began by breaking down the dish names into individual tokens. We then measured how often each token appeared and calculated its inverse frequency by dividing one by the frequency. We determined a cut-off point at the 95th percentile of these inverse frequencies. This percentile was chosen based on a manual review of a select group of 100 tokens. Tokens with an inverse frequency above this cut-off were marked as culturally unique (CU). For any token not previously seen, we assigned an inverse frequency of 1, indicating it was considered CU. We chose not to apply any smoothing techniques since we used inverse frequency for scoring against a fixed threshold, rather than for probability calculations.

**Limitations:** Our dataset was relatively small, containing only 4,562 entries with an average of 4.82 Chinese tokens per dish name. This small dataset size could lead to inaccuracies in frequency analysis. Additionally, our method only looked at the frequency of tokens individually. In CSI categories 2 and 3, many terms are not considered CSI on their own but are CSI when combined. Our current approach does not account for this aspect.

#### 4.3.2.3 Historical Significance (HS)

In this approach, we aimed to search for the history behind each token in the dish's name or the background of the entire dish using Wikipedia. We utilize the Wikipedia API to look up each term or the whole dish name. If a term appears with the "History" section of its Wikipedia page, we consider it to have historical significance. However, we excluded generic terms such as "chicken" or "sauce" in this method. Although these terms possess historical relevance, they lack the historical significance in relation to Chinese cuisine that is necessary to be classified as CSI.

**Limitations:** Wikipedia's coverage of Chinese cuisine and its historical context is limited, affecting the breadth of terms we can investigate.

### 4.3.3 CSI Identification with LLMs

As LLMs are trained with massive multilingual data and able to infer culturally aware context questions as shown in Yao et al. [2024], we used fine-tuned prompt in zero-shot to let GPT 3.5 and GPT4 to classify the token whether is CSI or not, prompt is illustrated in Figure 4.1.

We experimented with two different prompt formats:

- **With Dish Name Context:** This prompt includes the name of a dish associated with each token, providing specific cultural context. The purpose is to see how including a direct reference to a culturally significant dish affects the LLM's ability to classify tokens accurately.
- **Without Dish Name Context:** This prompt focuses solely on the tokens without any specific cultural context provided by dish names. It tests the LLM's ability to classify tokens based on their inherent cultural significance or lack thereof.

Both prompts ask the LLMs to consider a list of tokens extracted from various texts, identifying which tokens are culturally specific to Chinese culture—either because they are unique to this culture, hold particular significance, or lack a direct English equivalent. For each token, the LLMs are to provide a structured response that includes the token itself, a classification as "True" or "Fasle", and an explanation for the classification. We asked GPT to produce an explanation as we found out that it would lead to more consistent output and correct output that actually takes culture into account instead of simply letting GPT return true or false for CSI.

Figures illustrating the prompts for both approaches are provided to show how the LLMs are directed to analyze the tokens. The first Figure 4.1 demonstrates the prompt with dish name context, and the second Figure 4.2 shows the prompt without the dish name context.

**Prompt:** You are given a list of tokens related to Chinese culture. Your task is to analyze each token to determine if it represents a culture-specific item or concept.

For each token, provide a structured response that includes:
**Token:** Specify the token.
**Classification:** Classify in “True” or “False”
**Explanation:** Provide your reasoning.
**Examples:**

1. **Token:** [Token 1] in [Dish Name]
2. **Token:** [Token 2] in [Dish Name]
3. **Token:** [Dish Name] (entire phrase)

Figure 4.1: CSI Identification with Dish Name Context Prompt

**Prompt:** Analyze the given list of tokens, which may include items or concepts that are specific to Chinese culture, hold particular significance or lack direct English equivalents.

For each token, provide a structured response that includes:
**Token:** The token under examination.
**Classification:** Classify in “True” or “False”
**Explanation:** Offer your explanation for the classification.
**Examples:**

1. **Token:** [Token 1]
2. **Token:** [Token 2]

Figure 4.2: CSI Identification without Dish Name Context Prompt

## 4.4 Cross-Cultural Translation with external knowledge

Upon identifying CSIs, we explore methods to enhance translation accuracy by incorporating external knowledge. This approach aims to convey the meanings of CSIs more precisely and explicitly in the target language.

We propose utilizing detailed recipes to overcome the challenges of translating CSI-rich dish names. We used the most relevant recipe as external knowledge for each CSI (Recipe-based Translation), and we experimented this approach with two types of prompts in zero-shot:

- **CSI Preparation:** Provide recipe as an external knowledge to LLMs in one prompt.

- **CSI Explain-then-Translate:** Provide recipe as external knowledge to LLMs to first explain and extract relevant information; then, in the next prompt, we will ask LLMs to explain.

### 4.4.1 Recipe-based Translation:

Chinese dish names are mostly named from their cooking methods and ingredients, as discussed in Huo et al. [2020], Amenador and Wang [2022a]. Inspired by the naming conventions in Chinese, we propose an interesting method to feed the recipe as external knowledge for LLMs to decipher the meaning of CSI in the dish name. We decided to utilize the Xiachufang recipe database, by Liu et al. [2022], as a source of external knowledge. This database is the largest of its kind, with approximately 1.5 million Chinese recipes available.

Our translation process is structured into three key stages. Initially, we searched the recipe database using CSIs identified in the dish names. We then narrowed our search to the recipes that match these terms most closely, focusing on those whose names resemble the target dish name. Additionally, we employ the BM25 algorithm, developed in Robertson et al. [1994], as a method to rank documents based on the relevance of search terms. This algorithm is crucial for our process because it effectively filters out irrelevant content, such as advertisements and non-essential commentary found in many Xiachufang recipes, by prioritizing recipes where the CSIs are most relevant within the cooking instructions. This ensures we focus on the most relevant and significant recipe, enhancing the accuracy and quality of our translations.

Using recipes as external knowledge is particularly useful for translating CSI. Often, each part of a dish's name has a direct translation in the target language and culture. However, when combined, these parts may convey a meaning that significantly differs from the sum of their translations. A practical solution is to examine the dish's preparation method, ingredients, and instructions, as these elements provide vital clues about the dish's true meaning. For instance, a dish with a direct translation of "Racing Crab" might actually be better translated as "Crab-flavored Fish." This metaphorical name suggests that the taste of the fish resembles that of crab, indicating a cultural nuance that only becomes apparent with an understanding of how the dish is prepared. Therefore, using detailed recipes as a reference can enrich translation models with the necessary background knowledge, allowing for a more insightful interpretation of the dish names.

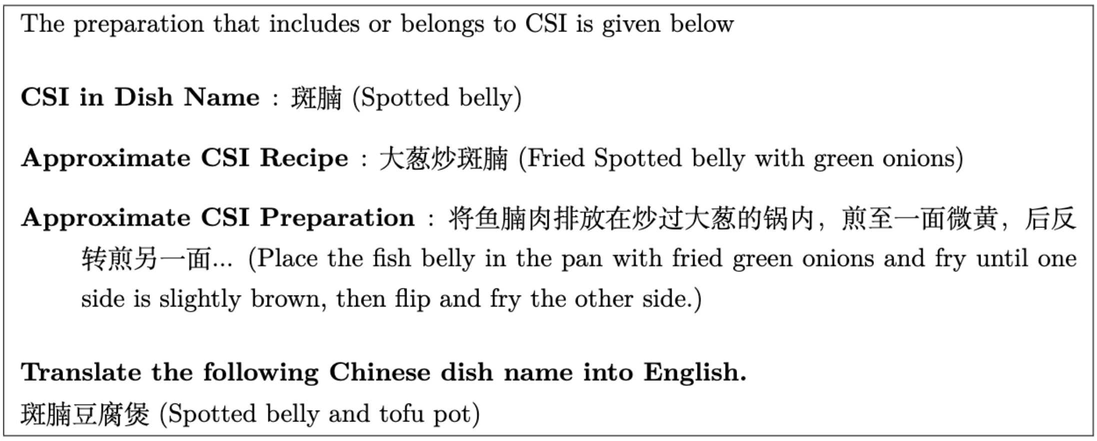
### 4.4.2 Prompt Strategy: CSI Preparation

In this prompt strategy, CSI preparation is centred on presenting the closest matching recipe that corresponds to the CSI identified. The strategy involves specifying the name and preparation details of the recipe while noting that the recipe provided might not exactly match the name of the dish. We categorize the CSIs encountered into two primary types:

1. **CSI with Common Words (Category 1):** This includes instances where one or several terms in the dish name are recognized as CSIs, yet the dish name as a whole does not solely consist of CSIs. An illustrative example of this scenario is

depicted in Figure 4.3, showcasing the recipe for “Spotted belly and tofu pot”, where “spotted belly” is the CSI.

2. **Figuratively and Abstract Combined Meaning (Categories 2 and 3):** This category pertains to cases where the entire dish name is likely viewed as a CSI, suggesting a more figurative or abstract meaning as a whole. An illustration of this is provided in Figure 4.4, demonstrating the recipe for “Ants Climbing the Tree”, where the entire dish name is CSI.



The preparation that includes or belongs to CSI is given below

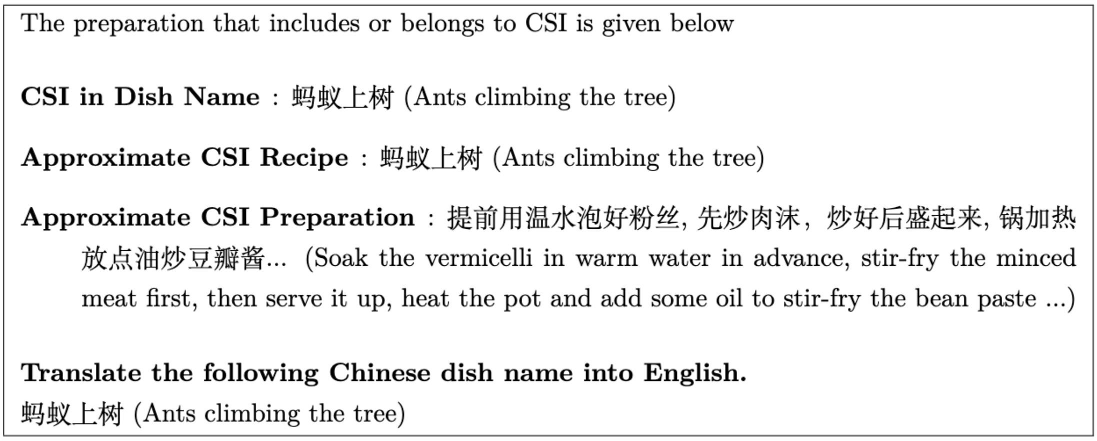
**CSI in Dish Name**：斑腩 (Spotted belly)

**Approximate CSI Recipe**：大葱炒斑腩 (Fried Spotted belly with green onions)

**Approximate CSI Preparation**：将鱼腩肉排放在炒过大葱的锅内，煎至一面微黄，后反转煎另一面... (Place the fish belly in the pan with fried green onions and fry until one side is slightly brown, then flip and fry the other side.)

**Translate the following Chinese dish name into English.**
斑腩豆腐煲 (Spotted belly and tofu pot)

Figure 4.3: CSI Preparation Prompt (Spotted belly and tofu pot)



The preparation that includes or belongs to CSI is given below

**CSI in Dish Name**：蚂蚁上树 (Ants climbing the tree)

**Approximate CSI Recipe**：蚂蚁上树 (Ants climbing the tree)

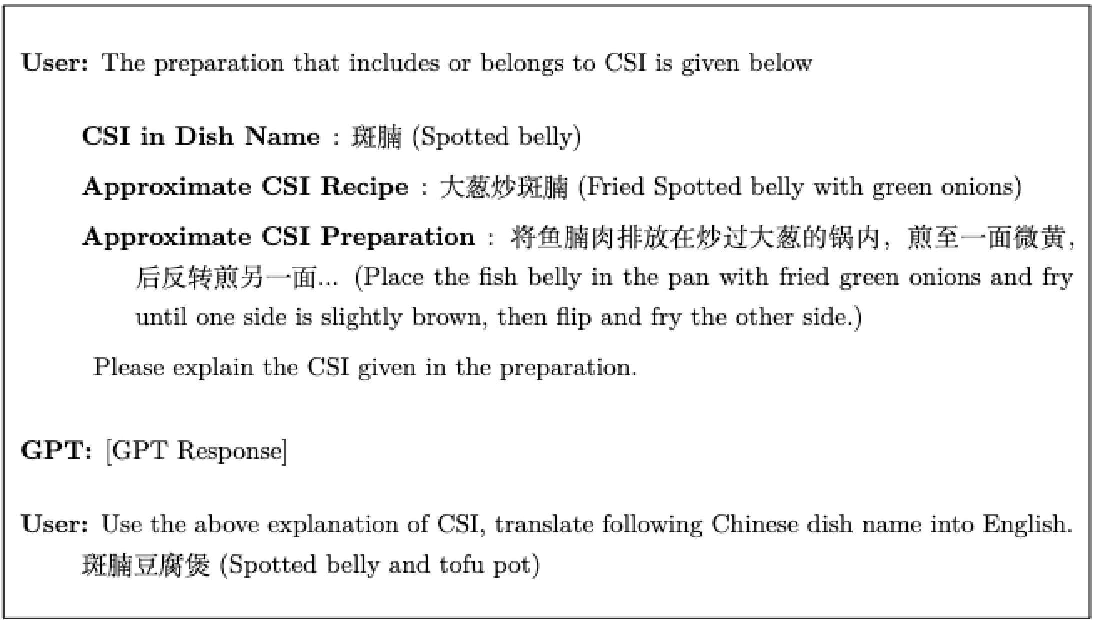
**Approximate CSI Preparation**：提前用温水泡好粉丝, 先炒肉沫，炒好后盛起来, 锅加热放点油炒豆瓣酱... (Soak the vermicelli in warm water in advance, stir-fry the minced meat first, then serve it up, heat the pot and add some oil to stir-fry the bean paste ...)

**Translate the following Chinese dish name into English.**
蚂蚁上树 (Ants climbing the tree)

Figure 4.4: CSI Preparation Prompt (Ants Climbing the Tree)

### 4.4.3 Prompt Strategy: CSI Explain-then-Translate

Inspired by the Chain of Thought (CoT) prompting strategy, as introduced by Wei et al. [2023] and Self-Explanation in Yao et al. [2024], we have developed a new approach in the prompt. This approach initially asks GPT to clarify the meaning of “CSI” as described in a recipe. Then, GPT uses the deduced meaning to translate the term more explicitly. This technique is used for complex reasoning tasks, such as interpreting dish names with CSIs not explicitly defined in the recipe. For example, a recipe might instruct to “cut it first, then stir fry” or note that “it can be very spicy” without explaining



**User:** The preparation that includes or belongs to CSI is given below

**CSI in Dish Name** : 斑腩 (Spotted belly)

**Approximate CSI Recipe** : 大葱炒斑腩 (Fried Spotted belly with green onions)

**Approximate CSI Preparation** : 将鱼腩肉排放在炒过大葱的锅内，煎至一面微黄，后反转煎另一面... (Place the fish belly in the pan with fried green onions and fry until one side is slightly brown, then flip and fry the other side.)

Please explain the CSI given in the preparation.

**GPT:** [GPT Response]

**User:** Use the above explanation of CSI, translate following Chinese dish name into English. 斑腩豆腐煲 (Spotted belly and tofu pot)

Figure 4.5: CSI Explain-then-Translate Prompt

the CSI. GPT's task is to infer the meaning of the CSI based on the recipe's instructions. The prompt is shown in Figure 4.5.

## 4.5 Experimental Setup

### 4.5.1 Entire BiMenu Dataset

The entire dataset is introduced in Section 3.2, which includes 4562 dish names, where we took only the Chinese and English names of each dish and removed the price, description, allergies and bounding box vertices of each text.

### 4.5.2 400 CSI Test Dataset Creation

Our BiMenu dataset were not labelled regarding the presence or absence of CSIs in each dish names. Fully labelling the dataset would have been time-consuming, a limitation we recognize. To manage this, we narrowed our focus to 400 selected dish names, split into two equal groups: 200 dishes identified with CSIs and 200 without. We aimed for a random selection process, picking dish names from the dataset one by one, labelling them as either containing CSIs or not until we had 200 in each category; the classification followed the guidelines in Section 4.1.

To ensure the accuracy of CSI identification, which can be subjective, we employed three native-Chinese speakers who are also fluent in English and have been exposed to Chinese culture and English Culture. They applied the identification guideline in Section 4.1 to label the datasets accordingly. We asked three annotators to classify Chinese dish name tokens as "With CSI" or "Without CSI". To quantitatively measure the consistency of their classifications and mitigate subjectivity, we used Fleiss' Kappa.

Fleiss' Kappa, proposed by Fleiss [1971], is a statistical method assessing inter-rater reliability, with values ranging from -1 (total disagreement) to 1 (perfect agreement), indicating consistency beyond chance. As a result, the Fleiss $k$ agreement score is 0.6385, indicating "Substantial agreement" in the 0.61 - 0.8 agreement score range. In the case of a disagreement among the annotators, we removed any CSI that did not get majority approval from three annotators. This helped us make sure the CSI in our test dataset was accurate and consistent.

### 4.5.3 150 CSI Category Test Dataset Creation

To better understand the impact of the three specific CSI categories outlined in Section 4.2, we tasked three annotators with categorizing 200 dishes from the "400 CSI Test Dataset" into three categories: CSI with Common Words, Figuratively Combined Meaning, and Abstract Combined Meaning.

We found that for each category, there were at least 50 dish names on which all annotators agreed. These dishes were then added to the CSI Category Test Dataset, resulting in a collection of 150 dish names related to CSI.

### 4.5.4 Models

We evaluated three machine translation models: Google Translate, GPT-3.5, and GPT-4. For GPT-3.5, we selected the "gpt-3.5-turbo-0125" version, while for GPT-4, we used the "gpt-4-0125-preview" version. These were chosen because they are the latest version offered by the OpenAI API, and they deliver their outputs in JSON format, making data extraction straightforward. For NMT, we employed Google Translate. It is recognized as the leading commercial solution, accessible via Google Translation API in Goole Cloud.

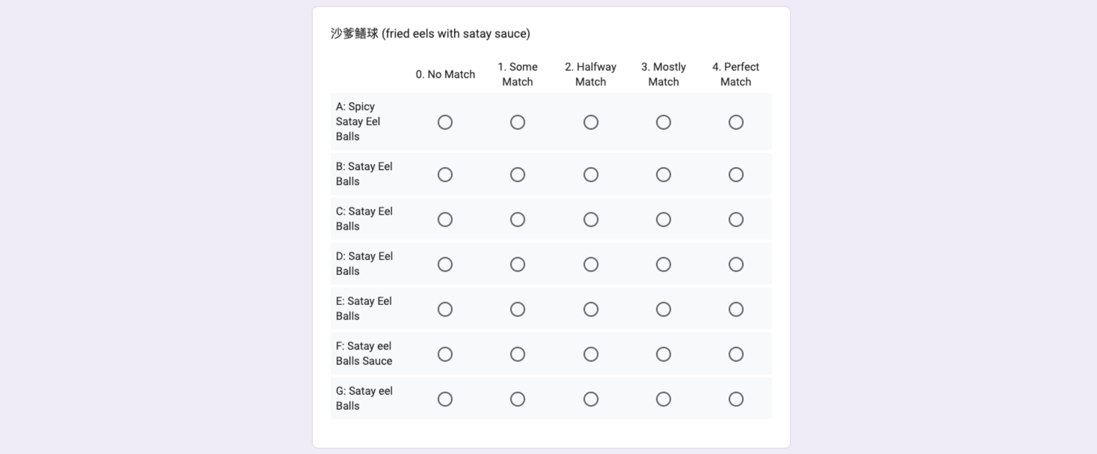
### 4.5.5 Automated Evaluation Metrics

The evaluation of translation quality is conducted through various recognized evaluation frameworks, including the **BLEU** proposed in Papineni et al. [2002], **COMET** (specifically "wmt-22-comet-da") proposed in Rei et al. [2022], **ChrF** proposed in Popović [2015], and **ROUGE-L** proposed in Lin [2004]. These frameworks provide a systematic approach for measuring the accuracy and quality of translations, enabling the comparison of different translation approaches on a consistent basis. BLEU and COMET are utilized to gauge the preciseness of translations and their congruence with standard translations. ChrF, which examines character-level accuracy and recall, and ROUGE-L, which focuses on the longest shared subsequence, offer further insight into the translation's fluency and its effectiveness in retaining the original meaning.

Furthermore, the study utilized Precision, Recall, and F1 Score for CSI identification evaluation:

- **Precision** quantifies the accuracy of class predictions, defined as:

  $Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$



沙爹鳝球 (fried eels with satay sauce)

| | 0. No Match | 1. Some Match | 2. Halfway Match | 3. Mostly Match | 4. Perfect Match |
|---|---|---|---|---|---|
| A: Spicy Satay Eel Balls | ○ | ○ | ○ | ○ | ○ |
| B: Satay Eel Balls | ○ | ○ | ○ | ○ | ○ |
| C: Satay Eel Balls | ○ | ○ | ○ | ○ | ○ |
| D: Satay Eel Balls | ○ | ○ | ○ | ○ | ○ |
| E: Satay Eel Balls | ○ | ○ | ○ | ○ | ○ |
| F: Satay eel Balls Sauce | ○ | ○ | ○ | ○ | ○ |
| G: Satay eel Balls | ○ | ○ | ○ | ○ | ○ |

Figure 4.6: Human Evaluation Questionnaire

- **Recall** measures the proportion of actual positives correctly identified, defined as:

  $Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$

- **F1 Score** represents the harmonic mean of Precision and Recall, defined as:

  $F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

### 4.5.6 Human Evaluation

Our study incorporates a vital human evaluation component to counterbalance the automated measures. We conducted a survey using Google Forms involving nine participants who are both native Chinese speakers and proficient in English. Their task was to evaluate a selection of 30 dish names and translations. These dish names were selected at random from a pool of 150 CSI Test Sets, with ten names from each category. The decision was made to ensure an equilibrium - wide-ranging enough to encompass diverse examples yet succinct enough to finish the survey within a 15-minute timeframe. This method allowed us to prevent participants from feeling overwhelmed and ensure that the responses we received were more accurate and reliable.

During the evaluation, participants were provided with the original Chinese dish names and their translations from restaurants as ground truth. We asked participants to evaluate the systems of Google Translate, GPT 3.5 (baseline, CSI Preparation, CSI Explain-then-Translate), and GPT 4 (baseline, CSI Preparation, CSI Explain-then-Translate). To prevent any bias, the translations were associated with system IDs rather than system names. The respondents were then asked to rate the accuracy of each translation on a scale from 0 (No Match) to 4 (Perfect Match). Figure 4.6 shows an example of the questionnaire.

To aid understanding, we provided examples of translated dish names before the participants began the rating process. Our goal was to make the evaluation as straightforward and unbiased as possible.

# Chapter 5

# Result and Analysis

This chapter presents a comprehensive analysis of the experimental findings focused on evaluating CSI translation difficulties, CSI identification methods, and translation quality across different technologies, as well as exploring the agreement between human evaluations and automated metrics.

## 5.1 Entire Dataset Translation Evaluation

We start by analysing how well current machine translation models perform when translating Chinese menu items into English. This establishes a baseline performance for the task. We compare three machine translation models, selecting the strongest commercial MT system (Google Translate) and the two most widely used LLMs, GPT3.5 and GPT4.

Performance was assessed on the entire BiMenu dataset in Section 4.5.1, totalling 4562 dish names, using COMET and BLEU scores alongside ChrF and ROUGE-L metrics.

| **Method** | **BLEU** | **ChrF** | **ROUGE-L** | **COMET** |
|---|---|---|---|---|
| Google Translate | 11.64 | 48.6 | 39.99 | 70.67 |
| GPT-3.5 | 10.82 | 46.36 | 38.01 | 70.57 |
| GPT-4 | **12.03** | **48.96** | **40.19** | **71.4** |

Table 5.1: Automated evaluation results using reference-based metrics: BLEU, ChrF, ROUGE-L and COMET. Higher scores indicate better performance on all metrics.

The results in Table 5.1 show that **GPT-4 outperforms both Google Translate and GPT-3.5** across all evaluated metrics. This suggests a nuanced capability of GPT-4 in handling the cultural specificities and complexities inherent in translating Chinese menu items. The fact that GPT-4 surpasses Google Translate is particularly intriguing. The latter has been highly optimized for machine translation tasks and, as existing literature points out, generally excels over LLMs across various language families, except for the low-resource language Romance. This was demonstrated in recent empirical analyses by Zhu et al. [2023], who highlighted Google Translate's dominance in multilingual translation tasks. Adding to this discourse, it's important to acknowledge the limitations of my dataset, which is highly specialized Chinese-English restaurant menu items with an average of 4.84 Chinese tokens and an average of 5.09 English tokens. While this dataset's specificity provides valuable insights into domain-specific translation capabilities of models like GPT-4, it may not fully encompass the broader linguistic and cultural variations in more generalized datasets used in studies such as Zhu et al. [2023]. Thus, while the findings from my dataset are promising, they highlight a niche aspect of machine translation that complements the broader understanding of model performances across diverse translation tasks.

The superiority of GPT-4 in this specific translation task suggests that the dataset poses unique challenges that NMT, even those as advanced as Google Translate, struggle with. These challenges likely stem from the need for a translation approach beyond the literal, requiring a deep understanding of cultural nuances and context. It's this complex, culturally rich nature of the dataset that demands a more sophisticated translation strategy—a demand that GPT-4 seems better equipped to meet.

## 5.2 Impact of CSI on Translation

In this section, we aimed to discuss the research question: **"How does the presence of CSI affect the performance and outcomes of translation in NMT systems and LLMs"**.

We evaluate this question using the 400 test set created in Section 4.5.2, which comprises 200 dish names with CSIs and 200 without CSIs.

Table 5.2: Evaluation Metrics for MT Systems With and Without CSI

| With CSI | | | | |
|---|---|---|---|---|
| System | BLEU | COMET | ChrF | ROUGE-L |
| Google Translate | 5.28 | 58.07 | 33.79 | 24.63 |
| GPT-3.5 | 6.21 | 60.11 | 32.68 | 24.50 |
| GPT-4 | **7.21** | **61.77** | **35.78** | **27.81** |

| Without CSI | | | | |
|---|---|---|---|---|
| System | BLEU | COMET | ChrF | ROUGE-L |
| Google Translate | **17.95** | **75.29** | 56.30 | **50.45** |
| GPT-3.5 | 15.03 | 74.74 | 51.76 | 44.92 |
| GPT-4 | 16.47 | 74.61 | **56.50** | 46.67 |

The evaluation highlights the distinct challenges CSIs pose. For dish names with CSIs, all systems exhibit a **notable decline in performance** across BLEU, COMET, ChrF, and ROUGE-L metrics compared to their performance on dish names without CSI. This decline underscores the difficulties in translating terms deeply rooted in specific cultural contexts. Among the evaluated systems, GPT-4 stands out for its adept handling of CSIs, achieving the highest scores across all metrics for dishes with CSIs. This points to GPT-4's enhanced ability to grasp and convey the nuanced cultural and contextual nuances that CSIs embody.

Conversely, in text without CSIs, Google Translate exhibits a remarkable proficiency, outperforming both GPT-3.5 and GPT-4 in nearly all metrics except slightly lower ChrF score compared with GPT 4. This superiority of Google Translate in text without CSI shows its strengths in translating straightforward, culturally neutral content. The results suggest that Google Translate's NMT may be optimized for translations where cultural depth and contextual nuances are less pronounced, benefiting from a comprehensive and diverse training dataset that prioritizes general language accuracy over deep cultural understanding.

In conclusion, the drastic difference in result of CSI presence, answer the research question that **CSIs significantly complex the translation tasks**. This finding reinforces the idea that the presence of CSIs is a key factor in making translations more challenging. It also highlights that GPT models perform better on texts with CSIs, while NMT models are more adept at handling texts without them.

### 5.2.1 Evaluation of Categorized CSI in Translation

In this section, our objective is to address the research question in CSI translation task, **"Which categories of CSI are most problematic for identification and translation, and what does this reveal about the intrinsic characteristics of these categories?"**

The evaluation was conducted utilizing a test set of 150 Categorized CSIs, as discussed in Section 4.5.3, wherein each category includes an equal number of 50 dish names. The assessment employed both COMET and BLEU scores in alongside with ChrF and ROUGE-L metrics for analysis.

**Analysis**

In Table 5.3, Category 1 (CSI with Common Words), GPT-4 outperforms both Google Translate and GPT-3.5 significantly in terms of BLEU scores, indicating a superior ability to maintain the meaning of the original text. The COMET scores are relatively consistent across the three systems, suggesting that all maintain a reasonable level of semantic accuracy. However, the higher ChrF and ROUGE-L scores for GPT-4 reflect its enhanced capability in capturing the nuances of the CSI and common word mix, ensuring a better-quality translation. The number of tokens across systems suggests that GPT-4 and GPT-3.5 generate slightly more accurate translations than Google Translate, potentially providing more context or clarity.

The translations of Category 2 (Figuratively Combined Meaning) demonstrate a distinct challenge, as they require the system to understand and convey figuratively combined meanings. The BLEU and COMET scores are lower across the board than Category 1, reflecting the inherent difficulty of translating figurative text. Despite this, GPT-4 shows a marginal improvement over the others, although the margins are narrower. This implies that although GPT-4 demonstrates superior proficiency in dealing with

Table 5.3: Comparative Analysis of Translation Challenges Across Culturally Specific Item Categories

CSI - Category 1 (CSI with Common Words)

| System | BLEU | COMET | ChrF | ROUGE-L | # Tokens |
|---|---|---|---|---|---|
| Google Translate | 7.40 | 65.87 | 38.14 | 32.16 | 24.47 |
| GPT-3.5 | 8.93 | 65.64 | 39.64 | 35.09 | 26.45 |
| GPT-4 | **16.52** | **65.88** | **46.55** | **39.76** | 26.49 |

CSI - Category 2 (Figuratively Combined Meaning)

| System | BLEU | COMET | ChrF | ROUGE-L | # Tokens |
|---|---|---|---|---|---|
| Google Translate | 6.31 | 59.65 | **34.86** | 26.48 | 23.21 |
| GPT-3.5 | 7.10 | 61.20 | 32.16 | 25.02 | 25.98 |
| GPT-4 | **8.13** | **61.74** | 34.57 | **26.89** | 24.65 |

CSI - Category 3 (Abstract Combined Meaning)

| System | BLEU | COMET | ChrF | ROUGE-L | # Tokens |
|---|---|---|---|---|---|
| Google Translate | 2.96 | 51.20 | 28.92 | 17.38 | 20.58 |
| GPT-3.5 | 4.82 | 53.33 | 29.88 | 20.27 | 24.81 |
| GPT-4 | **5.25** | **56.45** | **35.27** | **26.07** | 23.10 |

figurative language, the intricacy of translating metaphorical meanings is difficult for all systems. The lower ChrF and ROUGE-L scores across all systems compared to Category 1 highlight the struggle to maintain the figurative essence of the original dish names.

Category 3 (Abstract Combined Meaning) proves to be the most challenging for all systems, as indicated by the significantly lower BLEU, COMET, ChrF, and ROUGE-L scores. This category requires not only a deep understanding of cultural and culinary contexts but also the ability to interpret and translate metaphor, allegory, and symbolism. Despite these challenges, GPT-4 demonstrates a marginally better performance than its predecessors, particularly in COMET and ChrF scores, suggesting a slightly better grasp of the abstract meanings. The number of tokens indicates that GPT-3.5 tends to provide more verbose translations in this category, possibly in an attempt to capture the abstract meanings more fully.

In conclusion, the result reveals that as the **complexity of the CSI increases**, from Category 1 to 3, the **performance of all systems tends to decrease**. This trend underscores the difficulty of translating dishes with figurative and abstract combined meanings. This clear trend in translation accuracy emphasizes how well different categories are handled. Essentially, as we move from simpler to more abstract categories, language models like GPT-4 begin to outperform NMT like Google Translate by a wider margin. The improvement becomes more pronounced with the increase in abstraction, showcasing that LLMs are getting better at handling figurative and abstract translations. However, the challenge remains significant, particularly with abstract and figurative meanings that demand a deep cultural and linguistic understanding beyond the literal translation.

## 5.3 CSI Identification Evaluation

This section continues to investigate the research question on CSI identification: **“Which categories of CSI are most problematic for identification and translation, and what does this reveal about the intrinsic characteristics of these categories?”**

To this end, we evaluated eight CSI identification methods as proposed in Section 4.3. These methods include: CSI Combined, Back-translation, Cultural Uniqueness, Historical Significance, and GPT 3.5 and GPT 4 (both with and without context). Our evaluation involved analyzing these techniques against the 200 dish names with CSI in 400 Test Dataset labelled by human, discussed in Section 4.5.2. The analysis was conducted using Precision, Recall, and F1 score metrics.

### 5.3.1 Evaluation of Overall Results

Table 5.4 reveals that identifying CSI is a complex task, evidenced by all low F1 scores. Including context yields a marginal advantage for GPT in identifying CSI. All GPT models demonstrate a high recall rate, interestingly, with GPT 3.5 unexpectedly surpassing GPT 4. Moreover, the combined approach, which integrates back-translation, cultural uniqueness, and historical significance, delivers competitive results compared

Table 5.4: Comparison of CSI Identification Methods for Exact Match, where Combined denotes combined method of Back-translation, Cultural Uniqueness and Historical Significance, Context refers to giving entire dish name as context in prompt

| Method Type | Method | Precision (%) | Recall (%) | F1 Score (%) |
|---|---|---|---|---|
| GPT 3.5 | Without Context | 45.93 | **84.96** | 59.63 |
| | With Context | 47.40 | 80.53 | 59.67 |
| GPT 4 | Without Context | 48.46 | 76.55 | 59.35 |
| | With Context | 49.05 | 79.65 | 60.71 |
| Other | Combined | 79.70 | 46.90 | 59.05 |
| | Back-translations | 54.77 | 48.23 | 51.29 |
| | Cultural Uniqueness | 54.01 | 80.53 | **64.65** |
| | Historical Significance | **90.24** | 16.37 | 27.72 |

to individual GPT models. The ’Cultural Uniqueness’ method achieves the highest F1 score, near the top in precision and recall, implying a significant association between CSI and the distinctiveness of terms. In contrast, ’Historical Significance’ attains the peak in precision, affirming the strong link between CSI and historical context, though its low recall indicates that not every token carries a historical narrative, perhaps due to limitations in the dataset or inherent lack of historical relevance.

### 5.3.2 Evaluation of Categorized Three CSI

To analyse the effect of each CSI category, we further evaluated eight CSI identification techniques as proposed in Section 4.3. These methods include: CSI Combined, Back-translation, Cultural Uniqueness, Historical Significance, and GPT 3.5 and GPT 4 (both with and without context). Our evaluation involved analyzing these techniques against the 150 CSI Categorized Test Set in different categories (referenced in Section 4.5.3). The analysis was conducted using Precision, Recall, and F1 score metrics.

**Analysis**

The result in Table 5.5 highlights that knowing the **context**, like the dish name, is **crucial for accurately identifying CSI** as the complexity and category of CSI increases. For CSIs in Category 1 (CSI with Common Words), the dish name or context is not needed because these CSIs can be identified on their own, evidenced by lower F1 scores in GPT with context. However, as we move to more complex CSIs in Category 2, incorporating the dish name becomes beneficial, especially in GPT-4, enhancing the accuracy of CSI detection slightly. In Category 3, which involves the most complex and abstract CSIs, adding context greatly improves identification accuracy, with F1 scores jumping by 7% and 8%, highlighting the significant role of context in these challenging cases.

Moreover, our result shows that back-translation works well for Category 1, where CSIs are mentioned separately from dish names. This success fades as the complexity increases (Categories 2 and 3, involving more figurative and abstract meanings), where back-translation’s accuracy declines dramatically. This drop is due to its inability to

Table 5.5: Comparison of CSI Identification Methods Across Categories, where Combined denotes combined method of Back-translation, Cultural Uniqueness and Historical Significance, Context refers to giving entire dish name as context in prompt

| CSI - Category 3 (CSI with Common Words) | | | | |
|---|---|---|---|---|
| Method Type | Method | Precision (%) | Recall (%) | F1 Score (%) |
| GPT 3.5 | Without Context | 37.59 | **79.37** | 51.02 |
| | With Context | 37.10 | 73.02 | 49.20 |
| GPT 4 | Without Context | 42.11 | 76.19 | 54.24 |
| | With Context | 40.52 | 74.60 | 52.51 |
| Other | Combined | 83.33 | 47.62 | 60.61 |
| | Back-translations | 60.66 | 58.73 | 59.68 |
| | Cultural Uniqueness | 53.33 | 76.19 | **62.75** |
| | Historical Significance | **100.00** | 9.52 | 17.39 |

| CSI - Category 2 (Figurative Combined Meaning) | | | | |
|---|---|---|---|---|
| Method Type | Method | Precision (%) | Recall (%) | F1 Score (%) |
| GPT 3.5 | Without Context | 45.96 | **89.16** | 60.66 |
| | With Context | 46.81 | 79.52 | 58.93 |
| GPT 4 | Without Context | 45.99 | 75.90 | 57.27 |
| | With Context | 48.15 | 78.31 | 59.63 |
| Other | Combined | 82.69 | 51.81 | 63.70 |
| | Back-translations | 52.00 | 46.99 | 49.37 |
| | Cultural Uniqueness | 54.17 | 78.31 | **64.04** |
| | Historical Significance | **100.00** | 19.28 | 32.32 |

| CSI - Category 3 (Abstract Combined Meaning) | | | | |
|---|---|---|---|---|
| Method Type | Method | Precision (%) | Recall (%) | F1 Score (%) |
| GPT 3.5 | Without Context | 54.84 | 85.00 | 66.67 |
| | With Context | 58.82 | **87.50** | **70.35** |
| GPT 4 | Without Context | 58.49 | 77.50 | 66.67 |
| | With Context | 58.62 | 85.00 | 69.39 |
| Other | Combined | 75.00 | 41.25 | 53.23 |
| | Back-translations | 53.23 | 41.25 | 46.48 |
| | Cultural Uniqueness | 54.33 | 86.25 | 66.67 |
| | Historical Significance | **78.95** | 18.75 | 30.30 |

grasp the nuanced meanings of words in context. Simply put, while back-translation can handle direct translations, it falls short in interpreting the broader meaning of phrases, a critical aspect for accurately identifying CSIs in these more complex categories.

Finally, our combined CSI identification method outperforms GPT models in Categories 1 and 2, showing notable improvements (at least 9% in Category 1 and 3% in Category 2). However, it underperforms in Category 3, the most abstract category. Interestingly, considering cultural uniqueness (CU) yields highly competitive results across all categories, suggesting **CU is a strong indicator for CSI**. Even though combined metrics include CU, it requires agreement from at least two methods to confirm CSI; it did not fare as well, partly because CU often identifies CSIs that the other methods miss, evidenced by CU's significantly higher recall with back-translation and Historical Significance, lowering the overall score. This finding implies that CU is particularly effective but suggests a need for refinement in the combined metric approach to better capture CSI across various categories.

## 5.4 Evaluation of Recipe-based Translation

In this section, we aim to answer the research question, **"What are the most effective methods to translate CSIs into another language while being culturally accurate and understandable to the target culture?"**.

This section presents the assessment of our recipe-based translation methodology for accurately translating dish names with CSI. The evaluation was segmented into three categories of CSIs. Our method employed two prompt strategies discussed in Section 4.4, CSI Preparation and CSI Explain-then-Translate. We evaluated it against the 150 CSI Categorized Test Set (in Section 4.5.3).

### 5.4.1 Evaluation of Translation in Three Categories

The results from Table 5.6 indicate a significant decrease in performance for GPT models in Category 1 when using recipe-based translation, compared to the baseline, across all metrics. Additionally, we discovered that the decrease in accuracy can be attributed to over-reliance on provided recipes, resulting in inaccuracies when the translation task requires adaptability or flexibility to new food elements not included in the recipe reference. For instance, strictly following a recipe might lead to faulty translations, as seen in the situation where translating "Fried Fish Belly with Tofu" might mistakenly include onions based on a recipe for "Fried Fish Belly with Onions." This category's challenge lies in accurately translating dishes that combine common words with CSIs, where the recipe content may not properly align with the dish to be translated, thus affecting the translation accuracy.

Categories 2 and 3, however, depict a different picture. These categories focus on translating dish names with figurative or abstract meanings, where the precision in ingredient details is less critical than capturing the dish's cultural or figurative meaning. The result suggests that for Category 2, while improvements were observed, they were minimal and not statistically significant across different methodologies. However, for

Table 5.6: Evaluation Metrics for Recipe-based Translation

CSI - Category 1 (CSI with Common Words)

| System | Method | BLEU | COMET | ChrF | ROUGE-L | Tokens |
|---|---|---|---|---|---|---|
| Google Translate | Baseline | 5.77 | 65.24 | 38.33 | 29.96 | 24.53 |
| GPT 3.5 | Baseline | 8.93 | 65.99 | 38.91 | 31.31 | 28.15 |
| | CSI Preparation | 5.74 | 64.11 | 35.70 | 27.71 | 26.70 |
| | Explain-Translate | 8.37 | 64.79 | 36.02 | 30.31 | 30.81 |
| GPT 4 | Baseline | **10.99** | **66.14** | **42.84** | **33.09** | 25.87 |
| | CSI Preparation | 8.32 | 65.53 | 37.55 | 30.35 | 26.38 |
| | Explain-Translate | 7.32 | 65.47 | 36.79 | 30.06 | 26.72 |

CSI - Category 2 (Figurative Combined Meaning)

| System | Method | BLEU | COMET | ChrF | ROUGE-L | Tokens |
|---|---|---|---|---|---|---|
| Google Translate | Baseline | 6.36 | 58.08 | 32.76 | 25.09 | 21.33 |
| GPT 3.5 | Baseline | 8.11 | 60.63 | 34.25 | 27.97 | 23.50 |
| | CSI Preparation | 7.01 | 60.00 | 28.41 | 23.47 | 22.04 |
| | Explain-Translate | 8.45 | 62.45 | 32.35 | 27.85 | 22.81 |
| GPT 4 | Baseline | **10.09** | 60.31 | 36.87 | 29.11 | 23.08 |
| | CSI Preparation | 10.73 | **63.08** | **37.78** | **30.98** | 22.83 |
| | Explain-Translate | 7.87 | 61.70 | 32.35 | 25.57 | 23.54 |

CSI - Category 3 (Abstract Combined Meaning)

| System | Method | BLEU | COMET | ChrF | ROUGE-L | Tokens |
|---|---|---|---|---|---|---|
| Google Translate | Baseline | 3.41 | 49.62 | 27.88 | 18.18 | 22.38 |
| GPT 3.5 | Baseline | 4.12 | 50.42 | 27.20 | 20.21 | 25.54 |
| | CSI Preparation | 5.26 | 55.29 | 26.88 | 19.50 | 23.58 |
| | Explain-Translate | 5.32 | 57.64 | 25.14 | 21.45 | 25.98 |
| GPT 4 | Baseline | 5.37 | 55.17 | 32.64 | 26.15 | 24.75 |
| | CSI Preparation | 5.85 | 58.53 | 28.10 | 23.65 | 27.46 |
| | Explain-Translate | **6.64** | **58.54** | **29.63** | **26.72** | 27.33 |

Table 5.7: Comparative Evaluation of Translation Systems across Three CSI Categories Using Recipe-based Translation Methodology

Category 3, the abstract nature of dish names seems to favour the recipe-based translation approach, especially with the Explain-Translate method on GPT 3.5, demonstrating a significant enhancement in translation accuracy. This suggests that for abstract dish names, leveraging cultural and contextual insights, even with imperfect recipe matches, can considerably enhance translation effectiveness.

Lastly, the Explain-then-Translate (ETT) method did not consistently outperform the CSI Preparation approach across all scenarios. In our analysis, ETT allowed GPT-3.5 to surpass CSI Preparation, producing translations with a longer average length. This suggests that ETT prompts help GPT-3.5 to generate more detailed and localized English translations, evidenced by our manual check through 150 translations. However, for GPT-4, while there was an increase in the average translation length, this did not result in a performance improvement comparable to GPT-3.5's. This might implies that the more advanced reasoning capabilities of GPT-4 may make the explicit reasoning facilitated by ETT less effective.

### 5.4.2 Human Evaluation

Table 5.8: Comparative Analysis of Translation and Interpretation Performance (Scores out of 100, we adjusted user scores from a range of (0 to 4) to a scale of (0 to 100) by multiplying each score by 25 for easier comparison)

| System | Baseline | CSI Preparation | CSI Explain-then-Translate |
|---|---|---|---|
| Google Translate | 36.20 | N/A | N/A |
| GPT-3.5 | **41.52** | 49.57 | 50.57 |
| GPT-4 | 39.22 | **56.03** | **51.73** |

The result in Table 5.8 shows that using a recipe-based approach to significantly enhances translation quality. Specifically, GPT-3.5 and GPT-4's performance improved by about 7 and 8 points. This improvement aligns with findings from an automated evaluation in Section 2, which showed that the CSI Preparation result increased across all GPT models. Importantly, the "Explain-then-Translate" strategy further enhanced GPT-3.5's scores but resulted in a decrease for GPT-4 compared to the CSI Preparation approach.

# Chapter 6

# Conclusions

## 6.1 Summary

In this study, we collected the BiMenu dataset of bilingual Chinese-English restaurant menus, employing our document layout analysis parser that utilizes price tags for menu segmentation. This dataset was annotated by three individuals who labelled tokens with CSI, adhering to a set of guidelines we crafted to reduce subjective interpretation. The CSIs were subsequently divided into three novel categories we proposed, facilitating an analysis of their varying impacts on the processes of translation and identification. We further proposed an automatic CSI identification pipeline, incorporating a back-translation, cultural uniqueness, and historical significance, which has proven effective in identifying CSIs with LLM-based CSI identification in CSI categories 2 and 3.

We found that CSIs pose significant challenges to the translation process, as evidenced by the diminished performance across all systems when translating dish names containing CSIs. The evaluation across the three CSI categories revealed a trend: as the complexity of the CSIs increases, the performance of translation systems correspondingly declines. This trend emphasizes the critical need for developing and refining models capable of accurately interpreting and translating abstract, figurative, and complex culturally specific content.

To produce a more accurate translation of dish names, we proposed a recipe-based translation for LLMs and investigated external knowledge of recipe work with two prompt strategies. In automatic metrics evaluation, we found that external knowledge of recipes significantly improved in Categories 3 (Abstract Combined Meaning) and 2 (Figurative Combined Meaning), while Category 1 saw a decrease. In our human evaluation, we found that recipe-based translation shows significant improvement in both GPT models and correlates with automatic metrics.

We believe our datasets, proposed CSI categories, methods for parsing menus on the price tag, identifying CSI, and giving recipes as external knowledge will significantly contribute to the broader study of culture awareness and cross-cultural translation from Chinese to English.

## 6.2 Limitations

The primary limitation of this study lies in the utilization of a small test dataset for evaluating methodologies in CSI identification and translation. The dataset size of 400 needed to be increased for a comprehensive and reliable assessment of the various approaches. Furthermore, the diversity and number of human annotators constituted another significant limitation. Furthermore, the limited number and diversity of human annotators were another significant constraint. The complexity of the CSI annotation requires a significant subjective perspective, which requires more human annotators to minimise the subjective impact. However, the backgrounds of the three annotators needed more diversity to fully understand the intricate cultural aspects of Chinese cuisine in our study, which can significantly differ across different regions. The lack of diversity hindered our ability to capture the whole cultural context of each dish.

The omission of cultural context relevance metrics is a critical limitation in automatic CSI identification. We attempted semantic modelling but failed to accurately capture the significance of each token in the names of dishes, especially for abstract or metaphorical terms. This exclusion resulted in a critical aspect of human CSI annotation being overlooked, thus impacting the optimal performance of the automated system.

Additionally, the study encountered limitations during the data collection phase, particularly with our menu parser, which could only process menus containing price tags, thus excluding potentially valuable data from the analysis. Also, the complexity of the menu parser further requires manual validation of parsed menus against menu photos; this process was both time-consuming and laborious, limiting the size of our dataset of menu text collected from over 20,000 menus to 4,562 dishes

Lastly, our recipe-based translation did not include the aroma and visual information of each Chinese dish, as they are equally important as the cooking method. This oversight may result in significant understanding gaps, given the crucial role these sensory details play in identifying and appreciating dishes.

## 6.3 Future Work

The previous section highlighted several areas for improvement and provided directions for future research. Currently, our dataset is limited in size and scope, only focusing on translations between Chinese and English. By expanding our dataset to include more language pairs, we can gain more detailed insights into the translation of CSI in different cultural contexts. Moreover, increasing the number of annotators and ensuring their cultural backgrounds are diverse, especially from different regions of China, could reduce subjective bias in CSI annotations and more accurately reflect the elaboration of CSI meanings.

Additionally, we have developed a preliminary pipeline for automatic CSI recognition, although there is still room for improvement. Improvements include optimising the translation model to more effectively capture the subtle differences in CSIs, refining the method of estimating frequencies to make them more accurate and identifying more authoritative sources for terms with historical or cultural significance. Implementing these modifications will enhance the precision and reliability of the pipeline.

Furthermore, future research could explore different approaches to identifying CSIs, such as enhancing current models or integrating contextual learning to enhance the recognition of CSIs. Another intriguing research direction involves translating CSIs without relying on external knowledge, a method we proposed but did not investigate. This could be an approach to infer the meaning of CSIs based on the composition of the Chinese characters, as each character conveys its meaning. This method could be beneficial for translating dish names, as the characters can convey their meaning or ingredients, leading to a more instinctive understanding of CSI through character composition and visual clues.

# Bibliography

Idris Akinade, Jesujoba Alabi, David Adelani, Clement Odoje, and Dietrich Klakow. Varepsilon kú mask: Integrating Yorùbá cultural greetings into machine translation. In Sunipa Dev, Vinodkumar Prabhakaran, David Adelani, Dirk Hovy, and Luciana Benotti, editors, *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 1–7, Dubrovnik, Croatia, May 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.1. URL `https://aclanthology.org/2023.c3nlp-1.1`.

Idris Akinade, Jesujoba Alabi, David Adelani, Clement Odoje, and Dietrich Klakow. Varepsilon kú mask: Integrating Yorùbá cultural greetings into machine translation. In Sunipa Dev, Vinodkumar Prabhakaran, David Adelani, Dirk Hovy, and Luciana Benotti, editors, *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 1–7, Dubrovnik, Croatia, May 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.1. URL `https://aclanthology.org/2023.c3nlp-1.1`.

Román Álvarez and M Carmen Africa Vidal. *Translation, power, subversion*, volume 8. Multilingual Matters, 1996.

Kate Benedicta Amenador and Zhiwei Wang. The translation of culture-specific items (csis) in chinese-english food menu corpus: A study of strategies and factors. *SAGE Open*, 12(2):215824402210966, Apr 2022a. doi: https://doi.org/10.1177/21582440221096649.

Kate Benedicta Amenador and Zhiwei Wang. The translation of culture-specific items (csis) in chinese-english food menu corpus: A study of strategies and factors. *Sage Open*, 12(2):21582440221096649, 2022b. doi: 10.1177/21582440221096649. URL `https://doi.org/10.1177/21582440221096649`.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99, 2024.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Alok Das. Neural machine translation (nmt): Inherent inadequacy, misrepresentation, and cultural bias. *International Journal of Translation*, 32(1-2), 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. Training neural machine translation to apply terminology constraints. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1294. URL `https://aclanthology.org/P19-1294`.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. Training neural machine translation to apply terminology constraints. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1294. URL `https://aclanthology.org/P19-1294`.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

fxsjy, 2024. URL `https://github.com/fxsjy/jieba`.

James Paul Gee. *Introducing discourse analysis: From grammar to society*. Routledge, 2017.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. Dictionary-based phrase-level prompting of large language models for machine translation, 2023.

HyoJung Han, Jordan Boyd-Graber, and Marine Carpuat. Bridging background knowledge gaps in translation with automatic explicitation. Jan 2023. doi: https://doi.org/10.18653/v1/2023.emnlp-main.603. URL `https://aclanthology.org/2023.emnlp-main.603/`.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. How good are gpt models at machine translation? a comprehensive evaluation, 2023. URL `https://arxiv.org/abs/2302.09210`.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias

Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. Challenges and strategies in cross-cultural NLP. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.482. URL https://aclanthology.org/2022.acl-long.482.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime G Carbonell. Domain adaptation of neural machine translation by lexicon induction. *arXiv (Cornell University)*, Jan 2019. doi: https://doi.org/10.18653/v1/p19-1286. URL https://aclanthology.org/P19-1286/.

Junjie Hu, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig. Deep: Denoising entity pre-training for neural machine translation. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jan 2022. doi: https://doi.org/10.18653/v1/2022.acl-long.123. URL https://aclanthology.org/2022.acl-long.123/.

Jing Huang and Diyi Yang. *Culturally Aware Natural Language Inference*. 2023. URL https://aclanthology.org/2023.findings-emnlp.509.pdf.

Caiqiao Huo, Xiaomei Du, and Weichen Gu. The metaphor and translation of the dish names in chinese food culture. *Open Journal of Modern Linguistics*, 10(05): 423–428, Jan 2020. doi: https://doi.org/10.4236/ojml.2020.105025. URL https://www.scirp.org/journal/paperinformation?paperid=102807.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), mar 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation, 2020. URL https://arxiv.org/abs/2010.00710.

Young Jin Kim, Ammar Ahmad Awan, Alexandre Muzio, Andres Felipe Cruz Salinas, Liyang Lu, Amr Hendy, Samyam Rajbhandari, Yuxiong He, and Hany Hassan Awadalla. Scalable and efficient moe training for multitask multilingual models, 2021.

Claire Kramsch. Language and culture. *Aila Review*, 27:30–55, Dec 2014. doi: https://doi.org/10.1075/aila.27.02kra. URL https://www.jbe-platform.com/content/journals/10.1075/aila.27.02kra.

Daniel Liebling, Katherine Heller, Samantha Robertson, and Wesley Deng. Opportunities for human-centered evaluation of machine translation systems. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 229–240, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.17. URL `https://aclanthology.org/2022.findings-naacl.17`.

Lily Lim. A corpus-based study of braised dishes in chinese-english menus. In *Proceedings of the 32nd pacific asia conference on language, information and computation: 25th joint workshop on linguistics and language processing*, 2018.

Chin-Yew Lin. *ROUGE: A Package for Automatic Evaluation of Summaries*. Jul 2004. URL `https://aclanthology.org/W04-1013.pdf`.

Xiao Liu, Yansong Feng, Jizhi Tang, Chengang Hu, and Dongyan Zhao. Counterfactual recipe generation: Exploring compositional generalization in a realistic scenario. *arXiv preprint arXiv:2210.11431*, 2022.

Josep Marco. The translation of food-related culture-specific items in the valencian corpus of translated literature (covalt) corpus: a study of techniques and factors. *Perspectives*, 27(1):20–41, 2019. doi: 10.1080/0907676X.2018.1449228. URL `https://doi.org/10.1080/0907676X.2018.1449228`.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work?, 2022.

P. Newmark. *A Textbook of Translation*. English language teaching. Prentice-Hall International, 1988. ISBN 9780139125935. URL `https://books.google.co.uk/books?id=ABpmAAAAMAAJ`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. *BLEU: a Method for Automatic Evaluation of Machine Translation*. Jul 2002. URL `https://aclanthology.org/P02-1040.pdf`.

Ulrika Persson. Culture-specific items : Translation procedures for a text about australian and new zealand children's literature. 2015. URL `https://api.semanticscholar.org/CorpusID:146245406`.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL `https://aclanthology.org/W15-3049`.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Sev-*

*enth Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.wmt-1.52`.

Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. pages 0–, 01 1994.

Zejiang Shen, Ruochen Zhang, Melissa Dell, Lee Benjamin, Jacob Carlson, and Weining Li. Layoutparser: A unified toolkit for deep learning based document image analysis, 2021. URL `https://arxiv.org/abs/2103.15348`.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014a. MIT Press.

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

odjel za anglistiku Sveučilište u Zadru. [sic] – a journal of literature, culture and literary translation focuses on theoretical, empirical and artistic research in the fields of culture, literature and literary translation, 2024. URL `https://www.sic-journal.org/Article/Index/173`.

Zhengzhou University. Henan province chinese menu in english.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi Quoc, V Le, and Denny Zhou. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models Chain-of-Thought Prompting*. Jan 2023. URL `https://arxiv.org/pdf/2201.11903.pdf`.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. Benchmarking llm-based machine translation on cultural awareness, 2024.

Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: A case study, 2023. URL `https://arxiv.org/abs/2301.07069`.

H Zhao. Enjoy culinary delights: A chinese menu in english, 2008.

H Zhao. Xuhuiqu chinese menu in english, 2010.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis, 2023.

# Appendix A

# Participants’ information sheet

# Participant Information Sheet

| | |
|---|---|
| Project title: | Improving Menu Localization with Culture-Aware Language Models |
| Principal investigator: | Alexandra Birch Mayne |
| Researcher collecting data: | Zhonghe Zhang |
| Funder (if applicable): | |

This study was certified according to the Informatics Research Ethics Process, reference number 2024/125449. Please take time to read the following information carefully. You should keep this page for your records.

### Who are the researchers?

Researchers include Zhonghe Zhang, the principal investigator, and his project supervisor Professor. Alexandra Birch Mayne.

### What is the purpose of the study?

The objective of this study is to investigate the effectiveness of Large Language Models (LLMs) in enhancing menu translations by utilizing their capacity to comprehend and provide nuanced, cross-cultural translations from Chinese to English.

### Why have I been asked to take part?

You have been selected to participate in this study due to your unique perspective as a native Chinese speaker. Your role will involve evaluating translations generated by LLMs, focusing on their effectiveness in conveying the essence of culturally specific menu items accurately.

### Do I have to take part?

No – participation in this study is entirely up to you. You can withdraw from the study at any time, without giving a reason. Your rights will not be affected. If you wish to withdraw, contact the PI. We will stop using your data in any publications or presentations submitted after you have withdrawn consent. However, we will keep copies of your original consent, and of your withdrawal request.

**What will happen if I decide to take part?**

- Access to the Study:
  - Upon agreeing to participate, you will receive a link to an online platform specifically designed for this research. This platform will host the survey you'll be asked to complete.
- Survey Structure:
  - The survey will present you with a series of questions. For each question, you will be shown an original text in Chinese alongside multiple English translations of that text.
  - Your task will be to evaluate and rank these translations based on how accurately and effectively they capture the essence and meaning of the original Chinese text.
- Data Collection Details:
  - Types of Data Collected: The survey will focus on your perceptions and rankings of the translation accuracy of menu items from Chinese to English. It may also collect your general feedback on the translation process and any suggestions for improvement.
- Method of Collection:
  - The data will be collected through a structured online questionnaire, designed to be user-friendly and straightforward.
- Duration:
  - The entire process is expected to take between 10 to 15 minutes of your time, and you can resume and continue as you want.
- Audio/Video Recording:
  - Please note that there will be no audio or video recording during this survey. Your responses will be text-based only.
- Scheduling:
  - The survey can be completed at your convenience once the link is provided. It will be accessible online, allowing you to participate from anywhere at any time within the survey availability period.

### Are there any risks associated with taking part?

There are no significant risks associated with participation.

### Are there any benefits associated with taking part?

No benefits associated with, but your feedback will provide a large improvement in evaluating LLMs cross-cultural translation ability, leading better technology advancement.

### What will happen to the results of this study?

The results of this study may be summarised in published articles, reports and presentations. Quotes or key findings will be anonymized: We will remove any information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data may be archived for a minimum of 2 years.

### Data protection and confidentiality.

Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number rather than by name. Your data will only be viewed by the researcher/research team Zhonghe Zhang and Alexandra Birch Mayne.

All electronic data will be stored on a password-protected encrypted computer, on the School of Informatics' secure file servers, or on the University's secure encrypted cloud storage services (DataShare, ownCloud, or Sharepoint) and all paper records will be stored in a locked filing cabinet in the PI's office. Your consent information will be kept separately from your responses in order to minimise risk.

### What are my data protection rights?

The University of Edinburgh is a Data Controller for the information you provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

For general information about how we use your data, go to: edin.ac/privacy-research

### Who can I contact?

If you have any further questions about the study, please contact the lead researcher, Zhonghe Zhang, s2029523@ed.ac.uk.
If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

### Updated information.

If the research project changes in any way, an updated Participant Information Sheet will be made available on http://web.inf.ed.ac.uk/infweb/research/study-updates.

### Consent

By proceeding with the study, I agree to all of the following statements:

- I have read and understood the above information.
- I understand that my participation is voluntary, and I can withdraw at any time.
- I consent to my anonymised data being used in academic publications and presentations.
- I allow my data to be used in future ethically approved research.

# Appendix B

# Participants' consent form

Participant number:________________________

# Participant Consent Form

| | |
|---|---|
| Project title: | Improving Menu Localization with Culture-Aware Language Models |
| Principal investigator (PI): | Alexandra Birch Maybe |
| Researcher: | Zhonghe Zhang |
| PI contact details: | a.birch@ed.ac.uk |

By participating in the study you agree that:

- I have read and understood the Participant Information Sheet for the above study, that I have had the opportunity to ask questions, and that any questions I had were answered to my satisfaction.

- My participation is voluntary, and that I can withdraw at any time without giving a reason. Withdrawing will not affect any of my rights.

- I consent to my anonymised data being used in academic publications and presentations.

- I understand that my anonymised data will be stored for the duration outlined in the Participant Information Sheet.

**Please tick yes or no for each of these statements.**

**1.** I allow my data to be used in future ethically approved research.

□ **Yes** □ **No**

**2.** I agree to take part in this study.

□ **Yes** □ **No**

Name of person giving consent    Date dd/mm/yy    Signature

_______________________    _____________    _______________________

Name of person taking consent    Date dd/mm/yy    Signature

_______________________    _____________    _______________________

THE UNIVERSITY *of* EDINBURGH
informatics

If you had human participants, include information about how consent was gathered in an appendix, and point to it from the ethics declaration. This information is often a copy of a consent form.