

Article

Qiao Gan* and Jingyuan Ye

Contextualizing the variation in causal clause ordering in Mandarin Chinese: a multifactorial analysis

<https://doi.org/10.1515/cllt-2025-0024>

Received March 3, 2025; accepted July 25, 2025; published online August 11, 2025

Abstract: The sequential ordering of causal clauses in Mandarin Chinese displays variation, with reason connectives (e.g., *yīnwèi*, *yóuyú*) appearing either before or after the main clause (sentence-initial vs. sentence-medial positions). However, the interplay of syntactic, semantic, and discourse-pragmatic factors shaping causal clause ordering remains underexplored. This study addresses the gap through a corpus-based multifactorial analysis of spoken Mandarin, employing mixed-effects regression modeling. Our findings reveal that syntactic parsing factors, including clause length and complexity, influence clause positioning. Longer subordinate clauses favor sentence-medial ordering of reason connectives, while the interaction between main clause length and its complexity shows that simpler but longer main clauses favor sentence-medial placement. Semantically, emotional valence plays a role: positive subordinate clauses prefer sentence-medial positioning, while the emotional valence of main clauses interacts with genre. Specifically, neutral valence in dialogues and positive valence in monologues favor sentence-medial positioning, whereas negative valence in dialogues least favors it. Discourse-pragmatic factors further condition causal clause ordering. The presence of *yīnwèi* and the absence of *suǒyǐ* jointly favor sentence-medial positioning of reason connectives. This study provides novel insights into the multifaceted factors influencing causal clause ordering in Mandarin Chinese, thereby enhancing our understanding of sentence structure and discourse cohesion in the language.

Keywords: causal clause; contextual conditioning; linguistic variation; multifactorial analysis; spoken Chinese

***Corresponding author: Qiao Gan**, New Zealand Institute of Language, Brain and Behaviour, University of Canterbury, Private Bag 4800, 8140, Christchurch, New Zealand, E-mail: qiao.gan@canterbury.ac.nz
Jingyuan Ye, La Trobe University, Melbourne, VIC, Australia

1 Introduction

Language variation is ubiquitous and systematic in language use, conditioned by both linguistic factors (e.g., animacy, length of constituents) and social factors (e.g., age, gender). Since the 1960s, language variation has been widely studied across the English-speaking world (e.g., Labov 1963), forming the basis of the variationist paradigm, which has been later extended to research on variation in Spanish (Torres Cacoullos 1999), Portuguese (Bortoni-Ricardo 1985) and French (Sankoff and Cedergren 1971). These languages share typological and genealogical similarities, as they all belong to western branches of Indo-European, especially Germanic and Romance (Adli and Guy 2022). Non-Indo-European languages are investigated only infrequently, such as Chinese, whose grammatical system heavily relies on word order, which displays typological differences with Germanic and Romance languages in many linguistic aspects (Li and Thompson 1981; Wang 1984). As Adli and Guy (2022) pointed out, the generalizability of the variationist paradigm and its theoretical underpinnings should be tested with languages beyond Germanic and Romance (see also Stanford 2016). Another gap in the variationist paradigm is the long-standing focus on phonetics since its inception (cf. Gries 2001; Poplack and Tagliamonte 1999); comparatively fewer studies have examined variation beyond sound structures in languages other than English (although see recent variationist studies of Chinese morphosyntax by Li et al. 2023; Liao et al. 2024; Xu et al. 2024). To further address these gaps, our focus is on one morphosyntactic variable in spoken Mandarin Chinese, causal clause ordering, whose variation and probabilistic constraints have not been systematically examined to date.

Among the frequent types of adverbial clauses in Mandarin Chinese, including temporal, conditional, concessive, and causal clauses, the connectives for the first three types tend to precede the clause material they modify (Wang 1999). In contrast, causal clauses differ notably from other adverbial types in exhibiting considerable variability in the position of their connectives, which can appear either at the beginning or within the sentence (Biq 1995; Tsai 1996; Xing 2001). This positional variability is not unique to Chinese: a typological study of 60 languages, such as Chinese, English, German and Japanese, found that “causal adverbial clauses are structurally more independent of the associated main clause than other semantic types of adverbial clauses” (Diessel and Hetterle 2011: 24), resulting in greater cross-linguistic flexibility in the placement of reason connectives. In Mandarin specifically, causal clauses introduced by reason connectives (e.g., *yīnwèi* ‘because’, *jìrán* ‘since’, *yóuyú* ‘due to’, *jiànyú* ‘in view of’) may occur in either sentence-initial or sentence-medial positions (Chao 1968; Li and Thompson 1981: 635), as illustrated in Examples (1) and (2), respectively.

- (1) 因为 多 个 国家 都 出现 了 消极 比赛,
 yīnwèi duō gè guójiā dōu chūxiàn le xiāojí bǐsài,
 because many CL country all emerge PFV negative match,
 舆论 也 关注 规则 争议。
 yúlùn yě guānzhù guīzé zhēngyì.
 public also pay.attention.to rule dispute
 ‘Because multiple countries have experienced negative competition, public
 opinion is also focusing on the controversy over the rules.’
 (S1B_042)¹
- (2) 您 的 突破 口 在 精神 建设,
 nín de tūpò kǒu zài jīngshén jiànshè,
 you POSS breakthrough point at mental development,
 因为 物质 达到 了。
 yīnwèi wùzhì dá dào le.
 because material achieve PFV.
 ‘Your breakthrough lies in spiritual development, because material
 needs have been met.’

Studies from both generative grammar and functional grammar perspectives have not identified truth-conditional differences in the meanings of different causal clause orderings in Mandarin Chinese (e.g., Gasde and Paul 1996; Pan and Zhu 2023; Thompson et al. 2007). Instead, some studies have found that sentence-initial and sentence-medial causal clauses are pragmatically different, making causal clause ordering an interesting linguistic variable to test the semantic-pragmatic contestability (Song and Tao 2009). Furthermore, some linguistic factors have been found to constrain the ordering of causal clauses, such as the type of connective and the presence of the result marker *suǒyǐ* (‘so’) (Biq 1995; Kirkpatrick 1993; Li et al. 2016). However, existing studies only conducted monofactorial analysis, focusing on one or two linguistic factors at once and ignoring their interactions. Thus, no study has examined the effects of these co-occurring contextual constraints and their interactions on causal clause ordering in Mandarin Chinese.

To bridge these gaps, we adopt a variationist approach. First, variationist studies focus on how language is used by speakers in real-world contexts. Our dataset, consisting of spoken Chinese utterances in social settings, aligns well with the empirical requirements of this paradigm. Second, we aim to understand the variable

¹ This indicates the text’s ID. Here, ‘S1B’ represents public dialogues, while ‘042’ identifies the specific file. In the following examples, ‘S1A’ refers to private dialogues, ‘S2A’ to unscripted monologues, and ‘S2B’ to scripted monologues.

processes involved in causal clause ordering. A central task of the variationist approach is to define the variable envelope of a linguistic feature, identifying “alternative ways of saying the same thing”, within which meaningful comparisons can be made (Labov 1994: 550). In our study, variation in the ordering of causal clauses represents different ways of expressing causal relationships. Finally, this study employs quantitative reasoning to analyze the probabilistic structure of causal clause ordering. The variationist framework is particularly suited for this purpose because it considers the co-occurrence and distribution of multiple probabilistic constraints to probe the “orderly heterogeneity” in language use (Weinreich et al. 1968: 100).

2 Causal clause ordering in Mandarin Chinese

In expressing causal relations in Chinese, the users have the choices of using or not using connectives (Biq 1995: 49–50). When they do choose to express causal relations explicitly using connectives, there are at least five connectives that are frequently used, including *yīnwèi* (‘because’), *yīn* (‘cause’), *jìrán* (‘since’), *yóuyú* (‘due to’), and *jiànyú* (‘in view of’). Previous studies have argued that these reason connectives are substantially similar in their semantics, but some semantic nuances exist among them. Specifically, *yīnwèi*, *yīn*, *yóuyú*, and *jiànyú* are more similar to each other in terms of semantics compared to *jìrán*; however, this difference between the two categories is not truth-conditional, as all of them can indicate causal relations (Li et al. 2016; Xing 2003: 199). Thus, the locus of the variation in expressing causal relations through these connectives lies in the ordering of causal clauses that involve them. That is, these connectives may precede as well as follow the clause to which they are explicitly linked, signalling either forward or backward links (e.g., 我没办法去上学 (*wǒ méi bànfǎ qù shàngxué*; ‘I can’t go to school’) [result], 因为我肚子疼 (*yīnwèi wǒ dùzi téng*; ‘because my stomach hurts’) [reason] versus 因为我肚子疼 [reason], 我没办法去上学 [result]). In contrast to English where the preferred clause sequence in complex sentences is main to subordinate (Quirk et al. 1985), in Chinese, the modifier-modified principle of sequencing is widely accepted (Chao 1968; Tai 1975), indicating that “the main component comes at the end, and the subordinate component comes at the beginning” (Wang 1984: 96). This principle also applies to the causal clauses where reason connectives, such as *yīnwèi*, “exclusively signal forward links” (Kirkpatrick 1993: 426). Thus, forward linking has been accepted as the canonical ordering of causal clauses in Chinese (Chao 1968).

However, some other studies have found that forward linking is not necessarily the preferred order for expressing the causal relation in Chinese. For example, Biq (1995) examined the ordering of clauses involving *yīnwèi* that express an ideationally

determined causal relation. The results revealed that in conversation data, only 18 % ($n = 12/67$) of the tokens demonstrated forward linking, placing *yīnwèi* in the sentence-initial position, while the majority (82 %, $n = 55/67$) showed backward linking, placing *yīnwèi* in the sentence-medial position (Biq 1995: 53). In newspaper data, the rate of sentence-initial causal clauses is also lower compared to that of sentence-medial ones, but the distributional difference between these two variants is smaller (45 % vs. 55 %). Thus, in both spoken and written contexts, sentence-initial causal clauses are less favored, a finding that contrasts with the results of earlier studies (e.g., Chao 1968; Kirkpatrick 1993; Tai 1975; Wang 1984). In some follow up studies along the similar line, sentence-initial causal clauses have been consistently found to be less frequent than sentence-medial ones in Chinese (e.g., Song and Tao 2009; Wang 1999; Xiao et al. 2021). Thus, these studies suggest the existence of variation in causal clause ordering, but its patterning is not consistent across studies.

Nevertheless, examining the frequency differences between sentence-initial and sentence-medial causal clauses does not explain the mechanism of variation in causal clause ordering. To address this gap, some studies have tried to explain such variation by looking at the pragmatic functions of sentence-initial and sentence-medial causal clauses. For example, Song and Tao (2009: 95) found that sentence-initial causal clauses are more likely to function as an information-sharing device to provide background information for the main argument, while sentence-medial causal clauses tend to function as an interactional device to “mitigate dispreferred responses and address anticipated recipient doubts, surprises, or confusion”. Therefore, the pragmatic function of causal clauses has often been the reason used to explain the ordering of their connectives (Wang 1999, 2002).

Focusing on the frequency distribution of sentence-initial and sentence-medial causal clauses, along with their semantic and pragmatic connotations, risks oversimplifying these variants by categorizing them into rigid groups. This approach overlooks the variable processes underlying causal clause ordering and the nuanced contextual factors that shape the semantic, syntactic, and pragmatic relationships between causal clauses and their associated main clauses. Therefore, we shift our focus from merely analyzing frequency and function to adopting a variationist approach, which considers the influence of broader contextual factors on these two variants.

For contextual factors, it has been found that two factors – the type of connective and the presence of a paired result marker *suǒyǐ* – play important roles in influencing the ordering of causal clauses in Chinese. Specifically, anecdotal evidence suggests that *yīnwèi* and *yóuyú* are typically used to mark descriptive causality based on facts, while *jìrán* often conveys inferential causality based on reasoning (Xing 2001). A corpus-based analysis of three text categories – news reports, novels, and opinion pieces – further reveals that *jìrán* and sentence-medial *yīnwèi* consistently

exhibit subjective profiles across genres. In contrast, *yóuyú* and sentence-initial *yīnwèi* primarily express objective relations (Li et al. 2016). These findings suggest that the type of connective influences both the semantic meaning and the pragmatic functions of causal clauses. Additionally, the ordering of causal clauses may signal different types of causal relationships (e.g., Xiao et al. 2021). For the result marker *suǒyǐ*, it has been found that its presence generally promotes the use of sentence-initial causal clauses (Biq 1995). When *suǒyǐ* appears in the main clause, it tends to form a summary statement alongside reason connectives (e.g., *yīnwèi*, *yóuyú*) (Kirkpatrick 1993), which introduce background information in the subordinate clause, thus motivating the sentence-initial placement of these reason connectives. Although less frequent, *suǒyǐ* in the main clause can also appear in the sentence-initial position, followed by reason connectives in the subordinate clauses. This syntactic structure is often used to facilitate the progression of the main course of action, transitioning from a suspended state to the next action (Wang 2020: 328).

Another factor examined to explain the variation in causal clauses is the emotional valence of the main and subordinate clauses, although its impact on the ordering of causal clauses has yet to be explored (Yao 2007). Emotional valence here refers to “the speaker or writer’s attitude or stance towards, viewpoint or feelings about the entities and propositions that he or she is talking about” (Hunston and Thompson 2000: 5).² Adopting a three-way categorization of emotional valence in English (e.g., Partington 2004), Yao (2007, p. 53) annotated the attitudinal meanings of causal clauses with *suǒyǐ* and *jiěguǒ* in Chinese as positive (e.g., 所以深受市民支持 (*suǒyǐ shēn shòu shìmín zhīchí*; ‘so it is widely supported by the public’)), negative (e.g., 结果还是半途而废, 不了了之 (*jiéguǒ hái shì bàntú érfèi, bùliǎoliǎo zhī*; ‘so it was left unfinished and ended in failure’)), and neutral (e.g., 所以史书又叫汗青 (*suǒyǐ shǐshū yòu jiào hànqīng*; ‘so historical records are also called ‘汗青’ (an ancient term for historical records in Chinese)’)). The results revealed that 74 % of *suǒyǐ* clauses exhibited a neutral valence, while 14 % and 12 % displayed negative and positive valence, respectively. In contrast, 73 % of *jiěguǒ* clauses indicated a negative valence, while 8 % and 19 % showed neutral and positive valence, respectively. These patterns suggest that emotional valence is an important factor distinguishing between different types of causal clauses. This raises a new question, which we will address here: Does the emotional valence of clauses influence the ordering of causal clauses?

² Hunston and Thompson (2000) use the term “semantic prosody”, from which this definition is taken. Following a reviewer’s suggestion, we adopt “emotional valence” here as a more transparent term for this context.

Additionally, we introduce two new factors that have not yet been examined in studies of causal clause ordering in Chinese: the length and complexity of main and subordinate clauses. There are two key reasons for considering these factors. First, length and complexity have been extensively studied in English causal clauses and found to strongly influence their ordering (Diessel 2005). Second, given the substantial differences between Chinese and English syntax (Li and Thompson 1981; Wang 1984), it is likely that the effects of length and complexity on clause ordering in Chinese may follow distinct patterns. Thus, examining these factors in Chinese provides an opportunity to investigate potential cross-linguistic, typological universals in how syntactic features influence causal clause ordering.

For the length of main and subordinate clauses in English, it has often been measured by the number of words in each clause (including the subordinator, e.g., *because*). For example, Diessel (2005) looked at the ordering of multiple English adverbial clauses, including causal clauses, in three sources: conversational discourse, fiction, and scientific writing. Diessel found that when the adverbial clause is in sentence-initial position, it tends to be shorter than the main clause; conversely, in cases where the adverbial clause is in sentence-medial position, it tends to be longer than the main clause (see similar findings in Gries and Wulff 2021 regarding L1 English speakers). Diessel interprets this distribution of lengths as evidence for a general adherence to the short-before-long principle in English (Behaghel 1909), which reflects speakers' general attempts to minimize cognitive effort in utterance planning, according to parsing efficiency theories (Hawkins 1994, 2004).

For syntactic complexity, it focuses on the embeddedness of each clause which indicates whether the examined main or subordinate clause contains any embedded clauses (Diessel 2008), which has been widely used as a measure of complexity in studies of English. In English, syntactically complex constituents are more likely to appear towards the end of a sentence, as this allows the parser to quickly identify the high-level structure, thereby facilitating the analysis of subsequent constituents (Hawkins 1994, 2004; Wasow 1997). Given that causal clauses can vary in complexity, it is reasonable to assume that sentence-initial causal clauses are structurally less complex than sentence-medial ones. A corpus-based analysis of English "because" sentences revealed that subordinate causal clauses tend to be more complex with a greater degree of embeddedness, making them more likely to occur in sentence-medial positions (Xu and Kang 2022). These findings suggest that syntactic complexity influences the positioning of causal clauses, with more complex structures favoring sentence-medial positions, highlighting an intersection of clause complexity and causal clause ordering in English. Even though syntactic length and complexity have strong effects on the ordering of causal clauses in English, no studies have yet examined how these factors influence the positioning of causal clauses in Chinese, leaving a gap in our understanding of cross-linguistic patterns of clause ordering.

Furthermore, previous research on causal clause ordering in Chinese has shown that genre can influence the distribution of sentence-initial and sentence-medial variants. For example, Biq (1995) compared conversations and newspaper texts and found that sentence-medial causal clauses were more frequent in both genres, but occurred at a much higher rate in conversations than in newspapers (82 % vs. 55 %). Similarly, Song and Tao (2009), using the written Lancaster Corpus of Mandarin Chinese and the spoken CallFriend telephone conversation corpus, observed the same trend: sentence-medial causal clauses were more frequent in both genres, with a higher proportion in spoken than in written data (83 % vs. 68 %). These findings suggest that genre is a relevant factor in shaping causal clause positioning. These earlier studies were not grounded in a variationist framework and did not consider the role of contextual or probabilistic constraints, such as clausal complexity or emotional valence. Instead, they focused primarily on the functions of the two variants. This leaves open the question of how genre might interact with other probabilistic factors to condition clause ordering. Variationist studies in English morphosyntax, such as those on future temporal reference (*will* vs. *be going to*) (Engel and Szmrecsanyi 2023) and the dative alternation (e.g., *Tim gives Amy a book* vs. *Tim gives a book to Amy*) (Gan 2024), have often shown that genre interacts with broader probabilistic constraints. In this study, we test whether genre not only affects the overall distribution of sentence-initial and sentence-medial causal clauses but also interacts with other constraints identified in the literature above.

To summarize, variation in the ordering of causal clauses in Chinese is widely observed, with reason connectives such as *yīnwèi* and *yóuyú* either preceding or following the associated main clause, resulting in sentence-initial and sentence-medial positioning. This variation is not random but is systematically influenced by genre and contextual factors at syntactic, semantic, and discourse-pragmatic levels. However, no research has yet considered all these factors within a single multifactorial model, which would allow for concurrent measurement of the effect sizes and interactions among the different factors. In this study, we employ a large spoken corpus of Mandarin Chinese and adopt a variationist approach to examine the variation in causal clause ordering, focusing on the probabilistic influence of genre and multiple contextual constraints.

3 Methodology

3.1 Data

Spoken data has been considered “the primary data source for detecting the behaviour and evolution of grammar” (Biq 1995), so we sourced data from the

Diversified Spoken Chinese Uttered in Social Settings Corpus (DiSCUSS) for this analysis (Xu et al. 2022). DiSCUSS was compiled by the corpus research team led by Professor Jiajin Xu at Beijing Foreign Studies University, China (it can be accessed at <http://114.251.154.212/cqp/>). This is a one-million-word balanced corpus of spoken Mandarin Chinese, compiled following the data collection protocols of the widely used International Corpus of English (Greenbaum 1996). It covers 15 spoken genres, ranging from face-to-face conversations, phone calls, and broadcast interviews to spontaneous commentaries and legal presentations. These can be broadly categorized into two overarching types: dialogues and monologues. To avoid data sparsity issues in the regression model below, we use these two broader genre categories, dialogues and monologues, as a predictor of causal clause ordering. All recordings were transcribed *verbatim* on a character-by-character basis by over 15 trained team members and saved as plain text files (see the associated documents in Xu et al. 2022 for more details). For our analysis, regular expressions were used to systematically clean the text files – for example, by removing non-Chinese characters (e.g., code-switched English words) and inconsistent formatting. The Python package Jieba (<https://github.com/fxsjy/jieba>) was then employed for Chinese word segmentation. DiSCUSS is a reliable and widely used resource for researching spoken Mandarin Chinese. For instance, it has been used in studies on speech acts and motion verbs in Chinese by other researchers (Su and Fu 2023; Zhang and Xing 2024).

To extract tokens for analysis, we used AntConc 4.3.1 (Anthony 2022) to identify occurrences of *yīnwèi*, *yīn*, *yóuyú*, *jiànyú*, and *jìrán* from the entire DiSCUSS. Initially, we extracted 2,704 tokens involving the target reason connectives (*yīnwèi* = 2,417, *yīn* = 101, *yóuyú* = 98, *jiànyú* = 14, *jìrán* = 74). We then manually coded all these tokens to include only those that fit within the variable context.

3.2 Circumscribing the variable context

“The delimitation of the variable context within which the probabilistic constraints are operative” is crucial in conducting a variationist analysis (Torres Cacoullos and Travis 2019, p. 656). First, since reason connectives serve different functions, we limited the variable context to those that signal an ideational or propositional causal relation, as shown in Example 3. Following previous studies, tokens expressing all other functions – such as elaboration, discourse reflexive use, justification for requests/questions, and topic resumption – have been excluded (see Biq 1995: 51 for details of these functions). Second, tokens functioning as discourse markers have been excluded, as shown in Example 4. Third, to make the two genres in DiSCUSS, monologues and dialogues, more comparable, we included only sequential tokens produced by the same speaker in dialogues involving two or more participants.

Tokens that featured interruptions or interventions from interlocutors were excluded, as illustrated in Example 5. This approach ensured that, in both monologues and dialogues, our analysis focused on adjacent main and subordinate causal clauses, rather than on the broader discourse context. Fourth, following Liu and Zhu (2025), who conducted a logical and semantic comparison of the connective *jírán* with *yīnwèi* and *rúguǒ* (“if”) based on the factuality of the antecedents and consequents and the nature of their semantic relationships, we acknowledge that *jírán* can function in two distinct ways. Their analysis concluded that there are two types of *jírán* clauses: one aligning with causal connectives like *yīnwèi*, and the other with conditional connectives like *rúguǒ* (Liu and Zhu 2025: 18). Accordingly, in this study, we included only those *jírán* clauses that clearly function causally and excluded those that serve a conditional purpose, to maintain consistency in the analysis of causal clause structures. Fifth, 18 tokens containing *yīnwèi* are preceded by *zhī suǒyǐ*, forming the fixed expression *zhī suǒyǐ... shì/jiùshì/yěshì yīnwèi...*, all of which are sentence-medial. Given the formulaic nature of this construction and its lack of syntactic variability, these 18 tokens were excluded from the analysis.

After exclusions, 1,421 tokens remain for further analysis. Among them, 695 tokens (49 %) are sentence-initial, while 726 tokens (51 %) are sentence-medial, indicating that sentence-medial causal clauses are slightly more frequent overall. There are also distributional differences between monologues and dialogues: in dialogues ($n = 934$), 51 % of the tokens are sentence-initial and 49 % are sentence-medial, whereas in monologues ($n = 487$), 45 % are sentence-initial and 55 % are sentence-medial. This suggests that sentence-medial positioning is more common in monologues than in dialogues.

- (3) 你 不 能够 单纯 的 只是 百 分
 nǐ bù nénggòu dānchún de zhǐshì bǎi fēn
 you not be.able simply POSS only hundred percent
 百 的 复古, 因为 你 还 要 了解
 bǎi de fùgǔ, yīnwèi nǐ hái yào liǎojiě
 hundred POSS retro, because you still need understand
 当前 的 人 在 想 什么。
 dāngqián de rén zài xiǎng shénme.
 current POSS person at think what,
 ‘You cannot just be 100% retro, because you also need to
 understand what people today are thinking.’

(S1B_031)

- (4) 这里 很 好, 因为, 虽然 这里 很 安静,
 zhèlǐ hěn hǎo, yīnwèi,... suīrán zhèlǐ hěn ānjìng,
 here very good because although here very quiet
 我 也 放 松 不 了。
 wǒ yě fāngsōng bùliǎo.
 I also relax unable
 'It's very nice here because, although it's very quiet, I still can't relax.'
 (S1A_001)
- (5) 小 军: 还 有 这 个 水, 水 也 是 要 算。
 Xiǎo Jūn: hái yǒu zhège shuǐ, shuǐ yě shì yào suàn.
 also have this water, water also be need calculate
 'There's also this water; water should be counted as well.'
 小 峰: 为 什 么?
 Xiǎo Fēng: Wèishéme?
 why
 'Why?'
 小 军: 因 为 他 提 到 了 水, 还 有 这 个 是...
 Xiǎo Jūn: yīnwèi tā tídao le shuǐ, hái yǒu zhège shì...
 because he mention PERF water, also have this be...
 'Because he mentioned water, and also this one.'
 (S1A_011)

3.3 Annotating the probabilistic constraints

All of the target tokens were annotated to determine the contextual conditioning of the variation between sentence-initial and sentence-medial causal clauses. As discussed in Section 2, these constraints include genre (dialogue vs. monologue), clause length, clause complexity, clause emotional valence, connective type, and the presence of a paired connector *suǒyǐ*. Their annotations and distributions are detailed below. Since the variable context defined above is limited to adjacent main and subordinate causal clauses, we do not consider some very broad discourse-level factors such as information structure (e.g., new vs. given information) (see Chafe 1984 for the effect of information structure on adverbial clauses in English) or topic continuity and coreferentiality (e.g., same vs. different topics) (see Schiffrin 1985 for the effect of topic continuity on causal clauses in English). While examining these factors, which are (implicitly) embedded in a much broader context, would provide valuable insights into Chinese causal clauses, it may be more appropriate to

approach this using frameworks beyond the variationist model, such as discourse analysis or interactional linguistics.

3.3.1 Clause length

Clause length is measured by the number of characters in either the main or subordinate clause. This is exemplified in Example 6, where the main clause contains 8 characters, and the subordinate clause contains 25 characters.

- (6) 暂时 还 不 会, 还 不 会,
 zànshí hái bù huì, hái bù huì,
 temporarily still not happen still not happen,
 因为 我们 知道 这 一 艘 核潜艇,
 yīnwèi wǒmen zhīdào zhè yī sōu héqiántǐng,
 because we know this one CLF nuclear.submarine,
 并 不 像 印度 说 的 那么 神乎其神。
 bìng bù xiàng yìndù shuō de nàme shénhūqíshén.
 and not like India say REL that incredible.
 'For now, it's still not likely, still not likely, because we know that this nuclear submarine is not as extraordinary as India claims.'
 (S1B_038)

3.3.2 Clause complexity

We distinguished complexity in both main and subordinate clauses with a binary categorization. Any main or subordinate clause that is embedded in another clause ("a verb with its arguments" (see Cumming 1984: 374 for more details)) is coded as a complex clause; otherwise, it is coded as a simple clause, as illustrated in Example 7.

- (7) 我 支持 甲方, 因为 孩子 在 学校 里 更
 wǒ zhīchí jiǎfāng, yīnwèi háizi zài xuéxiào lǐ gèng
 I support party.A, because child at school inside more
 多 是 在 听, 回到 家里 更 多 是
 duō shì zài tīng, huí dào jiā lǐ gèng duō shì
 much is at listen, return.to home more much is

对话 和 交流。
 duìhuà hé jiāoliú.

conversation and communication.

‘I support Party A (**simple main clause**), because children listen more at school, while at home, there is more conversation and interaction (**complex subordinate clause**).’

(S1B_058)

3.3.3 Emotional valence

Emotional valence here refers to the overall emotional or evaluative tone conveyed by a clause. We adopt a widely used three-way categorization in English, classifying it as positive, negative, or neutral (e.g., Partington 2004). Following Yao (2007: 53), we focus on the semantic connotations of the “core words” in the predicate when annotating the tendencies of each clause’s emotional valence in Chinese. This annotation is illustrated in Example 8.

- (8) 对于 我们 来 讲, 每 增加 一个 键 都 非常
 Duiyú wǒmen lái jiǎng, měi zēngjiā yīgè jiàn dōu fēicháng
 for we come say, every add one key all very
 非常 纠结, 因为 我们 确实 希望 让 用户
 fēicháng jiūjié, yīnwèi wǒmen quèshí xīwàng ràng yòngù
 very tangled, because we indeed hope let user
 非常 简洁 能 操作。
 fēicháng jiǎnjié néng cāozuò.
 very concise can operate.
 ‘For us, adding each key is extremely stressful (**negative main clause**),
 because we truly hope to make it simple and easy to operate for the user
 (**positive subordinate clause**).’

(S2A_055)

3.3.4 Connective type

Previous research has argued that *yīnwèi* and *yóuyú* are more similar to each other, while *jìrán* shows some semantic differences (Li et al. 2016; Xing 2003: 199). This suggests that connective type may affect the ordering of causal clauses. To assess this, we annotated each token by connective type, as illustrated in Examples 9–11.

- (9) 因为 卖 出去 就 没有 稻草人 可以 打 了,
 yīnwèi mài chūqù jiù méiyǒu dàoǎorén kěyǐ dǎ le,
 because sell go.out then not.have scarecrow can hit PFV,
所以 会 多方 施压。
 suǒyǐ huì duōfāng shīyā.
 therefore will multiple.side apply.pressure.
 ‘Because once it’s sold, there will be no strawman to attack, so pressure will come from multiple sides.’

(S1B_029)

- (10) 由于 它 海拔 很 高,
 yóuyú tā hǎibá hěn gāo,
 due.to it altitude very high,
所以 它 在 环境 方面 更 接近 北方。
 suǒyǐ tā zài huánjìng fāngmiàn gèng jiējìn běifāng.
 therefore it in environment aspect more close.to north.
 ‘Because its altitude is very high, it is more similar to the north in terms of its environment.’

(S1B_008)

- (11) 既然 我 走路 不 方便, 儿子 你 去 自己 买。
 jìrán wǒ zǒulù bù fāngbiàn, érzi nǐ qù zìjǐ mǎi.
 since I walk not convenient, son you go self buy.
 ‘Since it’s inconvenient for me to walk, son, you go and buy it yourself.’

(S1B_012)

It should be noted that the distribution shows that *yīnwèi* is the most frequent connective in the dataset (94 %, $n = 1,339/1,421$) and tends to co-occur with sentence-medial causal clauses. In contrast, *jìrán*, *yóuyú*, and *jiànyú* show a strong preference for sentence-initial causal clauses. Given this distribution, we grouped *jìrán*, *yóuyú*, and *jiànyú* into a single category, non-*yīnwèi*, to facilitate the subsequent statistical analysis.

3.3.5 Presence of a paired connector *suǒyǐ*

The paired use of a reason connector, especially *yīnwèi*, and a result connector *suǒyǐ* has been argued to function as a form of double marking, signaling that the utterance serves as a summary statement (Biq 1995; Kirkpatrick 1993; Xiao 2020). To test the effect of this pairing on causal clause ordering, we annotated the presence or absence of *suǒyǐ* in main clauses, as shown in Examples 9–11 above.

3.4 Statistical analysis

We used mixed-effects generalized logistic regression modeling in R to perform multifactorial analyses on the data (R Core Team 2025). This approach enabled us to simultaneously estimate the effects of all identified contextual predictors on the alternation between sentence-initial and sentence-medial causal clauses. Additionally, it allowed for the inclusion of potential covariates, such as Text ID, thereby enhancing the generalizability of our findings to a broader user population within a specific language variety (Gries 2021).

To ensure a more meaningful comparison between the numerical predictors, `MainClauseLength` and `SubClauseLength`, and other categorical predictors, we standardized these two factors by dividing them by two standard deviations and centering them around the mean (Gelman 2008). For the random-effects structure, we incorporated the random intercept for `TextID`. Since some participants showed consistent usage of either sentence-initial or sentence-medial causal clauses without variation, including them individually in the model would not yield valuable insights (Gan 2024; Gan and Wang 2025). As a result, files containing only one token were grouped into an ‘other’ category, representing 14 % of the total files ($n = 32/233$).

After refining the factors, a comprehensive regression model was applied to examine all pairwise interactions between the fixed-effect factors: $\text{Ordering} \sim (\text{Genre} + \text{MainClauseLength} + \text{SubClauseLength} + \text{MainClauseComplexity} + \text{SubClauseComplexity} + \text{MainClauseEmotionalValence} + \text{SubClauseEmotionalValence} + \text{ConnectiveType} + \text{PresenceSuoyi})^2 + (1|\text{TextID})$. The regression analysis was performed using the `lme4` package (Bates et al. 2015). We followed a backward stepwise model selection procedure, starting with the full model and iteratively removing non-significant predictors in the fixed-effects structure. The `drop1()` function was used for this step, based on likelihood-ratio tests, with a significance threshold of 0.05. For all significant interaction effects, we used the `emmeans` package to perform post-hoc comparisons between the different levels of the predictors (Lenth 2025).

4 Results

4.1 Model summary

The final model identified three significant interaction effects predicting the ordering of causal clauses: `Genre * MainClauseEmotionalValence`, `MainClauseComplexity * MainClauseLength`, and `PresenceSuoyi * ConnectiveType`. Additionally, two significant main effects were found for `SubClauseLength` and

SubClauseEmotionalValence. On the other hand, SubClauseComplexity was excluded from the model as it did not contribute significantly to model fitting.

The model demonstrates good performance, with a concordance C index of 0.81, indicating a strong fit (Baayen 2008). It achieved an overall accuracy of 75 % (95 % CI: 73–77 %), which is significantly better than the no-information rate of 51 % ($p < 0.001$), indicating that the model performs substantially above chance level. The Cohen's Kappa coefficient was 0.50, reflecting moderate agreement between predicted and true classes beyond chance. Regarding class-specific performance, the model showed a sensitivity of 82 % for correctly identifying the sentence-medial class, and a specificity of 68 % for correctly identifying the sentence-initial class. The positive predictive value was 73 %, and the negative predictive value was 78 %. The balanced accuracy, which accounts for class imbalance, was 75 %. Overall, these results demonstrate that the model has satisfactory discriminatory ability for distinguishing between sentence-initial and sentence-medial clause orderings in context. The random effect in the model, TextID, accounts for a variance of 0.19 with a standard deviation of 0.44. Multicollinearity diagnostics show that the largest variance inflation factor (VIF) is 5.06, suggesting acceptable levels of collinearity, and the condition index k is 7.19, indicating no severe issues with multicollinearity in the model (Gries 2021).

4.2 Summary of effects

To facilitate interpretation of the results, we summarize the main and interaction effects from the fitted model in Table 1. Statistically significant effects are highlighted in grey. It is important to note that predictors such as MainClauseEmotionalValence, MainClauseLength, PresenceSuoyi, and ConnectiveType may appear to exhibit main effects but are, in fact, involved in interactions with other predictors. As such, they are not discussed here (see Gries 2021: 291 for details on this statistical decision). The predicted levels for all categorical variables in the model are shown in brackets, while the reference levels are as follows: Dialogue for Genre, Neutral for both MainClauseEmotionalValence and SubClauseEmotionalValence, Complex for MainClauseComplexity, Absence for PresenceSuoyi, and Yinwei for ConnectiveType.

Our analysis focuses on both the directions and magnitudes of these factors. In Table 1, positive z -values for each predictor indicate a preference for sentence-medial causal clauses, given that the predicted variant is sentence-medial causal clauses, while negative z -values suggest a preference for sentence-initial causal clauses. To aid interpretation, we also report odds ratios for each predictor to convey effect sizes. An odds ratio greater than 1 indicates a higher likelihood of sentence-medial causal clauses, whereas a value less than 1 signifies a greater likelihood of

Table 1: Output of logistic mixed-effects model predicting Medial. OR, odds ratio; SE, standard error.

	OR	SE	z	p-Value
(Intercept)	2.66	0.16	5.96	<0.001
Genre (Monologue)	0.64	0.23	-1.94	0.053
MainClauseEmotionalValence (Negative)	0.58	0.19	-2.86	<0.01
MainClauseEmotionalValence (Positive)	0.83	0.20	-0.93	0.353
MainClauseComplexity (Simple)	0.8	0.16	-1.40	0.162
MainClauseLength	0.35	0.20	-5.18	<0.001
PresenceSuoyi (Presence)	0.1	0.19	-11.73	<0.001
ConnectiveType (Non-yinwei)	0.03	0.61	-5.91	<0.001
SubClauseLength	1.84	0.14	4.39	<0.001
SubClauseEmotionalValence (Negative)	1.1	0.16	0.62	0.537
SubClauseEmotionalValence (Positive)	1.52	0.19	2.24	<0.05
Genre (Monologue): MainClauseEmotionalValence (Negative)	2.64	0.32	3.04	<0.01
Genre (Monologue): MainClauseEmotionalValence (Positive)	1.77	0.33	1.71	0.087
MainClauseComplexity (Simple): MainClauseLength	5.8	0.40	4.40	<0.001
PresenceSuoyi (Presence): ConnectiveType (Non-yinwei)	19.88	1.25	2.39	<0.05

sentence-initial causal clauses. For example, the predictor SubClauseLength has an odds ratio of 1.84, indicating that for every one-character increase in subordinate clause length, the odds of a causal clause being sentence-medial (as opposed to sentence-initial) increase by 84 %, holding all other predictors constant.

To further interpret the directions and effect sizes of significant predictors, all effects are visualized in Figures 1–4. Partial effects plots were utilized, offering insights into how individual predictors influence the outcome while holding other predictors constant or at their average levels. The x-axes display the predictor names, while the y-axes, ranging from 0 to 1 in each figure, represent the predicted probability of sentence-medial causal clauses as estimated by the model. Higher values on the y-axis indicate a greater probability of sentence-medial causal clauses compared to sentence-initial causal clauses.

As shown in Figure 1, the main effect of SubClauseLength, displayed on the left panel, reveals that as the length of the subordinate clause increases, the probability of sentence-medial causal clauses also increases. This suggests that longer subordinate clauses tend to favor sentence-medial causal clauses compared to shorter ones, as illustrated in Example 6 above. In the right panel, the main effect of SubClauseEmotionalValence shows that the probability of sentence-medial causal clauses is highest when the emotional valence is positive, compared to when it is neutral or negative. Therefore, when the emotional valence of subordinate clauses is

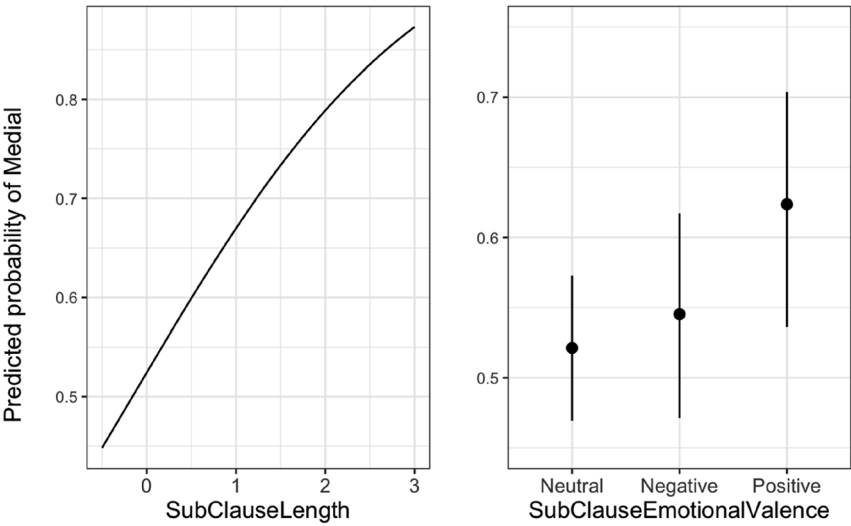


Figure 1: Plots of the main effects of subordinate clause length and emotional valence in the model.

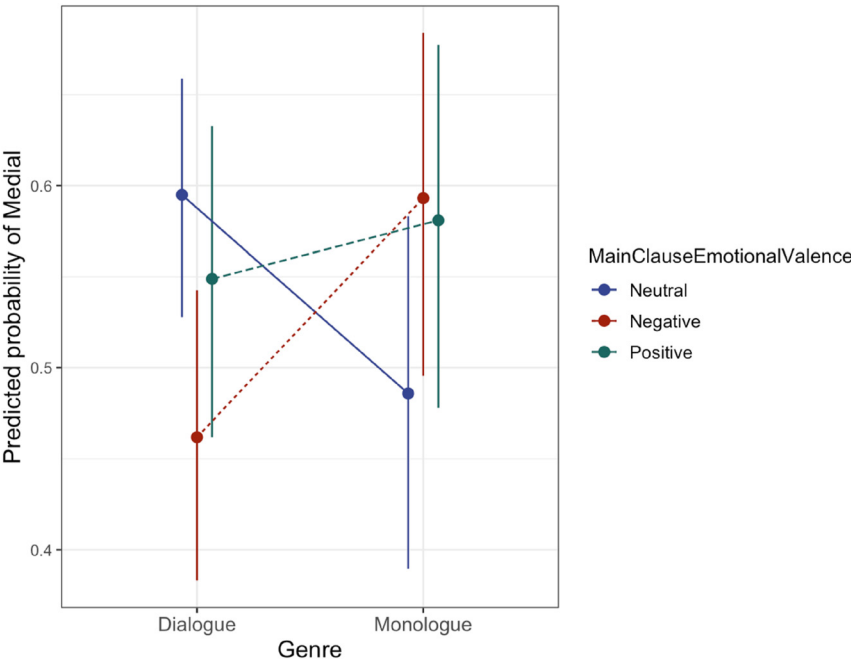


Figure 2: Plot of the interaction effect between genre and main clause emotional valence in the model.

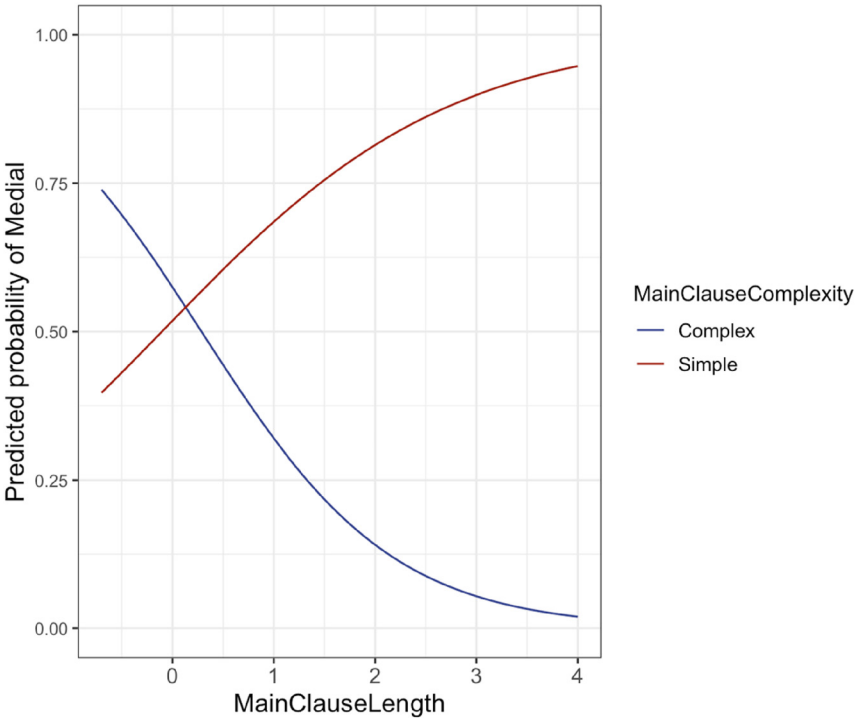


Figure 3: Plot of the interaction effect between main clause length and complexity in the model.

positive, sentence-medial causal clauses are most favored over sentence-initial ones, as demonstrated in Example 8 above.

Figure 2 illustrates the interaction effect between `MainClauseEmotionalValence` and `Genre`. Three types of emotional valence – neutral, negative, and positive – are represented by blue, red, and green lines, respectively. In dialogues, neutral valence of main clauses is the most favorable for sentence-medial causal clauses, whereas negative valence is the least favorable, with a significant difference between them, as confirmed by post-hoc comparisons using `emmeans` in R (estimate = 0.54, $p = 0.04$). Positive valence falls between neutral and negative valence in terms of favorability for sentence-medial clauses. In monologues, both positive and negative valence tend to favor sentence-medial causal clauses when compared to neutral valence; however, these differences do not reach statistical significance ($ps > 0.05$).

When comparing the effect of emotional valence between dialogues and monologues, the results show that with neutral valence, the probability of sentence-medial causal clauses is higher in dialogues than in monologues (estimate = 0.44,

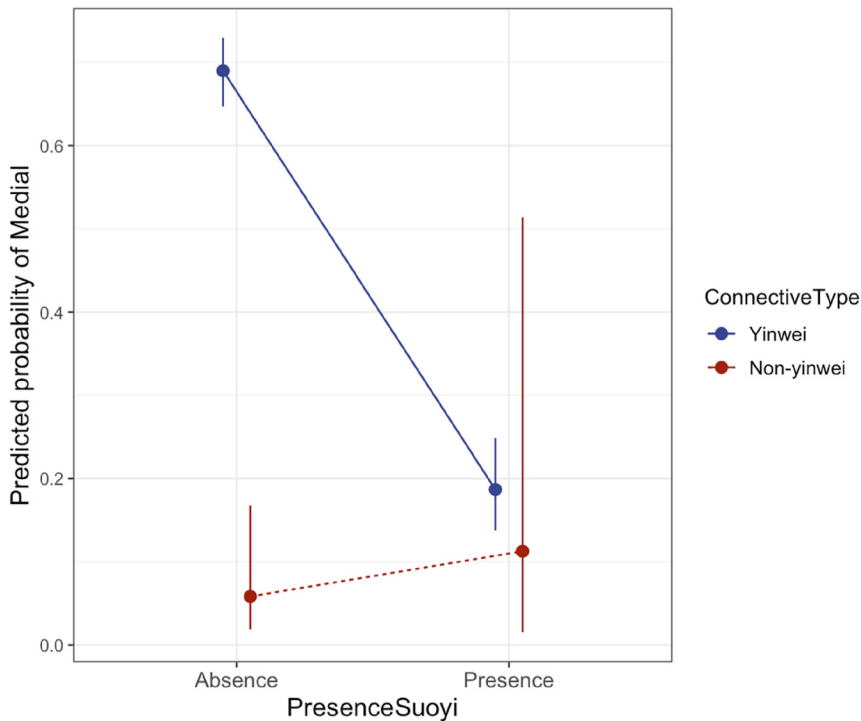


Figure 4: Plot of the interaction effect between the presence of *suǒyǐ* and connective type in the model.

$p = 0.38$). In contrast, with negative valence, the probability of sentence-medial causal clauses is greater in monologues than in dialogues (estimate = -0.53 , $p = 0.25$). Similarly, with positive valence, the probability of sentence-medial causal clauses is higher in monologues than in dialogues (estimate = -0.13 , $p = 1.00$). It should be noted that none of these differences reach statistical significance.

Figure 3 illustrates the interaction effect between main clause length and main clause complexity. The red line shows that when main clauses are syntactically simple, an increase in main clause length corresponds to a higher probability of sentence-medial causal clauses. This indicates that longer but simple main clauses favor sentence-medial over sentence-initial causal clauses. In contrast, the blue line shows that when main clauses are syntactically complex, an increase in main clause length corresponds to a lower probability of sentence-medial causal clauses. This suggests that shorter but complex main clauses favor sentence-medial over sentence-initial causal clauses.

Figure 4 illustrates the interaction effect between the presence of *suǒyǐ* in the main clause and the type of connective. When *suǒyǐ* is absent in the main clause and *yīnwèi* is used (blue line), sentence-medial causal clauses are most favored. In contrast, when *suǒyǐ* is absent in the main clause and non-*yīnwèi* connectives are used (red line), sentence-medial causal clauses are least favored. A post-hoc comparison confirms that these two contexts differ significantly (estimate = 3.58, $p < 0.0001$). When both *yīnwèi* and *suǒyǐ* are present, the probability of sentence-medial causal clauses is higher than in contexts where *suǒyǐ* is present but non-*yīnwèi* connectives are used. However, this difference is not statistically significant (estimate = 0.59, $p = 0.95$).

When connective types are compared, new patterns emerge. For *yīnwèi*, sentence-medial causal clauses are significantly more favored when *suǒyǐ* is absent in the main clause compared to when it is present (estimate = 2.27, $p < 0.0001$). In contrast, for non-*yīnwèi* connectives, the probability of sentence-medial causal clauses remains consistently low, regardless of whether *suǒyǐ* is absent or present. This suggests that non-*yīnwèi* connectives generally favor sentence-initial causal clauses over sentence-medial ones (estimate = -0.72 , $p = 0.94$).

5 Discussion

Adopting a variationist approach to analyze a large speech corpus of Mandarin Chinese, the data reveals that sentence-medial causal clauses occur slightly more frequently than sentence-initial ones (51 % vs. 49 %). While this difference is smaller than reported in some previous studies (Biq 1995; Li et al. 2016; Song and Tao 2009), it challenges earlier claims that sentence-initial positioning represents the canonical word order for causal clauses in Chinese (Chao 1968; Kirkpatrick 1993; Young 1994). This finding aligns with arguments from studies that emphasize the variation of word order in causal clauses (e.g., Biq 1995; Song and Tao 2009; Wang 1999). Moreover, this variation mirrors patterns observed in many other languages, such as English, German, Japanese, and Santali (Munda, India). In these languages, formally, causal clauses lack signs of syntactic embedding, and functionally, they act as independent assertions that explain or support the semantically associated main clause (Diessel and Hetterle 2011). Their discourse-pragmatic role contrasts with that of other adverbial clauses, which typically favor sentence-initial positions to establish broader discourse coherence. Thus, causal clauses are more likely to occur sentence-medially, where they remain largely independent of the preceding main clause (Diessel 2005).

We also compared the usage rates between dialogues and monologues, finding that sentence-medial causal clauses occur slightly more frequently in monologues

than in dialogues (55 % vs. 49 %). Sentence-medial causal clauses in Chinese have been argued to serve as an information-sharing device, adding context to the preceding discourse, whereas sentence-initial causal clauses function as an interactional device, setting a discourse frame for the upcoming discussion (Song and Tao 2009; Wang 1999). Therefore, in monologues, where speakers have more time and control over the flow of information, allowing for more detailed explanations, elaborations, or arguments, they tend to use more sentence-medial causal clauses.

Based on the multifactorial inferential analysis of the data, which considered multiple factors within a single statistical model, we observed several patterns regarding the syntactic factors influencing causal clause ordering in Chinese, which have been overlooked in previous research. Subordinate clause length exhibited a significant main effect, with longer subordinate clauses favoring sentence-medial positions over sentence-initial ones. This finding aligns with parsing efficiency theories (Hawkins 1994, 2004), which propose that longer constituents tend to appear towards the end of a sentence that facilitate efficient processing. However, subordinate clause complexity did not demonstrate either a main effect or an interaction effect with subordinate clause length. This suggests that, in processing Chinese causal clauses, subordinate clause length carries relatively greater cue weighting and serves as a more valid and salient cue for their positioning than complexity, according to the Competition Model in sentence processing (Bates and MacWhinney 1987; MacWhinney et al. 1984). The distinction between these two dimensions underscores the need for future studies to treat subordinate clause length and complexity as separate factors, not only in causal clauses but also in other types of adverbial clauses in Chinese. Furthermore, there is an interaction between main clause length and main clause complexity, which provides further insights into the variation of causal clause ordering. Specifically, when main clauses are syntactically simple, increased length correlates with a higher likelihood of sentence-medial causal clauses. This pattern likely reflects a cognitive preference for balancing the information load within a sentence. A long but simple main clause may ease processing demands, particularly when followed by a subordinate causal clause in a sentence-medial position. Conversely, when main clauses are syntactically complex, increased length decreases the probability of sentence-medial causal clauses. This suggests that the combination of complexity and length introduces a heavier cognitive load, making sentence-initial placement of subordinate causal clauses a preferable strategy to enhance comprehensibility.

These findings on syntactic length and complexity in Chinese only partially align with previous research on English (Diessel 2008; Wasow 1997; Wasow and Arnold 2003). In English, longer and more complex clauses tend to appear sentence-finally to facilitate parsing by allowing the main structure of the sentence to be processed first (Hawkins 1994; Wasow 1997). In contrast, Chinese demonstrates a distinct strategy:

while longer subordinate clauses are associated with sentence-medial positioning, complexity – particularly in subordinate clauses – does not significantly influence placement. This divergence between English and Chinese suggests that the latter employs unique strategies for managing parsing and processing demands, likely shaped by its typological characteristics, such as its topic-prominent structure and flexible word order (Gasde and Paul 1996; Li and Thompson 1976, 1981; Thompson et al. 2007; Wang 1984). Furthermore, the observed preference for sentence-medial positioning of causal clauses with long but simple main clauses, and sentence-initial positioning of causal clauses with short but complex main clauses in Chinese, highlights this distinctive approach to balancing syntactic and cognitive processing requirements. These findings underscore the importance of considering language-specific constraints when examining syntactic phenomena such as clause ordering. They also demonstrate that generalizations based on English may not always apply to Chinese, particularly regarding the interplay of length and complexity.

For the semantic factor, emotional valence, the regression model underscores its important role in modulating the positioning of causal clauses in Chinese, a factor also overlooked in previous studies. Specifically, the main effect of subordinate clause emotional valence indicates that positive valence favors sentence-medial positioning of causal clauses, suggesting that positive, favorable semantic content integrates more smoothly into the ongoing discourse than neutral or negative valence (Hunston 2007; see also Hunston and Thompson 2000; see Partington 2004 for similar functions of emotional valence in English). Additionally, an interaction between main clause emotional valence and genre further influences clause positioning. In dialogues, the preference for neutral valence of main clauses in sentence-medial positioning of causal clauses reflects the cooperative nature of conversational exchanges, where less emotive content promotes clarity and balance. In contrast, monologues tend to favor both positive and negative valence in sentence-medial positioning of causal clauses, reflecting a narrative style that foregrounds evaluative content within extended discourse. These findings highlight the nuanced effects of genre and emotional valence on the variation between sentence-medial and sentence-initial causal clauses across diverse communicative and emotional contexts, aligning with arguments in the register variation framework (Biber and Conrad 2019; Goulart et al. 2020).

For the factors at the discourse-pragmatic level, there is an interaction between the type of connective and the presence of the result marker *suǒyǐ*. Specifically, when *suǒyǐ* is absent and *yīnwèi* is used, sentence-medial causal clauses are most favored. This finding suggests a strong integration of *yīnwèi* with sentence-medial positioning, corroborating findings from other studies (Wang 1999; Xiao et al. 2021). This could

reflect the cooperation of *yīnwèi* and the absence of *suǒyǐ* allows for greater flexibility in clause positioning, facilitating a smoother connection between the main and causal clauses in ongoing discourse. As found in previous studies, when *suǒyǐ* appears in the main clause (which may mark a stronger or more explicit causal relationship), it tends to form a summary statement, alongside *yīnwèi* clauses, which introduce background information in the subordinate clause, motivating sentence-initial rather than sentence-medial positioning of *yīnwèi* (Biq 1995; Kirkpatrick 1993). However, when *suǒyǐ* is absent and non-*yīnwèi* connectives such as *jìrán*, *yóuyú*, and *jiànyú*, are used, sentence-medial clauses are least favored. This pattern may suggest that *certain* non-*yīnwèi* connectives are more likely to occur in sentence-initial position, possibly reflecting a preference for more explicit or rhetorically foregrounded causal structures (Li et al. 2016). While some of these connectives, such as *yóuyú* and *jiànyú*, have been associated with more objective causal relations (as demonstrated in Example 11 above and in Xiao et al. (2021)), this does not uniformly apply to all non-*yīnwèi* forms. For example, *jìrán* tends to encode subjective causal relations, indicating that the relationship between connective type and clause position is not strictly determined by the objective-subjective distinction (Xiao et al. 2021). Moreover, because non-*yīnwèi* connectives are generally less semantically prototypical for expressing causality than *yīnwèi* (Li et al. 2016; Xing 2003), they may be less integrated into the discourse flow, contributing to their tendency to appear sentence-initially. Overall, these results point to a complex interplay between connective type, the presence of *suǒyǐ*, and clause ordering preferences in spoken discourse.

It is worth noting that this study focuses exclusively on utterances containing explicit causal connectives (e.g., *yīnwèi*, *yóuyú*, *jìrán*) and does not include instances of implicit causal relations – those in which causality is inferred from context rather than overtly marked. While implicit causal relations are common in spoken Chinese discourse (cf. Biq 1995), their identification requires close, manual reading of each utterance and contextually grounded interpretation, which differs fundamentally from the automated extraction approach adopted in this study. The probabilistic constraints that govern clause ordering in implicit causal constructions may also diverge from those in explicitly marked ones. For instance, features such as double marking with *yīnwèi* and *suǒyǐ*, which are relevant to the analysis of explicit constructions, do not occur in implicit contexts (cf. Xiao 2020). Therefore, utterances with implicit causal relations are not directly comparable to those containing explicit connectives in terms of both extraction method and probabilistic constraints. An investigation of implicit causality in spoken Mandarin would require a separate, dedicated study using different theoretical frameworks and analytical procedures.

6 Conclusions

This corpus-based multifactorial analysis of spoken Mandarin Chinese provides compelling evidence for the variation in the sequential ordering of causal clauses, with reason connectives (e.g., *yīnwèi*, *yóuyú*) appearing either before or after the main clause (sentence-initial vs. sentence-medial positions). The analysis reveals that this variation is not random but is systematically constrained by syntactic, semantic, and discourse-pragmatic factors, aligning with the fundamental “orderly heterogeneity” argument within the variationist framework (Weinreich et al. 1968: 100). Notably, robust interactions were identified between factors. For instance, the interaction between two discourse-pragmatic factors, the type of connective and the presence of *suǒyǐ*, demonstrates that the preference for sentence-medial clauses with *yīnwèi* in the absence of *suǒyǐ* reflects a flexible, discourse-driven process of clause positioning. In contrast, the preference for sentence-initial clauses with non-*yīnwèi* connectives points to a more rigid structure governing certain types of causal relationships. Furthermore, there is an interaction between main clause length and its complexity, underscoring the nuanced role of syntactic factors.

Together, these findings contribute new evidence on the multifactorial conditioning of causal clause ordering in Mandarin Chinese, offering novel insights into sentence structure and discourse cohesion in the language. As this study focuses exclusively on utterances with explicit connectives, future research should explore implicit causality in Chinese, which is a complementary area for deepening our understanding of causal expression in discourse.

Acknowledgments: We are very grateful to Stefan Th. Gries, Co-Editor-in-Chief of CLLT, and to two anonymous reviewers for their constructive feedback on an earlier draft of this article.

References

- Adli, A. & G. R. Guy. 2022. Globalising the study of language variation and change: A manifesto on cross-cultural sociolinguistics. *Language and Linguistics Compass* 16(5–6). 1–15.
- Anthony, L. 2022. *AntConc (Version 4.0.10) [Computer Software]*. Tokyo, Japan: Waseda University Available at: <https://www.laurenceanthony.net/software>.
- Baayen, R. H. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
- Bates, D., M. Mächler, B. Bolker & S. Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.
- Bates, E. & B. MacWhinney. 1987. Competition, variation, and language learning. In B. MacWhinney (ed.), *Mechanisms of language acquisition*, 157–193. Hillsdale: Erlbaum.

- Behaghel, O. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25. 110–142.
- Biber, D. & S. Conrad. 2019. *Register, genre, and style*, 2nd edn. New York: Cambridge University Press.
- Biq, Y.-O. 1995. Chinese causal sequencing and *yinwei* in conversation and press reportage. In *Proceedings of the twenty-first annual meeting of the Berkeley linguistics society*, 47–60. Berkeley.
- Bortoni-Ricardo, S. M. 1985. *The urbanization of rural dialect speakers: A sociolinguistic study in Brazil*. Cambridge: Cambridge University Press.
- Chafe, W. 1984. How people use adverbial clauses. In *Proceedings of the tenth annual meeting of the Berkeley linguistics Society*, 437–449. Berkeley.
- Chao, Y. R. 1968. *A grammar of spoken Chinese*. Berkeley: University of California Press.
- Cumming, S. 1984. The sentence in Chinese. *Studies in Language* 8(3). 365–395.
- Diessel, H. 2005. Competing motivations for the ordering of main and adverbial clauses. *Linguistics* 43(3). 449–470.
- Diessel, H. 2008. Iconicity of sequence: A corpus-based analysis of the positioning of temporal adverbial clauses in English. *Cognitive Linguistics* 19(3). 465–490.
- Diessel, H. & K. Hetterle. 2011. Causal clauses: A cross-linguistic investigation of their structure, meaning, and use. In P. Siemund (ed.), *Linguistic universals and language variation*, 23–54. Berlin, New York: De Gruyter Mouton.
- Engel, A. & B. Szmrecsanyi. 2023. Variable grammars are variable across registers: Future temporal reference in English. *Language Variation and Change* 34(3). 355–378.
- Gan, Q. 2024. Different registers, different grammars in second language production? The dative alternation in spoken and written Chinese learner English. *Lingua* 309. 103790.
- Gan, Q. & M. Wang. 2025. Examining contextual constraints on the English dative alternation in L2 written production: A contrastive multifactorial analysis. *International Journal of Corpus Linguistics*. <https://doi.org/10.1075/ijcl.23044.gan>.
- Gasde, H.-D. & W. Paul. 1996. Functional categories, topic prominence, and complex sentences in Mandarin Chinese. *Linguistics* 34(2). 263–294.
- Gelman, A. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* 27(15). 2865–2873.
- Goulart, L., B. Gray, S. Staples, A. Black, A. Shelton, D. Biber, J. Egbert & S. Wizner. 2020. Linguistic perspectives on register. *Annual Review of Linguistics* 6(1). 435–455.
- Greenbaum, S. 1996. *Comparing English worldwide: The international corpus of English*. Oxford: Clarendon.
- Gries, S. T. 2001. A multifactorial analysis of syntactic variation: Particle movement revisited. *Journal of Quantitative Linguistics* 8(1). 33–50.
- Gries, S. T. 2021. *Statistics for linguistics with R: A practical introduction*, 3rd edn. Berlin, Boston: Mouton de Gruyter.
- Gries, S. T. & S. Wulff. 2021. Examining individual variation in learner production data: A few programmatic pointers for corpus-based analyses using the example of adverbial clause ordering. *Applied PsychoLinguistics* 42(2). 279–299.
- Hawkins, J. A. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.
- Hawkins, J. A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.
- Hunston, S. & G. Thompson. 2000. *Evaluation in text: Authorial stance and the construction of discourse*. New York: Oxford University Press.
- Hunston, S. 2007. Semantic prosody revisited. *International Journal of Corpus Linguistics* 12(2). 249–268.
- Kirkpatrick, A. 1993. Information sequencing in modern standard Chinese in a genre of extended spoken discourse. *Text & Talk* 13(3). 423–453.

- Labov, W. 1963. The social motivation of a sound change. *Word* 19(3). 273–309.
- Labov, W. 1994. *Principles of linguistic change. Volume 1: Internal factors*. Oxford: Blackwell.
- Lenth, R. V. 2025. emmeans: Estimated marginal means, aka least-squares means. *R package version 1.11.1*. Available at: <https://cran.r-project.org/package=emmeans>.
- Li, F., T. Sanders & J. Evers-Vermeul. 2016. On the subjectivity of Mandarin reason connectives: Robust profiles or genre-sensitivity? In N. Stukker, W. Spooren & G. Steen (eds.), *Genre in language, discourse and cognition*, 15–50. Berlin, Boston: De Gruyter Mouton.
- Li, Y., B. Szendrői & W. Zhang. 2023. The theme-recipient alternation in Chinese: Tracking syntactic variation across seven centuries. *Corpus Linguistics and Linguistic Theory* 19(2). 207–235.
- Li, C. N. & S. A. Thompson. 1976. Subject and topic: A new typology of languages. In C. N. Li (ed.), *Subject and topic*, 457–489. New York: Academic Press.
- Li, C. N. & S. A. Thompson. 1981. *Mandarin Chinese: A functional reference grammar*. Berkeley: University of California Press.
- Liao, S., S. T. Gries & S. Wulff. 2024. Transfer five ways: Applications of multiple distinctive collexeme analysis to the dative alternation in Mandarin Chinese. *Corpus Linguistics and Linguistic Theory*. <https://doi.org/10.1515/clt-2024-0033>.
- Liu, L. & G. Zhu. 2025. Jìrán jù de xìngzhì jí guīlèi wèntí (The nature and classification of “jìrán” clauses). *Linguistic Sciences* 24(1). 8–20.
- MacWhinney, B., E. Bates & R. Kliegl. 1984. Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior* 23(2). 127–150.
- Pan, V. J. & B. Zhu. 2023. On the syntax of causal clauses in Mandarin Chinese. In Ł. Jędrzejowski & C. Fleczonek (eds.), *Micro- and macro-variation of causal clauses: Synchronic and diachronic insights*, 221–249. Amsterdam, Philadelphia: John Benjamins Publishing Company.
- Partington, A. 2004. “Utterly content in each other’s company”: Semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9(1). 131–156.
- Poplack, S. & S. Tagliamonte. 1999. The grammaticization of *going to* in (African American) English. *Language Variation and Change* 11(3). 315–342.
- Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.
- R Core Team. 2025. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>.
- Sankoff, G. & H. Cedergren. 1971. Some results of a sociolinguistic study of Montreal French. In R. Darnell (ed.), *Linguistic diversity in Canadian society*, 61–87. Edmonton: Linguistic Research.
- Schiffrin, D. 1985. Multiple constraints on discourse options: A quantitative analysis of causal sequences. *Discourse Processes* 8(3). 281–303.
- Song, Z. & H. Tao. 2009. A unified account of causal clause sequences in Mandarin Chinese and its implications. *Studies in Language* 33(1). 69–102.
- Stanford, J. N. 2016. A call for more diverse sources of data: Variationist approaches in non-English contexts. *Journal of Sociolinguistics* 20(4). 525–541.
- Su, H. & Y. Fu. 2023. Local grammar approaches to speech acts in Chinese: A case study of exemplification. *Journal of Pragmatics* 212. 44–57.
- Tai, J. H.-Y. 1975. On two functions of place adverbials in Mandarin Chinese. *Journal of Chinese Linguistics* 3(2/3). 154–179.
- Thompson, S. A., R. E. Longacre & S. J. J. Hwang. 2007. Adverbial clauses. In T. Shopen (ed.), *Language typology and syntactic description*, 237–300. Cambridge: Cambridge University Press.
- Torres Cacoullos, R. 1999. Construction frequency and reductive change: Diachronic and register variation in Spanish clitic climbing. *Language Variation and Change* 11(2). 143–170.

- Torres Cacoullos, R. & C. E. Travis. 2019. Variationist typology: Shared probabilistic constraints across (non-) null subject languages. *Linguistics* 57(3). 653–692.
- Tsai, M.-C. 1996. A discourse approach to causal sentences in Mandarin Chinese. In *Proceedings of the 11th Pacific Asia Conference on Language, Information, and Computation (PACLIC 11)*, 93–98. Seoul, Korea.
- Wang, L. 1984. *Wang Li Wenji Vol. 1: Zhongguo Yufa Lilun [Collected Works of Wang Li, Vol. 1: Theory of Chinese Grammar]*. Jinan: Shandong Jiaoyu Chubanshe.
- Wang, Y.-F. 1999. The information sequences of adverbial clauses in Mandarin Chinese conversation. *Journal of Chinese Linguistics* 27(2). 45–89.
- Wang, Y.-F. 2002. The preferred information sequences of adverbial linking in Mandarin Chinese discourse. *Text & Talk* 22(1). 141–172.
- Wang, X. 2020. Managing a suspended course of action: A multimodal study of *suoyi* ‘so’-prefaced utterances in Mandarin conversation. *Chinese Language and Discourse* 11(2). 306–334.
- Wasow, T. 1997. End-weight from the speaker’s perspective. *Journal of Psycholinguistic Research* 26(3). 347–361.
- Wasow, T. & J. Arnold. 2003. Post-verbal constituent ordering in English. In G. Rohdenburg & B. Mondorf (eds.), *Determinants of grammatical variation in English*, 119–154. Berlin, New York: Mouton de Gruyter.
- Weinreich, U., W. Labov & M. I. Herzog. 1968. Empirical foundations for a theory of language change. In W. Lehmann & Y. Malkiel (eds.), *Directions for historical linguistics*, 95–195. Austin, London: University of Texas Press.
- Xiao, H. 2020. Subjectivity, causality and connectives in Mandarin Chinese: Converging evidence from written, spoken and social media discourse. *LOT*.
- Xiao, H., F. Li, T. J. M. Sanders & W. P. M. S. Spooren. 2021. How subjective are Mandarin REASON connectives? A corpus study of spontaneous conversation, microblog and newspaper discourse. *Language and Linguistics* 22(1). 166–210.
- Xing, F. 2001. *Hanyu fuju yanjiu [A study of Chinese complex sentences]*. Beijing: The Commercial Press.
- Xing, F. 2003. *Hanyu yufa sanbai wen [Three hundred questions for Chinese syntax]*. Beijing: The Commercial Press.
- Xu, J., T. Dong, M. Sun, Z. Chen, F. Liu, B. Wang, Y. Wang, Y. Li, Y. Wang, B. Ma, Z. Liu, Y. Qian, Z. Zhu, L. Quan & J. Lu. 2022. *The BFSU DiSCUSS (diversified spoken Chinese uttered in social settings) Corpus*. Available at: <http://corpus.bfsu.edu.cn/DiSCUSS.zip>.
- Xu, J. & H. Kang. 2022. Salience-simplification strategy for markedness of causal subordinators: “Because” and “since” in argumentative essays. *Lingua* 272. Article 103256.
- Xu, M., F. Li & B. Szmrecsanyi. 2024. Modeling the locative alternation in Mandarin Chinese: A corpus-based study. *International Journal of Corpus Linguistics* 29(2). 258–285.
- Yao, S. 2007. Lianci “jieguo” yu “suoyi” shiyong chayi de jiliang fenxi [A quantitative analysis of the usage differences between the conjunctions “jieguo” and “suoyi”]. *Journal of Ningxia University* 29(6). 51–53+72.
- Young, L. W. 1994. *Crosstalk and culture in Sino-American communication*. Cambridge: Cambridge University Press.
- Zhang, X. & X. Xing. 2024. The windowing and shifting of attention of motion verbs in the path event frame: A corpus-based contrastive study between 去 (qù, go) and “go” in Chinese and English. *Open Journal of Modern Linguistics* 14(4). 619–634.