Xiaoyu Tian, Weiwei Zhang, Dirk Speelman

Lectal variation in Chinese analytic causative constructions: What trees can and cannot tell us

1 Introduction

Years after its quantitative turn (cf. Janda 2013), Cognitive Linguistics experienced substantial methodological developments, which make it a suitable framework for exploratory studies that tap into large datasets and take into account multiple language-internal and language-external factors to model language variation (e.g. Colleman 2010; Zhang, Speelman, and Geeraerts 2011; Levshina, Geeraerts, and Speelman 2013a; Bernaisch, Gries, and Mukherjee 2014; Röthlisberger, Grafmiller, and Szmrecsanyi 2017). In addition to traditional hypothesis-testing regression modeling, more advanced statistical tools such as tree-based methods become more widely used to cope with the problems typically found in corpus data, such as data sparsity and collinearity (e.g. Tagliamonte and Baayen 2012; Bernaisch, Gries, and Mukherjee 2014; Szmrecsanyi et al. 2016). Recently, scholars have noticed the shortcomings of tree-based methods and proposed to combine them with regression models to yield more robust and interpretable results (cf. Strobl, Malley, and Tutz 2009; Gries 2019). In line with these methodological developments, this study explores the near-synonymous Chinese causative constructions from a cross-variety perspective using conditional random forests, conditional inference trees and multinomial logistic regression analysis.

Acknowledgement: The authors are grateful to the Linguistic Data Consortium for providing the corpus of "*Tagged Chinese Gigaword 2.0*". This project was supported by a China Scholarship Council grant to the first author (grant No.202006900017) and a Marie Skłodowska-Curie grant to the second author (European Union's Horizon 2020 research and innovation programme, agreement No. 793920). The usual disclaimers apply.

Xiaoyu Tian, Department of Linguistics, University of Leuven & Institute of Linguistics, Shanghai International Studies University, e-mail: xiaoyu.tian@kuleuven.be
Weiwei Zhang, Department of Linguistics, University of Leuven & Institute of Linguistics, Shanghai International Studies University, e-mail: weiwei.zhang@kuleuven.be
Dirk Speelman, Department of Linguistics, University of Leuven, e-mail: dirk.speelman@kuleuven.be

https://doi.org/10.1515/9783110687279-006

According to Talmy (2000: Ch. 7), causation is a force-dynamic pattern that involves two main participants: the antagonist (labeled as the CAUSER in this study) and the agonist (labeled as the CAUSEE in this study). The causer instigates a causing event or state, which affects the causee, who brings about the caused event. Linguistic means to express causation are called causatives or causative constructions. Based on the formal differences between the expressions of the cause and the effect, Comrie (1981) made a three-way distinction of causative constructions: morphological causatives, lexical causatives and analytic causatives. In morphological causatives, "the causative is related to the non-causative predicate by [productive] morphological means" (Comrie 1981: 167). When the relation between the cause and the effect is "handled lexically, rather than by any productive process", lexical causatives are involved (Comrie 1981: 168). Analytic causatives refer to cases where "there are separate predicates expressing the notion of causation and the predicate of the effect" (Comrie 1981: 167). The current study focuses on the analytic causative constructions in Chinese.

Chinese analytic causative constructions involve several different markers, among which the most used ones in contemporary written Chinese are *shi*, *ling* and *rang* (Liu 2000; Niu 2007), as in example (1):

(1)	他	使/令/让	我	想起	了	一	个	朋友。[1]
	Ta	***shi/ling/rang***	*wo*	*xiangqi*	*le*	*yi*	*ge*	*pengyou.*
	He	make	me	think of	PST	one	CLF	friend
	'He makes me think of a friend.'							
	CAUS	CAUSATIVE	CAUSEE	EFFECTED				
	ER	MARKER		PREDICATE				

Although extensive research has been carried out on Chinese analytic causative constructions, only few studies have attempted to investigate the choice of causative markers and their lectal variation using corpus data and advanced statistical tools (cf. Liesenfeld, Liu, and Huang 2020; Tian and Zhang 2020). This paper aims to address this gap by answering the following two questions: (1) What are the syntactic and semantic factors that affect the alternation of analytic causative constructions with the markers of *shi*, *ling* and *rang*? (2) What is

1 Example (1) is created by the first author through introspection. Other examples of this chapter are from the "Tagged Chinese Gigaword Version 2.0". Some of them are rephrased and unimportant details are omitted to save space.

the extent to which varieties of Chinese differ in the choice of *shi*, *ling* and *rang* as the causative marker?

The structure of the paper is outlined as follows: Section 2 reviews the previous studies on analytic causative constructions; Section 3 introduces the data and the methods; in Section 4, we present the statistical results, the implications of which are then discussed in Section 5; Section 6 provides some concluding remarks.

2 Previous studies on analytic causative constructions

2.1 Analytic causative constructions: A cross-linguistic review

Languages that have analytic causative constructions usually possess several different causative markers or auxiliaries, such as *make*/*have*/*cause* in English, *doen*/*laten* in Dutch, and *shi*/*ling*/*rang* in Chinese. In favor of the "Principle of No Synonymy" (Goldberg 1995: 67), cognitive linguists speculate that there should exist semantic or usage differences between different causative markers and investigate the alternation of analytic causative constructions. These studies mainly centered on English (e.g. Stefanowitsch 2001; Gilquin 2010) and Dutch (e.g. Verhagen and Kemmer 1997; Stukker 2005; Speelman and Geeraerts 2009; Levshina 2011). Some research then tackled the problem of lectal variation (e.g. Belgium Dutch vs. Netherlandic Dutch, cf. Speelman and Geeraerts 2009, Levshina 2011, Levshina, Geeraerts, and Speelman 2013a) or cross-linguistic variation (e.g. English vs. Dutch, cf. Levshina, Geeraerts, and Speelman 2013b) in analytic causative constructions.

To explain the difference between Dutch causative verbs *doen* ("do") and *laten* ("let"), Verhagen and Kemmer (1997) proposed the "(in)direct causation hypothesis", which was further developed by Stukker (2005). According to this hypothesis, if the causer produces the effected event directly without any inference "downstream", then direct causation is involved and speakers tend to choose *doen*; if besides the causer, the causee is the most immediate source of energy in the effected event and has some degree of "autonomy" in the causal process, then indirect causation is involved, and speakers tend to use *laten* (Stukker 2005: 50).

In light of the "(in)direct causation hypothesis", many scholars conducted quantitative research to investigate the alternation of analytic causative constructions. For instance, Speelman and Geeraerts (2009) employed logistic regression analysis to evaluate the effect of several linguistic internal and external predictors

(e.g. the animacy of the causer, the coreferentiality between the causer and the causee, the transitivity of the effected predicate, the genre and variety, etc.) on the choice of *doen* and *laten*. The statistical results showed that the "(in)direct causation hypothesis" cannot fully explain the alternation between *doen* and *laten*. They also found significant lectal variation in the distribution of *doen* and *laten*, with *doen* appearing more frequently in Belgian Dutch than in Netherlandic Dutch.

Then, Levshina, Geeraerts, and Speelman (2013a) enhanced the multivariate approach by adding some new semantic variables to the model, including the semantic class of both the causer and the causee, the semantic class of the event expressed by the effected predicate (physical or mental), etc. The result supports the "(in)direct causation hypothesis" and lectal factors only display weak influences indirectly in this study.

The study of analytic causative construction alternation is not limited to Dutch. For instance, Levshina, Geeraerts, and Speelman (2013b) created a common conceptual space of analytic causatives in English and Dutch and found that different animacy configurations of the causer and the causee constitutes the most important two dimensions in the conceptual space. They also pointed out some commonalities in the semantics of causation between English and Dutch, although there are no strict cross-linguistic correspondences between the two languages.

2.2 Chinese analytic causative constructions with *shi*, *ling* and *rang*

Previous research on Chinese analytic causative constructions mainly focused on their differences from other syntactic means expressing causation (e.g. Fan 2000; Xiang 2002; Xiong 2004; Zhou 2004) or the grammaticalization of causative markers (e.g. Xu 2003; Zhang 2005; Niu 2007; Cao 2011). Only a few studies explored the onomasiological choice of causative markers (e.g. Yang 2016; Liesenfeld, Liu, and Huang 2020) and its lectal variation (Tian and Zhang 2020) based on corpus data.

Interesting results regarding the function and usage of *shi*, *ling* and *rang* have been found in previous research. First, *shi* and *ling* have lost the imperative meaning and most frequently serve as causative markers, whereas *rang* is still intensively used to denote a permissive meaning (Niu 2007). Second, *ling* tends to co-occur with the word *ren* (meaning "people" or "person") followed by a single verb and the construction [*ling ren* + result] exhibits a lexical fixation tendency, which is more likely to occur in a relative clause (Niu 2007; Zhang 2005). Third, *shi* tends to co-occur with verbs that indicate changes of results, while *rang* is more likely to be followed by verbs referring to activities and movements

(Yang 2016). In addition, scholars also found that the distributions of *shi*, *ling* and *rang* are sensitive to register differences: *Shi* and *ling* appear more often in written language, while *rang* tends to occur in spoken language (Wan 2004; Miyake 2005, Yang 2016).

Ni (2012) explained the alternation between *shi* and *rang* in Chinese analytic causative constructions with the "(in)direct causation hypothesis" proposed by Verhagen and Kemmer (1997). She suggests that *shi* is similar to *doen* in Dutch and involves direct causation, whereas *rang* is similar to *laten* and often expresses indirect causation. However, her study is based on 250 observations and she only explored four variables (i.e. the transitivity of the effected predicate, the animacy of the causer and the causee, modal verbs co-occurring with *shi* and *rang*, whether the causative construction is *wh*-cleft), therefore, her conclusions still need to be tested with more data and more robust statistical methods.

In general, Chinese analytic causative constructions are still understudied in terms of the following aspects. First of all, researchers have not systematically examined the influence of various syntactic and semantic features of construction components on the choice of markers. Secondly, previous studies barely considered that people from different varieties of Chinese might differ in their choices of causative markers, whereas the lectal variation has been attested in studies on Dutch causatives (Speelman and Geeraerts 2009; Levshina 2011; Levshina, Geeraerts, and Speelman 2013a). Thirdly, methodologically speaking, the traditional approaches in Chinese causative studies largely involve introspection or small-scale corpus-illustrated description, based on which it is difficult to make further theoretical interpretations.

Therefore, with regiolectally-balanced corpus data, this study adopts advanced statistical methods, viz. conditional random forests, conditional inference trees and multinomial logistic regression analysis, to disentangle the syntactic, semantic and lectal factors that influence the choice of Chinese causative markers *shi*, *ling* and *rang*.

3 Data and methods

3.1 Data resource and extraction

For this study, we tapped into the corpus of "Tagged Chinese Gigaword Version 2.0" (Huang 2009),[2] which contains more than 800 million words of

2 The website of the corpus: https://catalog.ldc.upenn.edu/LDC2009T14.

newswire texts covering Mainland Chinese, Taiwan Chinese and Singapore Chinese (see Table 5.1).

Table 5.1: Information of the corpus of "Tagged Chinese Gigaword Version 2.0".

Source of data	Year	Number of characters
Central News Agency (Taiwan Chinese)	1991–2004	501,456,000
Xinhua News Agency (Mainland Chinese)	1991–2004	311,660,000
Lianhe Zaobao Newspaper (Singapore)	2000–2003	18,632,000
Total	/	831,748,000

For practical reasons, we restricted ourselves to a subset of the corpus by selecting around 2 million words from each variety.[3] In total, we retrieved 12,385 observations containing *shi*, *ling* and *rang* using *Antconc* (Anthony 2019) and then manually excluded spurious hits. The data cleaning criteria are as follows:

i) Delete repeated occurrences.
ii) Delete occurrences where *shi*, *ling* and *rang* are used as morphemes, e.g., *da-shi* 'ambassador', *ming-ling* 'order', *rang-zuo* 'offer seat to', etc.
iii) Delete occurrences where *shi*, *ling* and *rang* denote non-causative meanings. For instance, *ling* sometimes expresses 'order', and *rang* can be used to denote a permissive or passive meaning.

The third criterion involves some difficulties when it comes to distinguishing causative meanings from permissive meanings expressed by *rang*. For the difficulty cases, we rely on the Force Dynamic theory proposed by Talmy (2000): the causative category applies when the force of the Antagonist[4] overcomes that of the Agonist,[5] leading to a resultant state or activity of that "is the opposite of [the Agonist's] intrinsic actional tendency" (Talmy 2000: 418) in another situation, where the Antagonist that has been affecting the Agonist is removed and thus allows the Agonist to manifest its intrinsic tendency, the permissive category applies. We use (2) and (3) as examples for illustration:

3 The corpus consists of news texts monthly from 1991 to 2004 for Mainland and Taiwan Chinese and from 2000 to 2003 for Singapore Chinese. We selected one file from each variety to make sure that they are comparable in both time and size. The texts of Mainland and Singapore Chinese are from September 2003 and that of Taiwan Chinese are from October 2003.

4 The Antagonist in causative constructions is labeled as the CAUSER in this study.

5 The Agonist in causative constructions is labeled as the CAUSEE in this study.

(2) 如果 让 菊花 按照 生产 规律
Ruguo ***rang*** *juhua* *anzhao* *shengchan* *guilv*
If let chrysanthemum follow growing regularity
开放, 市民 将 无法 在 十月
kaifang, *shimin* *jiang* *wufa* *zai shiyyuechu'* *shiyue*
blossom citizens will not be able to in October
初
chu
beginning
赏花。
shanghua.
admire flowers
'If we let the chrysanthemums blossom naturally, the citizens won't be able to admire them in early October.'

(3) 高科技 让 菊花 提前 盛开。
Gaokeji ***rang*** *juhua* *tiqian* *shengkai.*
The high technology make chrysanthemum in advance blossom
'The high technology makes the chrysanthemums blossom in advance.'

In (2), the chrysanthemums will manifest their intrinsic tendency and blossom naturally if there is no other force affecting this process, so the permissive category of Talmy (2000) applies and *rang* expresses a permissive meaning. In (3), however, the force of the high technology overcomes that of the chrysanthemums and causes them to blossom in advance against their intrinsic tendency, so in this context the causative category of Talmy (2000) applies and *rang* is used as a causative marker.

The cleaning procedures leave us with more than 10,000 observations, which still involves tremendous manual work for the variable coding. Therefore, we randomly selected 30% of the observations with *shi*, *ling* and *rang* respectively for the data annotation and analysis (see Table 5.2).

Table 5.2: Overview of the numbers of extracted and randomly selected occurrences.

	Mainland Chinese	Taiwan Chinese	Singapore Chinese	Total	Randomly Selected (30%)
shi	1425	659	1290	3374	1012
rang	669	1954	2598	5221	1566
ling	249	351	1022	1622	486
Total	2343	2964	4910	10217	**3064**

3.2 Variable annotation

The dependent variable of this study is a categorical one involving three different levels, i.e. *shi*, *ling* and *rang*. Based on the literature (e.g. Speelman and Geeraerts 2009; Levshina 2011; Niu 2014), the 3,064 observations were annotated with 27 independent variables (see Table 5.3 for an overview). All variables except for Variety were coded manually by the first author following the annotation scheme. To evaluate the reliability of the manual annotation, two coders independently annotated the 26 variables other than Variety for 100 randomly selected observations. We then calculated the kappa statistic (Carletta 1996) of the inter-rater agreement,[6] and the kappa *k* values range from 0.879 to 1, which indicates an excellent inter-rater reliability (Orwin 1994: 152).

Due to the lack of space, in the following text we only illustrate the seven variables that show significance in the random forest model (see Section 4.1). A complete annotation scheme with detailed explanations and examples is provided at https://osf.io/342re/.

– PredSynt: it refers to the syntactic form of the effected predicate and has five levels of *tr* (transitive verb, cf. (4a)), *intr* (intransitive verb, cf. (4b)), *copula* (cf. (4c)), *adj* (adjective, cf. (4d)),[7] and *idiom* (i.e. a fixed expression, cf. (4e)).

(4) a. 我 觉得 有必要 让 他 了解 真相。
Wo juede youbiyao rang ta liaojie zhenxiang.
I think necessary let him know truth
'I think it is necessary to let him know the truth.'
(PredSynt = *tr*)

b. 新 技术 可 使 利润 提高。
Xin jishu ke shi lirun tigao.
New technology can make profit increase
'The new technology can make the profit increase.'
(PredSynt = *intr*)

6 This procedure is implemented using the {irr} package (Gamer et al. 2019) in R.

7 It is worth noting that Chinese analytic causative constructions allow the effected predicate to be a bare adjective. In other words, the three causative markers (i.e. *shi*, *ling* and *rang*) are interchangeable when a bare adjective serves as the effected predicate. It is different from English, where one can only use *make*, while other causative markers such as *let* or *have* do not work for a bare adjective effected predicate.

c. 她 使 中国队 成为 了 冠军。
Ta shi zhongguodui chengwei le guanjun
She make China Team become PST champion
'She made the China Team become the champion.'
(PredSynt = *copula*)

d. 这 使 我 难过。
Zhe shi wo nanguo.
This make me sad
'This makes me sad.'
(PredSynt = *adj*)

e. 这个 新闻 令 人 大跌眼镜。
Zhege xinwen ling ren dadieyanjing
This news make people drop glasses
'This news is extremely surprising.'
(PredSynt = *idiom*)

- Variety: it stands for language varieties. It was encoded automatically and has three possible values: *ml* (Mainland Chinese), *tw* (Taiwan Chinese), and *sg* (Singapore Chinese).
- CeSynt: it stands for the syntactic form of the causee. We assigned four possible values for this variable: *np* (noun phrase, cf. (5a)), *pron* (pronoun, cf. (5b)), *cl* (clause, cf. (5c)) and *ren* ("people/person" in Chinese, cf. (5d)). For practical reasons, *ren* is coded as a separate value since the previous studies (e.g. Wan 2004; Zhang 2005; Niu 2007) detected that *ling* is strongly collocated with *ren*. We expect that *ling* should be favored when the syntactic form of the causee is assigned the value of *ren*.

(5) a. 我 要 让 社区 更 美丽。
Wo yao rang shequ geng meili.
I want to make community more beautiful
'I want to make the community more beautiful.'
(CeSynt = *np*)

b. 他 让 我 失望。
Ta rang wo shiwang.
He make me disappoint
'He makes me disappointed.'
(CeSynt = *pron*)

c. 大雨 使 按期 完工 更 困难。
Dayu *shi* *anqi* *wangong* *geng* *kunnan.*
Intensive rainfall make on time complete project more difficult
'The intensive rainfall makes it more difficult to complete the project on time'
(CeSynt = *cl*)

d. 她的 邀请 令 人 无法 抗拒。
Tade *yaoqing* *ling* *ren* *wufa* *kangju.*
He invitation make people no way to resist
'Her invitation is irresistible.'
(CeSynt = *ren*)

- CsedProsody: this variable deals with the prosody of the caused event, i.e. the event expressed by the effected predicate. It has three possible values: *neg* (negative, cf. (5b)), *ntrl* (neutral, cf. (5d)) and *pstv* (positive, cf. (5a)).
- ClauseType: it refers to the clause type where the causative construction is found. We assigned five possible values to this variable: *avb* (adverbial clause, cf. (6a)), *cpl* (complemental clause, cf. (6b)), *cpd* (compound sentence, cf. (6c)), *main* (cf. (6d)), *rltv* (relative clause, cf. (6e)) and *smpl* (simple sentence, cf. (5d)). We expect that *ling* has a higher probability of occurring in relative clause as is observed in the literature (Niu 2007, Zhang 2005).

(6) a. 为 使 父母 高兴, 他 努力 学习。
Wei *shi* *fumu* *gaoxing,* *ta* *nuli* *xuexi.*
To make parents happy he hard study
'In order to make his parents happy, he studies hard.'
(ClauseType = *avb*)

b. 这个 广告 宣称 可 使 孩子 长高。
Zhege *guanggao* *xuancheng* *ke* *shi* *haizi* *zhanggao.*
This advertisement claim can make children grow taller
'This advertisement claims that it can make children grow taller.'
(ClauseType = *cpl*)

c. 网络 不仅 使 生活 方便, 也 提供
Wangluo *bujin* *shi* *shenghuo* *fangbian,* *ye* *tigong*
Internet not only make life convenient also provide
信息。
xinxi.
infomation

'The Internet not only makes life convenient, but also provides information.'
(ClauseType = *cpd*)

d.

由于	观众	很多,	使	场面	多次	失控。
Youyu	*guanzhong*	*henduo,*	*shi*	*changmian*	*duoci*	*shikong*
Because	audience	a lot	make	situation	many times	out of control

'Because there is a big audience, it makes the situation out of control several times.'
(ClauseType = *main*)

e.

昨天	真	是	令	人	难忘	的	一	天。
Zuotian	*zhen*	*shi*	*ling*	*ren*	*nanwang*	*de*	*yi*	*tian.*
Yesterday	truly	is	make	people	hard to forget	REL	one	day

'Yesterday truly was a memorable day.'
(ClauseType = *rltv*)

– PredSem: it stands for the semantic class of the effected predicate. There are three possible values for this variable: *atelic* (cf. (6b)), *telic* (cf. (7)), and *state* (cf. (6e)). The previous research shows that *ling* tends to co-occur with a stative predicate (Zhang 2005).

(7)

这	微笑	使	人	失去	了	判断力。
Zhe	*weixiao*	*shi*	*ren*	*shiqu*	*le*	*panduanli*
The	smile	make	people	lose	PST	judgement

'The smile made people lose their judgement.'
(PredSem = *telic*)

– CsedSemT: this variable refers to the semantic class of the target domain (i.e. the figurative meaning) of the caused event, as opposed to the source domain (i.e. the literal meaning) when metaphors are involved. It has three possible values: *ment* (mental caused event, cf. (6e)), *phy* (physical caused event, cf. (6b)) and *social* (social caused event, cf. (6d)).

Table 5.3: Independent variables.

Label	Predictor	Value
CrLocus	Locus of the causer	*adjacent, adjacent2, distant, implicit*
CrSynt	Syntactic form of the causer	*cl* (clause), *na, np* (noun phrase), *pron* (pronoun), *vp* (verbal phrase)
CrSem	Semantic class of the causer	*anim* (animate), *event, inanim* (inanimate), *na*
CrPers	Person of the causer	*1, 2, 3, na*
CrDef	Definiteness of the causer	*def* (definite), *indef* (indefinite), *na*
CrIntent	Intentionality of the causer	*intent* (intentional), *unintent* (unintentional)
CeSynt	Syntactic form of the causee	*cl, np, pron, ren* (meaning "people/person")
CeSem	Semantic class of the causee	*anim, event, inanim*
CePers	Person of the causee	*1, 2, 3, na*
CeDef	Definiteness of the causee	*def, indef*
CeControl	Whether the causee can control the caused event	*no, yes*
CeRole	Thematic role of the causee	*agent, befry* (beneficiary), *expcer* (experiencer), *patient, theme,*
CoRef	Coreference between the causer and the causee	*no, yes*
CseModality	Modal verb modifying the causative marker	*ability, incl* (inclination), nece (necessity), *none, poss* (possibility)
CseAdv	Adverb modifying the causative marker	*degree, none, oth* (other), *range, time*
CseNeg	Polarity of the causative marker	*no, yes*
PredSynt	Syntactic form of the effected predicate's	*adj* (adjective), *copula, idiom, intrans* (intransitive verb), *trans* (transitive verb)
PredSem	Semantic class of the effected predicate	*atelic, state, telic*
CsedProsody	Prosody of the caused event	*neg* (negative), *ntrl* (neutral), *pstv* (positive)
CsedModality	Modal verb modifying the effected predicate	*ability, incl* (inclination), nece (necessity), *none, poss* (possibility)

Table 5.3 (continued)

Label	Predictor	Value
CsedAdv	Adverb modifying the effected predicate	*degree, manner, none, range, time*
CsedNeg	Polarity of the effected predicate	*no, yes*
CsedSemS	Semantic class of caused event (source domain)	*ment* (mental), *phy* (physical), *social*
CsedSemT	Semantic class of caused event (target domain)	*ment* (mental), *phy* (physical), *social*
ClauseType	Clause type where the causative construction is found	*avb* (adverbial), *cpl* (complemental), *cpd* (compound), *main, rltv* (relative), *smpl* (simple)
Structure	Number of the effected predicate	*mul* (multiple), *sg* (single)
Variety	Language variety	*ml* (Mainland), *sg* (Singapore), *tw* (Taiwan)

3.3 Statistical analyses

The statistical analyses of this study are conducted in R software (R Core Team 2014) and the code can be found in the Appendix of this chapter. Given that our study looks into an extensive set of predictor variables, which moreover may interact in potentially ways, we opted not to directly feed all predictor variables into a massive logistic regression model. Instead, we chose to first try and identify which are the most important predictors by means of a conditional random forest analysis, an approach which has gained popularity in recent multifactorial studies (e.g. Tagliamonte and Baayen 2012; Bernaisch, Gries, and Mukherjee 2014; Szmrecsanyi et al. 2016; Deshors and Gries 2016). Compared to logistic regression models, tree-based methods such as conditional inference trees and conditional random forests are less affected by predictor collinearity and data sparsity, which is typically exhibited in corpus data (cf. Gries 2019). Moreover, tree-based methods can process a large number of variables simultaneously and can handle high-order interactions well (Strobl, Malley, and Tutz 2009), which suits the characteristics of our data.

We first built a conditional random forest model with all the 27 independent variables using the {party} package in R (Hothorn, Hornik and Zeileis 2006). A conditional random forest is an ensemble of multiple conditional inference trees.

A conditional inference tree partitions the data into two subsets based on whichever predictors that co-vary most strongly with the responses and recursively repeats this process, each time picking the predictor that works best at that point in the tree, until no further split can be made to significantly improve the classification accuracy. Random forest models add two layers of randomness in this procedure to reduce the variability of prediction: first, each tree in the forest is grown based on randomly bootstrapped data; second, each split of each tree is decided based on a randomly selected subset of the predictors (cf. Gries 2019). The researcher needs to specify two parameters when creating a random forest model: the number of trees that the forest grows (*ntree*) and the number of variables considered when making each split of each tree (*mtry*). By aggregating predictions of all the conditional inference trees, a random forest model assigns a variable importance score to each predictor[8](as in Figure 5.1).

After having the variable importance of each predictor, our next step was to build a single conditional inference tree, again using the {party} package, this time only using the variables that showed significant effects in the random forest. Previous studies using tree-based methods (e.g. Bernaisch, Gries, and Mukherjee 2014; Szmrecsanyi et al. 2016) usually start by building a conditional inference tree with all the variables in the first step and then create a random forest, using the conditional inference tree as a first tentative, but possibly fragile model, and using the forest for a more robust assessment of the variable importance of the predictors. Indeed, individual conditional inference trees are known to be easily affected by the problem of instability because each split of the tree is made based on the previous splits and a small change in the data may alter the whole structure (Strobl Malley, and Tutz 2009; Kuhn and Johnson 2016: 174; Levshina, in press). We acknowledge this vulnerability, and by no means consider the conditional inference tree we build in the second step to be superior to the random forest we build in the first step. It merely helps interpret the random forest, because it visualizes how exactly the important predictors affect the alternation and how they interact, which is information that is hidden in the forest analysis. We accept, however, that the conditional inference tree analysis remains potentially less robust than the random forest analysis. We do believe, though, that building the tree with just the predictors deemed important by the forest can somewhat reduce undesired variability of the results of the tree analysis.

Both tree-based methods described above are powerful tools. Unfortunately, however, both have their limitations. On the one hand, conditional random

8 There are different kinds of variable importance measures, e.g. "Gini importance" and "permutation accuracy importance". For more details, see Strobl, Malley, and Tutz (2009).

forest analysis is a "black-box" method whose result is hard to interpret (Strobl, Malley, and Tutz 2009). It shows us which variables are important, but not how exactly they are important. On the other hand, an individual conditional inference tree might yield less robust results and, moreover, also may keep some of the global patterns in the data hidden, because of the way each node in the tree is the result of a 'winner takes it all' kind of procedure. More specifically, at each point in the tree, only the single most important pattern in that part of the tree is highlighted, potentially hiding other important patterns.

Therefore, our third and final step is to complement the tree-based methods with logistic regression modelling, as is suggested and implemented in several studies (e.g. Strobl, Malley, and Tutz 2009; Gries 2019; Deshors and Gries 2020). This procedure serves two purposes: the first one is to verify whether the results of tree-based methods and of logistic regression modelling are consistent; the second one is to further investigate the interactions between Variety and other variables. The latter goal is driven by our research interest, which lies in the lectal variation. Given that the dependent variable of our study is a categorical one that includes three values, i.e., *shi*, *ling* and *rang*, we built a multinomial logistic regression model.

Multinomial logistic regression models can be built in R software using any of the packages {mlogit} (McFadden 1973, 1974; Train 2009), {polytomous} (Arppe 2008, 2009) or {nnet} (Ripley 1996; Venables and Ripley 2002) (cf. Levshina 2015: Section 13 for the differences between these packages). We opted for the {nnet} package in this study. We started out with a model with the seven significant variables that are significant in the random forest model as well as the two-way interactions between Variety and the other six variables and used a backward model selection procedure to obtain a final model, which is presented in Section 4.3.

The annotated 3064 observations were randomly divided into a training dataset (70%) and a test dataset (30%). All the three models were created based on the training dataset and then were used to predict the responses in the 30% test dataset to evaluate their performance.

4 Results

4.1 Conditional random forests

We created a random forest model with 1000 inference trees (ntree =1000) and each split of each tree is made based on five randomly selected variables

(mtry = 5). The model yields importance scores for all the variables, as is shown in Figure 5.1.[9]

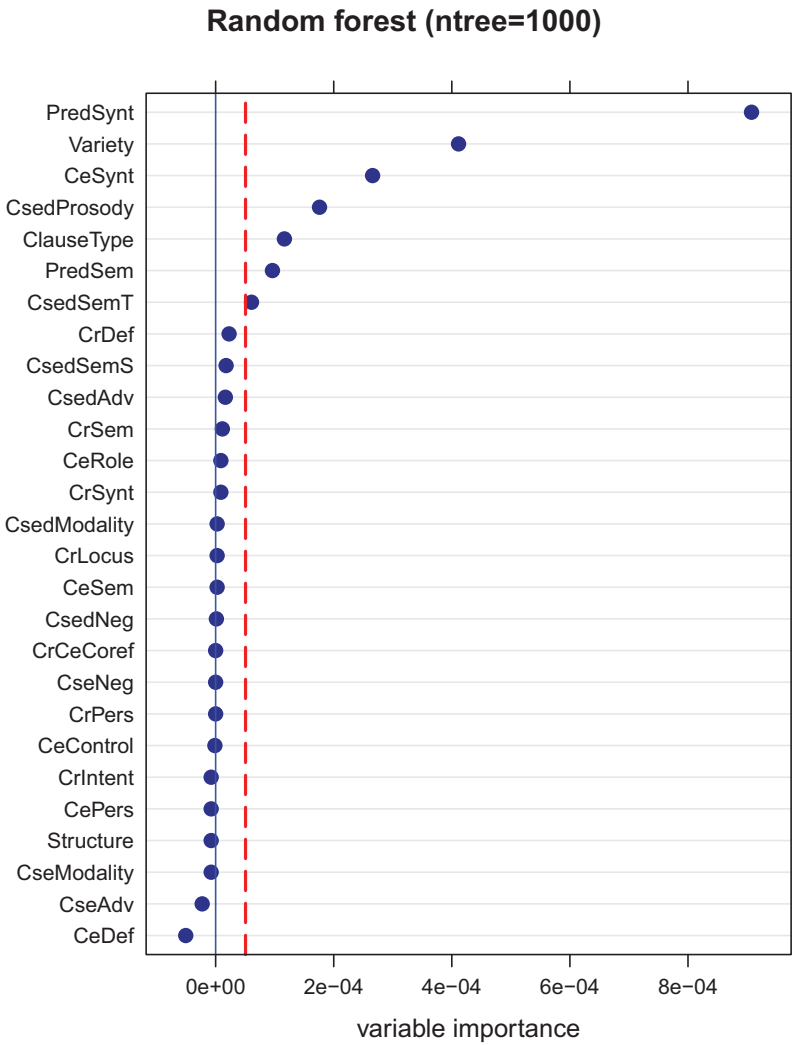

Figure 5.1: Importance measure of all variables.

9 We built five random forest models by using different seeds when splitting the data into the training dataset and the test dataset. In all these five models, PredSynt, Variety, CeSynt, CsedProsody and ClauseType are always significant, which means that the results of the random forest models are highly stable and reliable.

According to the random forest model, the syntactic form of the effected predicate (PredSynt) is the most important variable, followed by Variety, the syntactic form of the causee (CeSynt), the prosody of the caused event (CsedProsody), the type of the clause in which the causative construction is found (ClauseType), the semantic class of the effected predicate (PredSem) and the semantic class of the caused event (target domain) (CsedSemT). The other variables do not have a significant effect on the choice of *shi*, *ling* and *rang*.

The prediction accuracy of the random forest on the 30% test dataset is 71.43% (Table 5.4), which is better than the accuracy rate of always choosing the most frequent marker (i.e. *rang*) (51.15%) and much better compared to 33.33%, the correct chance if the responses are chosen randomly.

Table 5.4: Confusion matrix of the conditional random forest model on 30% test dataset.

		Observed		
Predicted		*ling*	*rang*	*shi*
	ling	82	47	16
	rang	17	381	71
	shi	1	110	192

The importance scores assigned to the variables by the random forest model are hard to interpret. For instance, we know from Figure 5.1 that PredSynt is the most important variable, but the model does not show how different values of this variable affect the probability of choosing *shi*, *rang* or *ling*. Therefore, we complement this method with a conditional inference tree (Section 4.2) and a multinomial logistic regression model (Section 4.3).

4.2 Conditional inference trees

We built a conditional inference tree with the seven variables that turned out to be significant in the random forest model. The result is shown in Figure 5.2.

We can see from the model that the alternation of *shi*, *ling* and *rang* involves a complex interplay of different variables. The first split is made based on the syntactic form of the causee (CeSynt), where the most important reason of that split seems to be that *shi* is hardly ever used when the causee is expressed by the word *ren* (CeSynt = *ren*). Whether in the latter situation the chosen causative is *rang* or

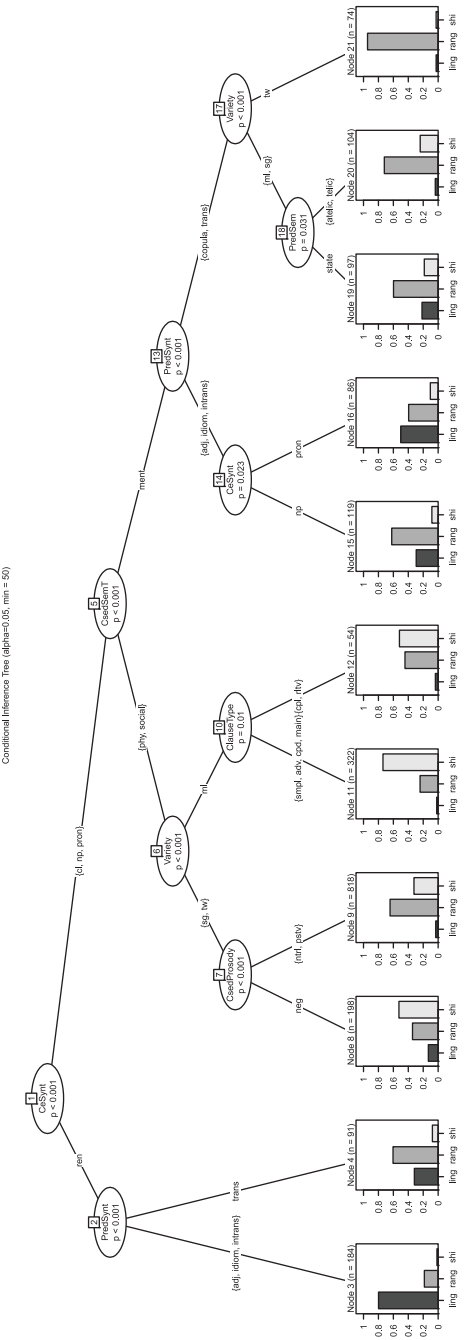

Figure 5.2: Result of the conditional inference tree. Splits are made if $p < 0.05$. Minimal number of observations in terminal nodes equaled 50.

ling, turns out, according to the model, to correlate strongly with the syntactic form of the effected predicate (PredSynt, cf. Node 2). More specifically, if the effected predicate is an idiom, an adjective or an intransitive verb, *ling* is strongly favored (Node 3), as in (8). If it is a transitive verb, the proportion of *rang* exceeds that of *ling* (Node 4).

(8) 这个 奇迹 令 人 叹为观止。
Zhege qiji ling ren tanweiguanzhi.
This miracle make people knock one's socks off
'This miracle knocked people's socks off.'
(CeSynt = *ren*; PredSynt = *idiom*)

When the causee is expressed by other forms than *ren*, the interactions of variables are much more complex, as is shown in the right side of the first split. The first variable coming into play in that part of the tree is the semantic class of the caused event (target domain) (CsedSemT, Node 5), which divides the local subset of the data into a subset to the left (CsedSemT = *physical* or social) with relatively more *shi* and less *ling*, compared to the second branch, and a subset to the right (CsedSemT = *mental*), with relatively more *ling* and less *shi*, compared to the first branch.

If a physical or social event is expressed by the effected predicate, Variety comes into play, where Mainland Chinese (the second branch of Node 6) shows a stronger preference for *shi* than Singapore and Taiwan Chinese (the first branch of Node 6). In the latter two varieties, the proportion of *shi* only slightly exceeds that of *rang* when the event is negative (Node 8) while *rang* is predominant if the event is positive or neutral (Node 9). In Mainland Chinese, on the other hand, *shi* always takes up the highest proportion, although its advantage over *rang* lessens when the causative construction is in a complemental or a relative clause (Node 12).

The picture is more complex when the effected predicate expresses a mental event, where the syntactic form of the effected predicate (PredSynt) shows an effect again (Node 13). In the first branch of Node 13, *ling* shows again its preference for an effected predicate expressed by an idiom, an adjective or an intransitive verb, and it becomes the most favored marker if the causee is a pronoun (Node 16).

In the second branch of Node 13, i.e., when the syntactic form of the effected predicate is a copula or a transitive verb, *rang* is always the most favored marker, although the preferences of the three varieties differ again. In Taiwan Chinese, *rang* is predominant and the other two markers are barely used (Node 21), whereas in Mainland and Singapore Chinese, the proportion of *ling* notably increases when the semantic class of the effected predicate is stative (Node 19).

The prediction accuracy of the conditional inference tree on the test dataset is 65.21% (Table 5.5), which is better than the accuracy rate of always choosing *rang* (51.15%), the most frequent marker and much better compared to 33.33%, the correct chance if the responses are chosen randomly.

Table 5.5: Confusion matrix of the condition inference tree model on 30% test data.

		Observed		
Predicted		*ling*	*rang*	*shi*
	ling	77	48	20
	rang	36	361	72
	shi	3	140	160

4.3 Multinomial logistic regression

Then we fitted a multinomial logistic regression model with the seven variables that are significant in the random forest model as well as the two-way interactions between Variety and the other six variables. We started out with a model with all these predictions and interactions and continued by removing non-significant terms, each time removing the term with the highest *p*-value, until all remaining terms were significant at an alpha-level of 0.01.[10] The final model is:

> Item ~ PredSynt + Variety + CeSynt + CsedProsody + ClauseType + CsedSemT + Variety: CsedSemT[11]

The reference value of the dependent variable is *ling*. An odds ratio greater than 1 indicates that the probability of *rang* or *shi* increases compared with *ling*. An odds ratio smaller than 1 indicates the opposite, viz. a decrease of the probability of *rang* or *shi*, compared to *ling*.

10 This procedure is implemented with the function Anova() in the R package {car}. The output of the model is created using the function tab_model() in the R package {sjPlot}.

11 We noticed that there are only seven observations with the value CeSynt = *cl*. In order to avoid the effect of data sparsity, we removed the seven observations from the data and built the multinomial logistic regression model.

Table 5.6: Summary of the multinomial logistic regression model.

Predictors	*Odds Ratios*	*CI*	*p*	*Response*
(Intercept)	1.43	0.50 – 4.08	0.504	*rang*
PredSynt [copula]	2.87	0.63 – 13.01	0.172	*rang*
PredSynt [idiom]	0.87	0.49 – 1.54	0.634	*rang*
PredSynt [intrans]	0.94	0.57 – 1.55	0.818	*rang*
PredSynt [trans]	3.69	2.27 – 6.00	**<0.001**	*rang*
Variety [sg]	0.74	0.41 – 1.36	0.335	*rang*
Variety [tw]	0.97	0.49 – 1.93	0.925	*rang*
CeSynt [pron]	1.04	0.68 – 1.61	0.853	*rang*
CeSynt [ren]	0.16	0.10 – 0.24	**<0.001**	*rang*
CsedProsody [ntrl]	3.43	2.31 – 5.09	**<0.001**	*rang*
CsedProsody [pstv]	3.00	2.06 – 4.38	**<0.001**	*rang*
ClauseType [compl]	4.98	1.07 – 23.15	0.041	*rang*
ClauseType [cpd]	1.53	0.44 – 5.30	0.503	*rang*
ClauseType [main]	1.46	0.42 – 5.10	0.551	*rang*
ClauseType [rltv]	0.19	0.07 – 0.49	**0.001**	*rang*
ClauseType [simple]	0.69	0.29 – 1.62	0.396	*rang*
CsedSemT [phy]	2.93	0.60 – 14.17	0.182	*rang*
CsedSemT [social]	1.19	0.49 – 2.87	0.703	*rang*
Variety [sg] * CsedSemT[phy]	0.81	0.15 – 4.39	0.805	*rang*
Variety [tw] * CsedSemT[phy]	2.55	0.36 – 18.02	0.349	*rang*
Variety [sg] * CsedSemT[social]	2.16	0.81 – 5.77	0.126	*rang*
Variety [tw] * CsedSemT[social]	4.15	1.34 – 12.87	0.014	*rang*
(Intercept)	1.96	0.63 – 6.04	0.244	*shi*
PredSynt [copula]	5.96	1.31 – 27.15	0.021	*shi*
PredSynt [idiom]	0.58	0.30 – 1.15	0.120	*shi*
PredSynt [intrans]	0.75	0.42 – 1.34	0.334	*shi*
PredSynt [trans]	2.27	1.29 – 3.99	**0.004**	*shi*
Variety [sg]	0.42	0.20 – 0.87	0.020	*shi*

Table 5.6 (continued)

Predictors	*Odds Ratios*	*CI*	*p*	*Response*
Variety [tw]	0.09	0.02 – 0.29	**<0.001**	*shi*
CeSynt [pron]	0.95	0.59 – 1.53	0.833	*shi*
CeSynt [ren]	0.08	0.04 – 0.15	**<0.001**	*shi*
CsedProsody [ntrl]	1.36	0.88 – 2.12	0.167	*shi*
CsedProsody [pstv]	1.60	1.06 – 2.41	0.026	*shi*
ClauseType [compl]	2.93	0.61 – 14.10	0.179	*shi*
ClauseType [cpd]	1.29	0.36 – 4.72	0.695	*shi*
ClauseType [main]	1.65	0.46 – 5.96	0.441	*shi*
ClauseType [rltv]	0.18	0.06 – 0.53	**0.002**	*shi*
ClauseType [simple]	0.58	0.24 – 1.40	0.225	*shi*
CsedSemT [phy]	12.92	2.65 – 63.13	**0.002**	*shi*
CsedSemT [social]	8.66	3.45 – 21.74	**<0.001**	*shi*
Variety [sg] * CsedSemT[phy]	0.47	0.08 – 2.65	0.394	*shi*
Variety [tw] * CsedSemT[phy]	5.80	0.64 – 52.59	0.118	*shi*
Variety [sg] * CsedSemT[social]	0.73	0.25 – 2.08	0.554	*shi*
Variety [tw] * CsedSemT[social]	7.40	1.63 – 33.58	**0.009**	*shi*
Observations	2141			
R^2 Nagelkerke	0.476			

According to the model, when the syntactic form of the effected predicate is a transitive verb (PredSynt = *trans*), the probabilities of both *rang* and *shi* significantly increase ($p < 0.01$). When the causee is expressed by the word *ren* (CeSynt = *ren*), the probabilities of both *rang* and *shi* significantly decrease ($p < 0.001$). Another variable configuration that disfavors *rang* ($p = 0.001$) and *shi* ($p = 0.002$) is when ClauseType = *rltv* (relative), i.e., when the causative construction occurs in a relative clause. When the effected predicate expresses neutral or positive events, the probability of *rang* will significantly increase ($p < 0.001$). The probability of using *shi* significantly increase when the effected predicate expresses physical ($p = 0.002$) or social events ($p < 0.001$). These main effects in the multinomial logistic regression model corroborate the result of the conditional inference tree.

The interaction between Variety and CsedSemT is illustrated in Figure 5.3.[12]

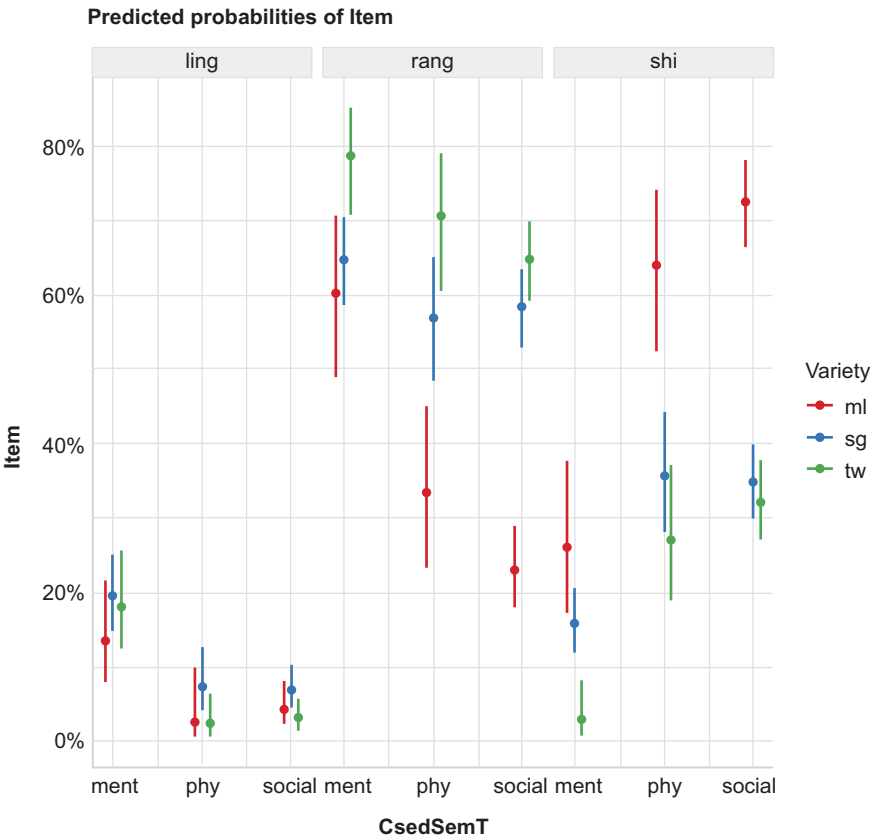

Figure 5.3: The interaction plot between Variety and CsedSemT.

Figure 5.3 presents the probabilities of *ling* (left side), *rang* (middle) and *shi* (right side) in Mainland Chinese (red), Singapore Chinese (blue) and Taiwan Chinese (green) with different values of CsedSemT, i.e. the semantic class of the caused event (target domain).

We can see from the plot that the probability of *shi* is always higher in Mainland Chinese than in the other two varieties, whereas Singapore Chinese and especially Taiwan Chinese prefer to choose *rang*, irrespective of the values of CsedSemT. The probability of *ling* remains low in all three varieties, with a

12 The plot is generated with the functions ggeffect() in the R package {ggeffects}and plot().

notable increase when the semantic class of the caused event (target domain) is a mental activity.

The variability of the probability of *shi* and *rang* with different values of CsedSemT is bigger in Mainland Chinese than in Singapore Chinese and Taiwan Chinese. Although the probabilities differ, the effect of CsedSemT is similar in Mainland Chinese and Taiwan Chinese. More specifically speaking, in Mainland Chinese and Taiwan Chinese, the probability of *shi* is the lowest when it is followed by mental caused event (target domain) and the highest when it is followed by social caused event (target domain), and the probability of *rang* is exactly the opposite. In Singapore Chinese, on the other hand, the probabilities of *shi* are nearly the same when it is with social or physical caused event (target domain), and this finding also holds for *rang*.

The prediction accuracy of the multinomial logistical regression in 30% test dataset is 67.14%. The *C* values of the model when predicting *ling*, *rang* and *shi* are 0.92, 0.75 and 0.81 respectively, indicating a good predictive power (Hosmer and Lemeshow 2000: 162).

5 Discussion

With the help of the advanced statistical methods, viz. conditional random forests, conditional inference trees and multinomial logistic regression analysis, this study manages to simultaneously investigate multiple language-internal and language-external factors based on corpus data and achieve a more realistic model of the variation in Chinese analytic causative construction alternation. The model results unveil the complex interplay between the syntactic, semantic and lectal factors that affect the choice of Chinese causative markers *shi*, *ling* and *rang*, yielding some important findings that speak to the research questions laid out in Section 1.

(1). What are the syntactic and semantic factors that affect the choice of *shi*, *ling* and *rang*?

Based on bottom-up data analytics, this study provides objective and verifiable evidence which confirms some previous findings regarding the use of *shi*, *ling* and *rang* while providing some new insights.

Both conditional inference tree and multinomial logistic regression analyses confirm that the [*ling ren* + result] construction, where *ling* and *ren* form a fixed expression while the result is expressed by an effected predicate in form of an adjective or an intransitive verb, has become the most prevalent usage for *ling* (Figure 5.2, Node 3). Given the low frequency of *ling* occurring in other

contexts (the terminal nodes other than Node 3 in Figure 5.2) and the low probability of using *ling* in general (Figure 5.3), we concur with Niu (2007) in that *ling* is losing its status of functioning as a causative marker and is on the process of becoming a morpheme in the fixed expressions of [*ling ren* + adj./intransitive verb]. This tendency of *ling* also explains why it is significantly more likely to occur in a relative clause (see Table 5.6), which is normally shorter than other clause types in Chinese and has a similar function with adjectives in terms of modifying a noun. Therefore, the fixed expressions of [*ling ren* + adj./intransitive verb] are more ideal choices than other longer and more complicated causative constructions. In addition, the *ling*-construction tends to express a mental event instead of a physical or social event (see Figure 5.3), which supports the findings on the usage of *ling* in Zhang (2005).

While studies on Chinese analytic causative constructions usually compare *shi* with *rang* (e.g. Hu 2002; Chen 2005; Ni 2012; Yang 2016), our models show that *shi* tends to occur in a different context with *ling* rather than *rang*. For instance, the frequency of *shi* is extremely low when *ling* is the most favored marker (Node 3 in Figure 5.2) and vice versa (Node 11 in Figure 5.2). As *shi* and *ling* grammaticalized into causative markers in similar historical periods and their grammaticalization occurred much earlier than *rang* (cf. Xu 2003; Cao 2011), we speculate that there may exist a competitive relationship between *shi* and *ling* during their grammaticalization processes, which caused the division of labor between these two markers. Of course, more diachronic research is required to verify this speculation.

Rang, on the other hand, as a much younger and more versatile marker, covers more usage contexts, as is shown by the relatively high proportions of *rang* in many terminal nodes of the conditional inference tree (see Figure 5.2). However, this does not mean that *rang* is equally distributed in all the contexts. For instance, *rang* shows a strong preference for effected predicates that are transitive verbs (see Figure 5.2 Node 4, 19–21 & Table 5.6). This finding confirms Ni (2012)'s assessment that *rang* has some similarity with the Dutch causative marker of *laten* (Speelman & Geeraerts 2009; Levshina, Geeraerts, and Speelman 2013a), which also favors transitive effected predicates.

As for Ni (2012)'s proposition of accounting for the distribution of *shi* and *rang* with the "(in)direct causation hypothesis" (Verhagen and Kemmer 1997), our study, by investigating more variables, calls for caution to reach such a conclusion. Firstly, the "(in)direct causation hypothesis" distinguishes direct and indirect causation by identifying whether the causer or the causee constitutes the source of energy over the whole causation process (cf. Verhagen and Kemmer 1997 and Section 2.1 of this paper), which assumes that different syntactic and semantic configurations of the causer and the causee should be the

most important factors. However, our models suggest that when taking multiple variables into account, the features of the causer and the semantic features of the causee do not stand out as the most influential ones affecting the choice of *shi* and *rang*, which means the "(in)direct causation hypothesis" may not be a suitable theory to explain the distribution of these two markers. Secondly, as is discussed above, the competition between *shi* and *ling* in different contexts is more intense than that between *shi* and *rang*, and the difference between *shi* and *rang* can be attributed to language external factors, such as register (cf. Wan 2004; Miyake 2005, Yang 2016) and language variety (see below for a detailed discussion).

(2). What is the extent to which varieties of Chinese differ in the choice of analytic causative constructions?

We explored the regional variation of Chinese analytic causative constructions by incorporating the variable of Variety into our models. The random forest model determines whether Variety stands out in the competition of all factors affecting the choice of *shi*, *ling* and *rang*, while the conditional inference tree and the multinomial logistic regression model can provide more information by showing the interactions between Variety and the language-internal factors.

All three models presented in Section 4 point to significant lectal differences in Chinese analytic causative construction alternation. More specifically, Variety ranks the second most important variable in the random forest model; in the conditional inference tree analysis, Variety manifests complex interactions with other variables by showing up twice in the splits; the multinomial logistic regression model also reveals a significant interaction between language varieties and the semantic class of the caused event.

A closer look at the results shows that the lectal variation mainly lies in the choices of *shi* and *rang*, whereas the frequency of *ling* remains rather low in all the three varieties and is mainly affected by the language-internal factors. According to the conditional inference tree and the multinomial logistic regression analysis, Mainland Chinese favors *shi* while Singapore and especially Taiwan Chinese favor *rang* (see Figure 5.2 & 5.3).

There are two possible explanations for this lectal variation. First, previous studies have reported that register plays an important role in the distributions of *shi* and *rang*. More specifically, *shi* is frequently used in written Chinese, while in spoken Chinese people prefer to choose *rang* (Wan 2004; Miyake 2005, Yang 2016). Although we controlled the register in the current study by only looking at newswire data for all the three varieties, the news articles from the different varieties may display stylistic variation. For instance, the news articles

in Taiwan and Singapore Chinese may be more informal than in Mainland Chinese, leading to a lectal difference in the distributions of *shi* and *rang*. A second explanation is that like the usage of *doen* and *laten* in Netherlandic Dutch and Belgian Dutch (Speelman and Geeraerts 2009; Levshina, Geeraerts, and Speelman 2013a), the division of labor between *shi* and *rang* differs across the three varieties. In Mainland Chinese, *shi* and *rang* are both frequently used and show different preferences for the semantic class of the caused events, whereas in Taiwan and Singapore Chinese, *shi* becomes an obsolescent marker that has been gradually replaced by *rang*. However, more cross-variety and cross-register investigations are needed to verify which explanation reflects the real picture of language use.

6 Concluding remarks

To conclude, this study contributes to the discussion on Chinese analytic causatives by exploring syntactic and semantic factors that constrain the choice of *shi*, *ling* and *rang*, which are the three most frequently used causative markers in contemporary written Chinese. By incorporating a cross-variety perspective, we also found significant lectal variation in the alternation of Chinese analytic causative constructions. As a case study that explores multiple factors based on a large dataset, this study provides a showcase of how bottom-up data analytics in the framework of Cognitive Linguistics can help to draw new insights on construction alternation studies. More specifically, it provides empirical evidence pertaining to Chinese analytic causative constructions on the benefits of combining tree-based methods and logistic regression modelling.

However, the results reported here should be considered in the light of some potential limitations. The first limitation concerns the design of variables. In this case study, we only included one language external factor, i.e., Variety, in our models. Other variables which have been discussed in the literature of Chinese analytic causatives (e.g, the register of the texts) should also be explored in future studies to achieve a more adequate account. Second, we point out that the competition between *shi* and *ling* may be attributable to their grammaticalization processes, however, without an empirical study based on diachronic data, this conclusion should be taken with caution.

Appendix: R codes of the study

```
# Activating necessary packages:
> library(readr); library(party); library(caret); library(dplyr);
library(lattice); library(pdp); library(nnet); library(car); library
(ggeffects); library(sjmisc)

# Preparing the data:
> data <- read_csv("dataname.csv")
> slr <- data %>%
Select (Item, CrLocus, CrSynt, CrSem, CrPers, CrDef, CrIntent,
CseModality, CseAdv, CseNeg, CeSynt, CeSem, CePers, CeDef, CeControl,
CeRole, CsedModality, CsedAdv, CsedNeg, Coref, PredSynt, PredSem,
CsedProsody, CsedSemS, CsedSemT, Structure, ClauseType, Variety)
> slr[] <- lapply(slr, factor)

# Data splitting (70% training set, 30% test set):
> set.seed(18)
> trainsamples <- createDataPartition(slr$Item, times = 1, p = 0.70)
> trainsamples <- unlist(trainsamples)
> data_train <- slr[trainsamples, ]
> data_test <- slr[-trainsamples, ]

# Creating a random forest using {party} package:
> m_cf <- cforest(Item ~ ., data=data_train, control = cforest_unbiased
(ntree= 1000, mtry = 5))
> m_cf.varimp <- varimp(m_cf, conditional=TRUE)
> dotplot(sort(m_cf.varimp), main="Random forest (ntree=1000)",
xlab="variable importance", panel=function(x,y){
    panel.dotplot(x,y,col="darkblue", pch=16, cex=1.2)
    panel.abline(v=abs(min(m_cf.varimp)), col="red", lty="longdash",
    lwd=2)
    panel.abline(v=0, col="blue")})

# Prediction of random forest on the 30% test data:
> pred <- predict (m_cf, newdata = data_test, OOB = TRUE, type = "response")
> table(observed=slr$Item[-trainsamples], predicted=pred)
```

```
# Building a conditional inference tree with the variables evaluated to be
significant by the random forest using {party} package:
> ctree_model <- ctree(Item~Variety + CsedSemT + CeSynt + CsedProsody +
ClauseType + PredSynt + PredSem, data=data_train, controls =
ctree_control(testtype = "MonteCarlo", mincriterion = 0.95, minbucket = 50))
> ctree_model
> plot(ctree_model, main="Conditional Inference Tree (alpha=0.05, min = 50)")

# Prediction and model evaluation on testing dataset (30%):
> data_test$pred <- predict(ctree_model, data_test[,-1])
> confusionMatrix(data_test$Item, factor(data_test$pred))

# Building a multinomial logistic regression model with the seven variables
proved to be significant in the random forest model as well as their
interactions with the variable Variety:
> fit0 <- multinom(Item ~ PredSynt + Variety + CeSynt + CsedProsody +
ClauseType + PredSem + CsedSemT + Variety:PredSynt + Variety:CeSynt + Variety:
CsedProsody + Variety:ClauseType + Variety:PredSem + CsedSemT:Variety,
data=data_train)
> Anova(fit0)

# Removing the insignificant variables and fitting the final model:
> fit <- multinom(Item ~ PredSynt + Variety + CeSynt + CsedProsody + ClauseType
+ CsedSemT + Variety:CsedSemT, data=data_train)
> Anova(fit)
> summary(fit)
> tab_model(fit)

> data_test$pred <- predict(fit, data_test[,-1])
> confusionMatrix(data_test$Item,factor(data_test$pred))

# Drawing the interaction plot:
> Var_CsedSemT<-ggeffect(fit, type = "pred", terms = c("CsedSemT",
"Variety"), ci.lvl = 0.95)
> plot(Var_CsedSemT)
```

References

Anthony, Laurence. 2019. *AntConc* (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

Arppe, Antti. 2008. *Univariate, bivariate and multivariate methods in corpus-based lexicography – a study of synonymy*. Publications of the Department of General Linguistics, University of Helsinki, No. 44. URN: http://urn.fi/URN:ISBN:978-952-10-5175-3.

Arppe, Antti. 2009. Linguistic choices vs. probabilities – How much and what can linguistic theory explain? In Featherston, S. & S. Winkler (eds.) *The Fruits of Empirical Linguistics. Volume 1: Process*. Berlin: de Gruyter. 1–24.

Bernaisch, Tobias, Gries, Stefan Th. & Joybrato Mukherjee. 2014. The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide*, *35* (1): 7–31.

Carletta Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, *22* (2): 249–254.

Cao, Jin. 2011. Shilingju cong shanggu hanyu dao zhonggu hanyu de bianhua [The change of *shi/ling* causative construction from Old Chinese to Middle Chinese]. *Yuyan Kexue (6)*: 602–617.

Chen, Xiaoying. 2005. Dai jianyu de *shi* yu *rang* zhi bijiao. [The comparison of *shi* and *rang* with pivotal constructions]. *Guangxi Social Sciences* (*2*): 156–158.

Colleman, Timothy. 2010. Beyond the dative alternation: The semantics of the Dutch aan-Dative. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in Cognitive Semantics: Corpus-driven approaches*, 271–303. Berlin/New York: De Gruyter Mouton.

Comrie, Bernard. 1981. *Language universals and linguistic typology: Syntax and morphology*. Chicago: The University of Chicago Press.

Deshors, Sandra C. & Stefan Th. Gries. 2016. Profiling verb complementation constructions across New Englishes: A two-step random forests analysis of -ing vs. to- complements. *International Journal of Corpus Linguistics 21*(2). 192–218.

Deshors, Sandra C. & Stefan Th. Gries. 2020. Mandative subjunctive vs. *should* in world Englishes: A new take on an old alternation. *Corpora 15* (2). 213–241.

Fan, Xiao. 2000. Lun zhishi jiegou [About causative constructions]. In Chinese Language Magazine (ed.). *Yufa Yanjiu he Tansuo*. Beijing: The Commercial Press. 135–151.

Gamer, Matthias, Jim Lemon, Ian Fellows & Puspendra Singh. 2019. IRR: Various coefficients of interrater reliability and agreement. *R package version, 0.84.1*. https://CRAN.R-project.org/package=irr.

Gilquin, Gaëtanelle. 2010. *Corpus, cognition and causative constructions*. Amsterdam / Philadelphia: John Benjamins.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Gries, Stefan Th. 2019. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory*.

Hosmer, David W. & Stanley Lemeshow. 2000. *Applied Logistic Regression*. New York: Wiley.

Hu, Yunwan. 2002. Dai jianyu de *shi* he *rang* zhi bijiao yanjiu. [The comparison research of *shi* and *rang* with pivotal constructions]. *Songliao Journal* (1): 86–86, 93.

Huang, Chu-Ren. 2009. *Tagged Chinese Gigaword Version 2.0 LDC2009T14. Web Download*. Philadelphia: Linguistic Data Consortium.

Hothorn, Torsten, Kurt Hornik & Achim Zeileis. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15* (3): 651–674.

Janda, Laura A. 2013. *Cognitive Linguistics-The Quantitative Turn: The Essential Reader*. Berlin/New York: Walter de Gruyter.

Kuhn, Max & Kjell Johnson. 2016. *Applied Predictive Modeling*. New York: Springer.

Liesenfeld Andreas, Meichun Liu & Chu-Ren Huang. 2020. Profiling the Chinese causative construction with rang (讓), shi (使) and ling (令) using frame semantic features. *Corpus Linguistics and Lingustic Theory*. Aop.

Levshina, Natalia. 2011. *Doe Wat Je Niet Laten Kan: A Usage-based Analysis of Dutch Causative Constructions*. Leuven: Catholic University of Leuven dissertation.

Levshina, Natalia. 2015. *How to Do Linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins.

Levshina, Natalia. In press. Conditional inference trees and random forests. In: Magali Paquot & Stefan Th. Gries (eds.), *Practical Handbook of Corpus Linguistics*. New York: Springer.

Levshina, Natalia, Dirk Geeraerts & Dirk Speelman. 2013a. Towards a 3D-grammar: Interaction of linguistic and extralinguistic factors in the use of Dutch causative constructions. *Journal of Pragmatics*, (52): 34–48.

Levshina, Natalia, Dirk Geeraerts & Dirk Speelman. 2013b. Mapping constructional spaces: A contrastive analysis of English and Dutch analytic causatives. *Linguistics 51*(4): 825–854.

Liu, Yonggeng. 2000. Shilinglei dongci he zhishici [Imperatives and causatives]. *Journal of Xinjiang University (Social Science)* (*1*): 93–96.

McFadden, Daniel. 1973. Conditional Logit Analysis of Qualitative Choice Behaviour. In Paul Zarembka (ed.). *Frontiers in Econometrics*. New York: Academic Press.

McFadden, Daniel. 1974. The measurement of urban travel demand. *Journal of Public Economics*, *3*(4): 303–328.

Miyake, Takayuki. 2005. A usage-based analysis of the causative verb shi in Mandarin Chinese. In Takagaki, T., Zaiman S., Tsuruga, Y., Moreno-Fernandez, F. & Kawaguchi, Y. (eds.). *Corpus-based Approaches to Sentence Structures*. Amsterdam/Philadelphia: John Benjamins Publishing Company. 77–94.

Ni, Yueru. 2012. *Categories of Causative Verbs: A Corpus Study of Mandarin Chinese*. Utrecht: Utrecht University MA thesis.

Niu, Shunxin. 2007. Putonghua zhishici de sange yufahua jieduan [Three grammaticalization phrases of mandarin causative verbs]. *Shehui Kexuejia* (*3*): 206–209.

Niu, Shunxin. 2014. *Hanyuzhong Zhishi Fanchou de Jiegou Leixing Yanjiu [A Typological Study of Causatives in Chinese]*. Tianjin: Nankai University Publishing.

Orwin, Robert G. 1994. Evaluating coding decisions. In Harris Cooper & Larry V. Hedges (Eds.), *The handbook of research synthesis*. Russell Sage Foundation. 139–162.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. URL http://www.R-project.org/.

Ripley, Brian D. 1996. *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Röthlisberger, Melanie, Jason Grafmiller & Benedikt Szmrecsanyi. 2017. Cognitive indigenization effects in the English dative alternation. *Cognitive Linguistics*, *28*(4): 673–710.

Speelman, Dirk & Dirk Geeraerts. 2009. Causes for causatives: The case of Dutch doen and laten. In T. Sanders & E. Sweetser (eds.). *Linguistics of Causality*, 173–204. Berlin: Mouton de Gruyter.

Stefanowitsch, Anatol. 2001. Constructing causation: A Construction Grammar approach to analytic causatives. Houston, TX: Rice University dissertation.
Strobl, Carolin, James Malley & Gerhard Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods 14* (4): 323–348.
Stukker, Ninke. 2005. *Causality marking across levels of language structure*. University of Utrecht dissertation.
Szmrecsanyi, Benedikt, Jason Grafmiller, Benedikt Heller & Röthlisberger Melanie. 2016. Around the world in three alternations: Modeling syntactic variation in varieties of English. *English World-Wide 37*(2). 109–137.
Tagliamonte, Sali A. & Harald R. Baayen. 2012. Models, forests, and trees of York English: Was/ were variation as a case study for statistical practice. *Language Variation and Change 24*(2). 135–178.
Talmy, Leonard. 2000. *Toward a Cognitive Semantics Vol.I Concept Structuring Systems*. Cambridge, MA.: The MIT Press.
Tian, Xiaoyu & Weiwei Zhang. 2020. Hanyu biantizhong fenxixing zhishi goushi bianyi yanjiu: Duofenlei luojisidi huigui jianmo [Chinese analytic causative constructions and their lectal variation: A multinomial logistic regression]. *Foreign Languages and Their Teaching* (3). 22–33.
Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.
Venables, W. N. & Brian D. Ripley. 2002. *Modern Applied Statistics with S (fourth edition)*. Berlin: Springer.
Verhagen, Arie & Suzanne Kemmer. 1997. Interaction and causation: Causative constructions in modern standard Dutch. *Journal of Pragmatics*, (27): 61–82.
Wan, Xinzheng. 2004. *Xiandai Hanyu Zhishi Goushi Yanjiu* [Research of Causative Constructions in Modern Chinese]. Shanghai: Fudan University PhD dissertation.
Xiang, Kaixi. 2002. Hanyu de shuangshili goushi [Double-force constructions in Chinese]. *Yuyan Yanjiu* (2): 70–77.
Xiong, Zhongru. 2004. *Xiandaihanyuzhongde zhishi jushi [Causative constructions in Modern Chinese]*. Hefei: Anhui University Press.
Xu, Dan. 2003. *Shi* ziju de yanbian, jiantan *shi* de yufahua [The grammaticalization of *shi*-construction and the evolution of *shi*]. In Wu (ed.). *Yufahua yu Yufa Tansuo*. Beijing: The Commercial Press.
Yang, Jiangfeng. 2016. *Hanyu Yuhui Zhishi Jiegou de Duoweidu Yanjiu* [A Multi-Dimensional Study of Periphrastic Causative Construction in Mandarin Chinese]. Doctoral dissertation, Zhejiang University.
Zhang, Lili. 2005. Cong shiyi dao zhishi [From order to causation]. *Taida Wenshi Zhexue Bao:* 119–152.
Zhang, Weiwei, Dirk Speelman & Dirk Geeraerts. 2011. Variation in the (non) metonymic capital names in Mainland Chinese and Taiwan Chinese. *Metaphor and the Social World 1*(1). 90–112.
Zhou, Hong. 2004. *Xiandaihanyu Zhishi Fanchou Yanjiu* [Research of Causative Category in Modern Chinese]. Shanghai: East China Normal University PhD dissertation.