Lectal Variation of the Two Chinese Causative Auxiliaries

Cing-Fang Shih, Mao-Chang Ku, Shu-Kai Hsieh
Graduate Institute of Linguistics, National Taiwan University
r08142004@ntu.edu.tw, d08142002@ntu.edu.tw, shukaihsieh@ntu.edu.tw

摘要

本文旨在從語料庫的觀點研究中文兩個使役助動詞*使* ‘cause’ 和*讓* ‘let’ 之間的差異。我們對從兩個語料庫中提取的中文語料進行邏輯迴歸分析，認為此兩個助動詞之間的差異可視為 Verhagen and Kemmer (1997) 所提出的直接/間接使役區分。回歸模型得到的結果表明，直接/間接使役的理論為動詞的特徵和詞義提供了合理的解釋。我們指出，動詞*使*與「直接使役」相關，因為它通常使用於涉及無生命參與者的使役事件中，在這種情況下，起因論旨角色必然且直接地導致受使役者的結果狀態。另一方面，*讓*應該被歸類為「間接使役」，因為它通常用於涉及有生命參與者的場景，並且除了使動者之外，亦有其他一些驅動來源也導致使役事件的發生。

Abstract

This paper aims to investigate the variation between two Chinese causative auxiliaries *shi* ‘使’ and *rang* ‘讓’ from a corpus-based perspective. We conduct a logistic regression analysis to the Chinese data extracted from two corpora and propose a direct/indirect distinction (Verhagen and Kemmer 1997) between the two auxiliary verbs. The results retrieved by the regression model show that the theory of direct/indirect causation provides a reasonable account for the characteristics and lexical meanings of the verbs. We indicate that the verb *shi* is correlated with “direct causation” because it is typically used when inanimate participants are involved in the causing event, in which the force initiated by the cause inevitably and directly leads to the resulted stage of the causee. On the other hand, the verb *rang* should be classified as “indirect causation” because it is typically used in scenarios where animate participants are both involved, and some extra force besides the causer also plays a role in the effected event.

關鍵詞：語言變異，使役結構，邏輯迴歸，R 語言統計

Keywords: language variation, causation, logistic regression, R statistics.

1. Introduction

The causative construction has been a debatable subject in linguistic studies. It is widely accepted that there are two participants encoded in a causative construction, which are the causer and the causee. The causing event led by the causer, and the caused event formed by the cause, are two components of a causative construction [1]. Verhagen and Kemmer [2] described the causative verb as a 'causal predicate', and the infinitive in the construction is called 'effected predicate', which includes two varieties: intransitive and transitive. In Mandarin Chinese, causative verbs *shi* '使' and *rang* '讓' can form causative constructions, see (1).

(1) a. 你又說了幾句讓我印象深刻的話

nǐ yòu shuō-le jǐ-jù ràng wǒ
you again say-PERF several-CL make me
yìnxiàng shēnkè de huà
impression deep MOD words
'You again say something that has deeply impressed me.'

b. 現代通訊科技使我們可以天天通話

xiàndài tōngxùn kējì shǐ wǒmen
modern communications technology make us
kěyǐ tiāntiān tōnghuà
able every.day call
'Modern communications technology enables us to call every day.'

In (1), the subject before *shi* or *rang* is the causer, and the object after the predicate is the causee. Constructions with causal predicates *shi* and *rang* are categorized as direct and indirect causation, respectively. Most of the time, direct causation is more likely to indicate non-human interaction than the indirect one is. To clarify the usages of the two causal predicates, this study is going to demonstrate a corpus-based regression analysis to explore the word choice between *shi* and *rang*. Furthermore, the regression analysis explains how the property of the causal

predicates influences the tendency of choosing direct or indirect causation.

This paper is organized as follows. Section 1 is the introduction. Section 2 briefly reviews related literature. Section 3 describes our research methods. Section 4 presents our results. A direct/indirect dichotomy is argued for and a comparison between Chinese and Dutch causative predicates is made. Section 5 concludes this paper.

2. Literature Review

The structure of causative construction reflects human's real-world experience of the relationship between the cause and the result. It is widely discussed from the typological aspect and the cognitive aspect. From the typological point of view, causatives are widely classified into three different types: (i) lexical causatives, (ii) morphological causatives, and (iii) analytic causatives [3]. From the cognitive point of view, Croft [4] explained the Idealized Cognitive Model (ICM) based on Lakoff [5]. Croft [4] views the causative construction as a single event, and it falls into three categories: (i) causative, (ii) inchoative, and (iii) stative. Both Comrie's [3] and Croft's [4] classifications of causative construction are defined as a continuum, which expresses that a linguistic expression does not always neatly fall into one of the three types. Instead, it can fit in between the two adjacent types.

Croft [6] schematized the causation types proposed by Talmy [7, 8], as shown in Figure 1. Two dimensions distinguish the four causation types. The first dimension makes distinctions between the initiator and endpoint in a causative construction. The other dimension shows differences between the animate and inanimate. Animates are seen as the mental dimension, and inanimates are physical. As demonstrated by Figure 1, the two arrows starting from the physical entity, which are affective and physical, are rather straight and direct. It shows that physical entities can act on other entities directly. On the other hand, the two arrows starting from the mental entity are not straightforward. The arrow of mental-on-mental causation, which is inducive, is rather bent. Also, the arrow of mental-on-physical causation, which is volitional, is slightly bent. It shows that mental entities cannot act on others as directly as physical entities.

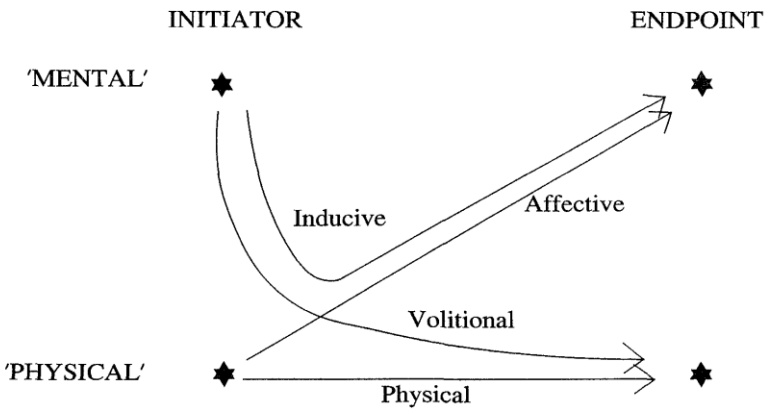

Figure 1. A Model of Causation Types (Croft 1991: 167; based on Talmy 1976)

The model of causation types ([6, p. 167]; based on [7]) is used in the study of Verhagen and Kemmer [2] to analyze the causative constructions in modern standard Dutch formed by *doen* and *laten*. According to the estimate of Verhagen and Kemmer [2], *laten* should indicate inducive (mental-on-mental) causation and should have more animate causers than inanimate ones, for it forms indirect causation. By contrast, *doen* should have more inanimate causers, for it is thought to be the component of direct causation.

3. Research Methods

To understand the usages of the two predicates under different circumstances, the data that contains *shi* and *rang* were extracted. The traditional Chinese data is collected from Academia Sinica Balanced Corpus of Modern Chinese, and the simplified Chinese data is from The Chinese Web Corpus (zhTenTen). The data contains four categories, which are traditional *shi*, traditional *rang*, simplified *shi*, and simplified *rang*. Two hundred items of each category are selected randomly for further analysis. After removing the data in which *shi* and *rang* are not used as causative verbs, the remaining 606 data were annotated.

First of all, whether the subject and the object of the data are mental or non-mental will be decided. If the subject, usually a human being or an institute that is operated by humans, can conduct the causing event of their own will, it is a mental subject. By contrast, if the subject is inanimate, it is marked as non-mental. A mental object is decided when it spontaneously executes the caused event. Otherwise, it will be considered as non-mental. The properties of

the causal predicates are decided by subjects and objects in causative constructions. If the causative construction contains a mental subject and a mental object, its property is annotated as inducive, as exemplified in (2), repeated from (1a).

(2) 你又說了幾句讓我印象深刻的話

nǐ	yòu	shuō-le	jǐ-jù	ràng	wǒ
you	again	say-PERF	several-CL	make	me
yìnxiàng	shēnkè	de	huà		
impression	deep	MOD	words		

‘You again say something that has deeply impressed me.’

In (2), the causer is *nǐ* ‘you’, which is a mental subject, and the causee is *wǒ* ‘me’, which is a mental object. The causer can directly influence the causer, while the causer can decide to perform the influence of his or her own will.

As shown in (3), volitional causation is defined when the construction contains a mental subject and a non-mental object.

(3) 他讓事件的終點等於起點

tā	ràng	shìjiàn	de	zhōngdiǎn	děngyú	qǐdiǎn
he	make	event	MOD	end.point	equal.to	starting.point

‘He makes the endpoint of the event equal to its starting point.’

In (3), the causer is *tā* ‘he’, and the causee is ‘the endpoint of the event’. This causative construction contains a mental subject the performs influence on the non-mental object.

If the construction features a non-mental subject and a mental subject, it is considered affective, as demonstrated in (4), reproduced from (1b).

(4) 現代通訊科技使我們可以天天通話

xiàndài	tōngxùn	kējì	shǐ	wǒmen
modern	communications	technology	make	us
kěyǐ	tiāntiān	tōnghuà		
able	every.day	call		

‘Modern communications technology enables us to call every day.’

The example given in (4) contains a non-mental subject *xiàndài tōngxùn* ‘modern communications’ and a mental object *wǒmen* ‘us’. The non-mental subject does not voluntarily act on the object; however, the object has influenced.

Finally, physical causation is found when both the subject and the object are non-mental, as shown in (5).

(5) 長壽能使文化承繼較完整

chángshòu	néng	shǐ	wénhuà	chéngjì	jiào	wánzhěng
longevity	can	make	culture	inheritance	more	complete

‘Longevity can make cultural inheritance more complete.’

Both the subject and object in (5) are non-mental. It presents an indirect act which is done by the inanimate subject to the object that is also inanimate.

After the properties are recorded, the transitivity variable is annotated by the transitivity of verbs after the causal predicates. The verb expresses the function of an ‘effected predicate’ [2], and it can be transitive or intransitive. If the verb requires an object, it is marked transitive (TR). Otherwise, it will be considered intransitive (INTR).

Finally, the varieties of the data are being marked for further analysis. There are two varieties, Chinese Traditional (CHT) and Chinese Simplified (CHS), based on their sources.

The annotated data is then being fitted to a logistic regression model (cf. Levshina [9], Geeraerts [10]). We choose to adopt the model because logistic regression is suitable for modeling a set of binary dependent variables. In this study, the statistics returned by the logistic regression model will be examined for the analysis of the word choice between two auxiliaries.

4. Results

4.1. Evaluation

The output retrieved by the regression model is given below in Table 1, which contains several columns with different statistics.

Table 1. A Logistic Regression Analysis to the Two Chinese Causative Auxiliaries

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	606	LR chi2	152.46	R2	0.297	C	0.777
rang	278	d.f.	5	g	1.324	Dxy	0.555
shi	328	Pr(> chi2)	<0.0001	gr	3.759	gamma	0.600
max |deriv|	3e-07			gp	0.274	tau-a	0.276
				Brier	0.190		

	Coef	S.E.	Wald Z	Pr(>|Z|)
Intercept	0.4308	0.2172	1.98	0.0473
Property=inducive	-1.3660	0.3091	-4.42	<0.0001
Property=physical	1.5052	0.2146	7.01	<0.0001
Property=volitional	0.1508	0.3704	0.41	0.6839
Transitivity=TR	-0.2606	0.1913	-1.36	0.1733
Varieties=CHT	-0.8006	0.1971	-4.06	<0.0001

As illustrated in Table 1, the column on the upper left reports the total number of observations and the frequency of each verb in our dataset.

The “Model Likelihood Ratio Test” column in the middle of the upper part of Table 1 provides an overall picture of whether the model is significant in general. In this column, one can find the Likelihood Ratio test statistic, the number of degrees of freedom, and the *p*-value. Since the *p*-value is smaller than 0.05 (< 0.0001), our model is significant, i.e. at least one predictor is significant in our model.

The rightmost column of the upper part of Table 1 contains the concordance index C, which is the proportion of the times when the model predicts a higher probability of *shi* for the sentence with *shi,* and a higher probability of *rang* for the sentence with *rang*. The statistic C in our model is 0.777. This means that for 77.7% of the pairs of *shi* and *rang* examples, the

predicted probability of *shi* is higher for the sentence where the speaker actually used *shi* than for the example where *rang* occurred. According to the scale proposed by Hosmer and Lemeshow [11, p. 162] given in Table 2 below, the discrimination in our result is acceptable.

Table 2. A Scale for the Index C (Hosmer and Lemeshow 2000: 162)

$C = 0.5$	no discrimination
$0.7 \leqq C < 0.8$	acceptable discrimination
$0.8 \leqq C < 0.9$	excellent discrimination
$C \geqq 0.9$	outstanding discrimination

Finally, the lower part of Table 1 contains the figures of coefficients. These values represent the estimated log odds of the outcome when all predictors are at their reference levels, which correspond to affective causation, intransitive effected predicates, and CHS materials.

If the coefficient is positive, the level specified in the table boosts the chances of *shi* and decreases the odds of *rang.* If the coefficient is negative, the specified level decreases the odds of *shi* and boosts the chances of *rang.* For the predictor of Causation Property, the reference level is 'affective'. We can see that only inducive causation has negative coefficients. This means that inducive causation decreases the odds of *shi*, and, conversely, boosts the chances of *rang,* in comparison with affective causation. Physical causation has the biggest positive estimate, so it seems to significantly boost the chances of *shi*, i.e. has a strong preference for choosing *shi* instead of *rang*, in comparison with the reference level. Transitive effected predicates seem to disfavor *shi* when compared with intransitives, though the difference is merely subtle. The odds of *shi* in the CHS variety are much higher than those in the CHT variety.

These findings can be nicely accounted for if we adopt a direct/indirect distinction [2] between *shi* and *rang*. As the verb *shi* is correlated with "direct causation", it is typically used when inanimate participants are involved in the causing event, in which the force initiated by

the cause inevitably and directly leads to the resulted stage of the causee. Therefore, the fact that physical causation, characterized as having both a non-mental causer and a non-mental causee, particularly favors the use *shi* but not *rang* is not difficult to imagine.

In contrast, since the verb *rang* should be regarded as "indirect causation", it is typically used in scenarios where animate participants are both involved, and some other force besides the causer becomes the most immediate source of energy in the effected event. This explains why inducive causation, which features both a mental causer and a mental causee, has a strong tendency for choosing *rang* rather than *shi*.

A plot for the outliers and discrepancy values in our dataset is provided in Figure 3 below. One can see that there are a few observations with large discrepancies and large Cook's distance values distributed around the borders of the plot. The outliers are extracted in Table 3.

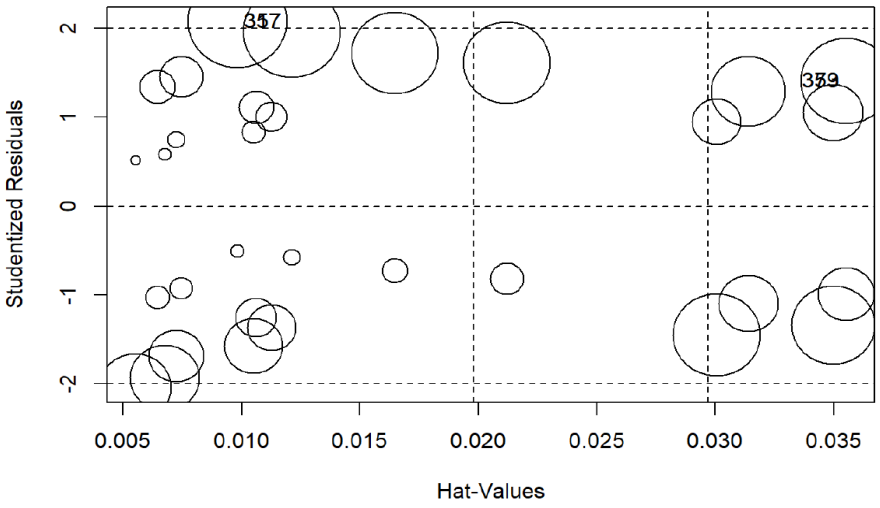

Figure 3. Plot with Outliers and Discrepancy Values

Table 3. Outliers in the Dataset

ResuPred	Property	Transitivity	Varieties
shi	inducive	TR	CHT
shi	inducive	TR	CHT

ResuPred	Property	Transitivity	Varieties
shi	volitional	TR	CHT
shi	volitional	TR	CHT

Table 3 presents contexts that are not typical of *shi*. As mentioned previously, *shi* is typically used with physical causation but not inducive or volitional causation. This is an indicator that the dataset we collected may be too coarse-grained for subtle conceptual differences, a common problem for corpus-based semantic studies.

To avoid undermining the value of the logistic regression model, overfitting is also tested, and its performance on new data is checked. The methods used in this study is to validate the model with bootstrapping (cf. Levshina [9]). The function refits the model 200 times, and the results are shown in Table 4.

Table 4. The Results of Testing for Overfitting

```
##           index.orig training   test optimism index.corrected   n
## Dxy           0.5451   0.5462 0.5392   0.0070          0.5381 200
## R2            0.2893   0.2954 0.2822   0.0132          0.2761 200
## Intercept     0.0000   0.0000 0.0041  -0.0041          0.0041 200
## Slope         1.0000   1.0000 0.9691   0.0309          0.9691 200
## Emax          0.0000   0.0000 0.0077   0.0077          0.0077 200
## D             0.2423   0.2486 0.2355   0.0131          0.2292 200
## U            -0.0033  -0.0033 0.0002  -0.0035          0.0002 200
## Q             0.2456   0.2519 0.2353   0.0166          0.2290 200
## B             0.1922   0.1905 0.1942  -0.0037          0.1959 200
## g             1.3011   1.3168 1.2700   0.0467          1.2543 200
## gp            0.2693   0.2695 0.2642   0.0053          0.2640 200
```

The model is more likely to be overfitted if the ‘optimism’ of the estimates is high. As shown in Table 4, the optimism value is 0.0386 in the line with ‘Slope’, which is relatively small. It indicates that the estimates of the regression coefficients should be trustworthy.

4.2. A Comparison between Chinese and Dutch

This subsection showcases a comparison between our results and Levshina’s [9] work on the

two causative verbs *doen* and *laten* in modern Dutch. First of all, Chinese and Dutch behave very differently with respect to the transitivity of the matrix verb. Generally speaking, in Chinese, the CHT variety seems to favor the use of *rang*, i.e. indirect causation, while the CHS variety prefers to use *shi*, i.e. direct causation. However, CHT speakers tend to use *rang* when the matrix verb is transitive, whereas CHS speakers are more inclined to use *shi* when the matrix verb is transitive.

In Dutch, the indirect variant *laten* is more frequently used than the direct variant *doen* in both dialects, the reason why Geeraerts [10] regarded *laten* as the default form in causative constructions. Besides, the two dialects behave the same with respect to transitivity as both dialects are particularly more likely to choose *laten* when the main verb is transitive. The interactions between the predictors of Varieties and Transitivity in Chinese and Dutch are schematized below in Figure 2 and Figure 3, respectively.

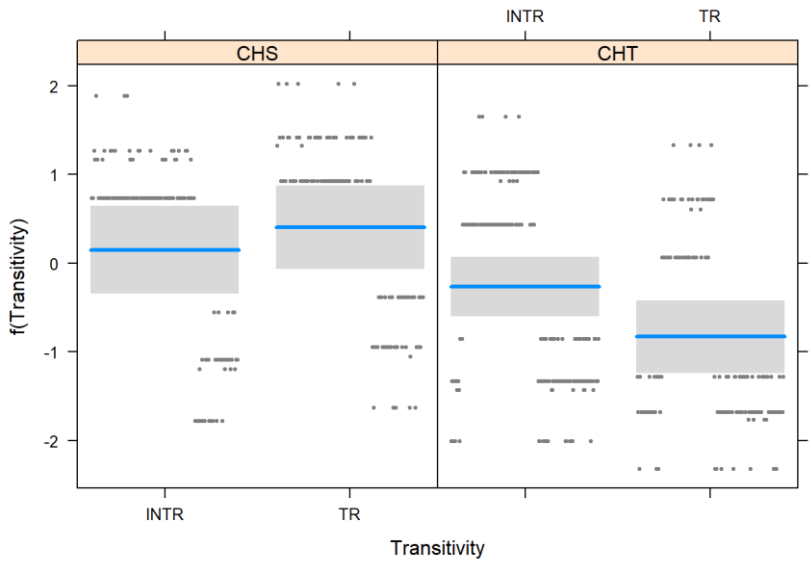

Figure 2. Interaction between Varieties and Transitivity in Chinese

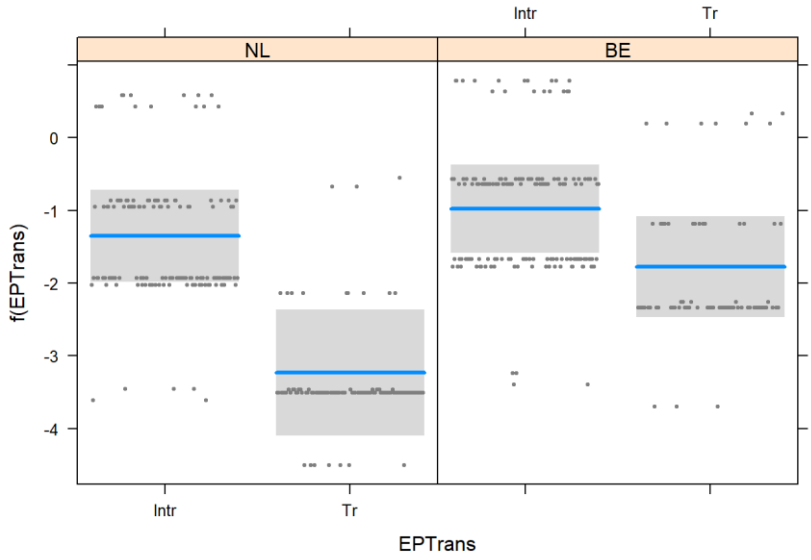

Figure 3. Interaction between Varieties and Transitivity in Dutch (Levshina 2015: 269)

The second difference between Chinese and Dutch has to do with causation types. In Chinese, physical causation tends to use *shi*, while inducive causation will opt for *rang*. Affective and volitional causation, however, have no obvious preference. No obvious difference between the two dialects is observed either. On the other hand, in Dutch, affective and physical causation are more likely to choose *doen*, while inducive and volitional causation favor *laten*. Again, no clear dialectal difference can be observed. The interactions between the predictors of Varieties and Causation Types in Chinese and Dutch are schematized below in Figure 4 and Figure 5, respectively.

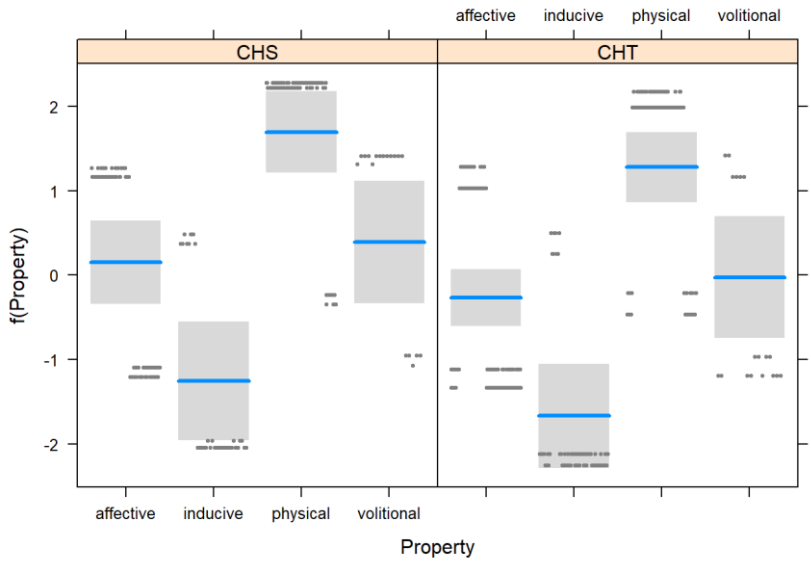

Figure 4. Interaction between Varieties and Causation Types in Chinese

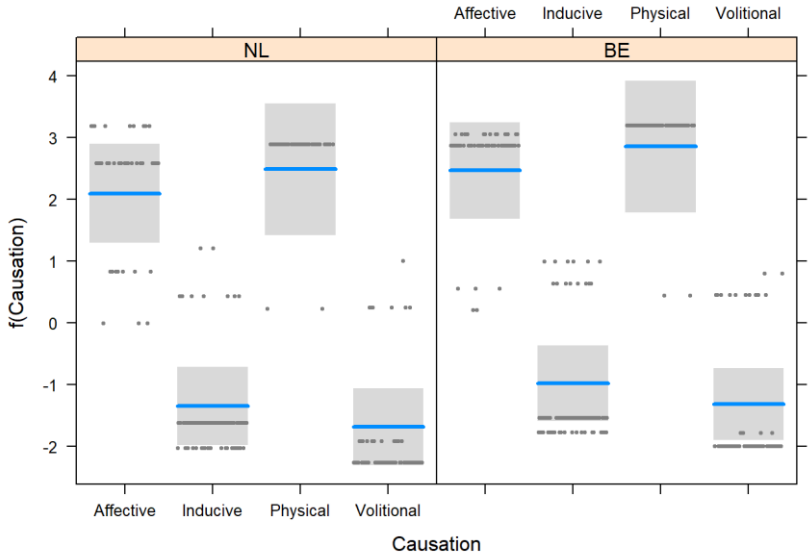

Figure 5. Interaction between Varieties and Causation Types in Dutch

5. Conclusion

Although the causative construction has been a frequently discussed subject in linguistic studies and has been widely studied in the literature, it remains unclear when it comes to the difference between its variants. This paper managed to fill the gap by providing a direct/indirect dichotomy in differentiating causation types.

We conducted a logistic regression analysis of the two Chinese causative auxiliary verbs *shi* and *rang*. The results retrieved by the regression model showed that the theory of direct/indirect causation provides a reasonable account for the characteristics and lexical meanings of the verbs. We propose that the verb *shi* is correlated with "direct causation" because it is typically used when inanimate participants are involved in the causing event, in which the force initiated by the cause directly gives rise to the resulted stage of the causee. On the other hand, the verb *rang* should be considered "indirect causation" because it is typically used in situations when animate participants are involved, and some extra force besides the causer also participates in the causal event.

In natural language processing tasks, it can be difficult to recognize how causers interact with causees in causative constructions. The results of this study explain the different usages

of direct *shi* and indirect *rang*. It is hoped that this research as well as the annotated data can make improvements to the performance in other related tasks.

To conclude, the findings help shed light on the nature of the two Chinese causative predicates. We demonstrate that the word choice between the two verbs in different contexts is influenced by the intimate relation between cognitive factors, pragmatic contextual effects, and even lexical semantics as well. A cross-linguistic survey on causation in more other languages is also necessary for future work with a view to verifying our proposal.

References

[1] M. Shibatani, "The grammar of causative constructions: A conspectus," in *The Grammar of Causative Constructions*, M. Shibatani, Ed. New York: Academic Press, 1976, pp. 1-42.

[2] A. Verhagen and S. Kemmer, "Interaction and causation: Causative constructions in modern standard Dutch," *Journal of Pragmatics*, vol. 27, no. 1, Jan., pp. 61-82, 1997.

[3] B. Comrie, *Language Universals and Linguistic Typology: Syntax and Morphology*, Chicago: University of Chicago Press, 1989.

[4] W. Croft, "Possible verbs and the structure of events," in *Meanings and Prototypes: Studies in Linguistic Categorisation*, S. Tsohatsidis, Ed. London & New York: Routledge, 1990, pp. 58-83.

[5] G. Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, Chicago: University of Chicago Press, 2008.

[6] W. Croft, *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*, Chicago & London: University of Chicago Press, 1991.

[7] L. Talmy, "Semantic causative types," in *The Grammar of Causative Constructions*, M. Shibatani, Ed. New York: Academic Press, 1976, pp. 43-116.

[8] L. Talmy, “Force dynamics in language and cognition,” *Cognitive Science*, vol. 12, no. 1, Jan., pp. 49-100, 1988.

[9] N. Levshina, *How to Do Linguistics with R: Data Exploration and Statistical Analysis*, Amsterdam & Philadelphia: John Benjamins Publishing Company, 2015.

[10] D. Geeraerts, *Ten Lectures on Cognitive Sociolinguistics*, Leiden & Boston: Brill, 2017.

[11] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, New York: Wiley, 2000.