Hypothesis Testing and beyond:
A Corpus-based Case Study of Chinese Analytic Causatives

Yanan Hu, Dirk Geeraerts and Dirk Speelman
Research Unit Quantitative Lexicology and Variational Linguistics, Department of Linguistics,
University of Leuven

I. Introduction

1. General background

Cognitive Linguistics starts to embrace corpus-based quantitative studies these years (Gries 2012, Gries & Divjak 2010, etc.). With a new emphasis on language usage, it puts Corpus Linguistics in use to bridge the gap between actual linguistic occurrences and abstract theories. This brings together method and theory, the former of which used to be regarded as number crunching and the latter of which used to criticized due to its introspective technique of research. As the usage-based researches take a bottom-up perspective, most of them are exploited to test the well-established top-down theories. The results turn out to either verify or falsify those theories, or even partially falsify, that is, to find the working theories need amending in other words (Speelman & Geeraerts 2009). All the above-mentioned advantages make Cognitive Linguistics more scientific and precise.

The case study that we demonstrate in the following sections is a Chinese case of applying quantitative techniques for the purpose of hypothesis testing. The hypothesis that we address is the (in)direct causation hypothesis.

2. Theoretical background

Causality is the relationship between an event (the cause) and a second event (the effect), where the second event is a consequence of the first. The issues related to causality and causation remain a recurrent topic. And its linguistic construal has gained much interest in many a field of linguistics too. Previous studies include the definition and categorization of causatives, e.g. morphological, periphrastic/productive/analytic and lexical causatives (Shibatani 1976), classification of causation types, e.g. physical, affective, volitional and inductive causations (Croft 1991), differentiation between the concepts pertinent to causation, such as CAUSE, PERMIT and PROHIBT (Wolff, Song & Driscoll 2002), and a great number of studies devoted to caused motion construction in the framework of Construction Grammar (Goldberg 1995). Besides, different notions have been proposed to distinguish causatives and causations in lexicology and Cognitive Semantics. Rather than the concepts like compactness (Song 1996), implicativity (Shibatani 1976), we focus on the conceptual distinction of direct/indirect causation in the present study.

The (in)direct causation hypothesis was first formulated by Suzanne Kemmer and Arie Verhagen (Verhagen & Kemmer 1992, Kemmer & Verhagen 1994, Verhagen & Kemmer 1997, Verhagen 1998, Verhagen 2000) and was more analyzed by Ninke Stukker (2005). It crucially involves the flow of energy in the causative event. We use an example in Speelman and Geeraerts's study in 2009 to explain:

The professor	[made]	the students	follow the scientific method.
NP1	CAUSE	[NP2	V NP3]
subject (matrix sentence)		subject (embedded sentence)	object (embedded sentence)

causer	causee	affectee
direct causal instigator	the intermediary	ultimately affected entity

The (in)direct causation hypothesis states that the choice for either *doen* or *laten* in Dutch is influenced by the degree of involvement of the causee. Stukker (2005) further claims that the causer of *doen* produces the effected event directly so there is no intervening energy source "downstream" while, in the case of *laten*, besides the causer, the causee is the most immediate source of energy in the effected event. In other words, the causee has some degree of "autonomy" in the causal process. Therefore *laten* expresses indirect causation. However, this hypothesis didn't stop there. Ni's corpus research (2012) follows its reasoning and extends to Chinese analytic causatives *shi* and *rang*. In Ni's words, "the event descriptions with *shi* categorize as direct causation in that the causer directly brings about the result of causee, thus inanimate entities are involved, while *rang* categorizes as indirect causation in that a more immediate force in working in the causal event other than the causer to bring about the result."

Yet Speelman and Geeraerts (2009) derives 6 predictions from the assumption that *doen* is associated with direct causation, *laten* with indirect causation, builds a model with corresponding predictors, applies the statistical analysis of corpus materials to come to a conclusion that the (in)direct causation hypothesis is doubtful. The results show that most of the 6 predictions are falsified, and that it is therefore necessary to pursue a different hypothesis about the causes for choosing either *doen* or *laten*. Based on the suggestion of this study, Levshina (2011) depicts a more refined picture of the division of semantic and lectal labour of *doen* versus *laten* ('Lect' is a cover term for all types of language varieties, like dialects, regiolects, sociolects, registers and so on). These studies also cast doubt on the validity of the (in)direct causation hypothesis for Chinese *shi* and *rang*. Does this hypothesis provide a clear insight into the nature of *shi* and *rang*, as Ni (2012) mentions? Or is the Chinese another case that cannot be fairly explained only by the conceptual difference?

In the next section of this paper, we will introduce the research target and our design of the study.

II. Research target and design of the study

2.1 Research target

When we talk about Chinese analytic causatives, it is necessary at this point to have a look at the analytic causative construction in Chinese:

我	[让]	客 人	围	着	桌 子	坐 下。
wǒ	ràng	kè rén	wéi	zhe	zhuō zi	zuò xià
I	cause	the guests surround (present tense marker) the table sit down				
I	asked	the guests to sit around the table.				
NP1	CAUSE	[NP2	V	NP3]		
causer	CAUSE	causee	caused event			

In the analytic causative construction mentioned above, the slot fillers of CAUSE are analytic (periphrastic/productive) causatives that we are interested in here. In Mandarin Chinese nowadays, we find 7 forms of realization for further investigation. They are 使 *shǐ*, 令 *lìng*, 让 *ràng*, 叫 *jiào1*, 教 *jiào2*, 给 *gěi* and 要 *yào*. For the purpose of simplicity and the consistency between the interpretation and the immediate output given by R, we choose not to indicate the tones when we refer to them, i.e. *shi*, *ling*, *rang*, *jiao1*, *jiao2*, *gei* and *yao*. Amidst the seven of them, two of them, *jiao*, are homophones. We give label 1 and 2 to distinguish them.

2.2 The materials

The data in our study were retrieved from the second edition of the UCLA Written Chinese Corpus, UCLA2 (http://www.lancaster.ac.uk/fass/projects/corpus/UCLA/, Tao & Xiao 2012). The corpus is a compilation of written modern Chinese available from the internet during the period of 2000-2012. The total number of tokens in UCLA2 is 1,119,930, which covers 15 genres. The hits of the characters we extracted are shown in Table 2.2.1.

Characters	*shi*	*ling*	*rang*	*jiao1*	*jiao2*	*gei*	*yao*
Hits	730	254	1,507	576	161	1,425	3,028

Table 2.2.1 Number of tokens of each character in UCLA2

Then we filter to find the occurrences, in which the characters are used as analytic causatives, shown in Table 2.2.2.

Causatives	*shi*	*ling*	*rang*	*jiao1*	*jiao2*	*gei*	*yao*
Occurrences	686	205	1,200	77	11	130	100

Table 2.2.2 Number of sentences with causal use of the characters

However, we find in the dataset a phenomenon of Chinese language, which can be illustrated in Example 2.2.

(2.2) a. 因此， 陷 入 社 会 公 共 危机 的 政 府，
yīn cǐ xiàn rù shè huì gōng gòng wēi jī de zhèng fǔ
therefore, driven to societal public crisis (adjective signifier) government,
必须 采取 强 有 力 执 法 手 段，严 格执 法 ，
bì xū cǎi qǔ qiáng yǒu lì zhí fǎ shǒu duàn yán gé zhí fǎ
must adopt powerful forceful enforce law means, strictly enforce law,
将 公 众 的 行 为 约 束 在 法律的 范围 内，
jiāng gōng zhòng de xíng wéi yuē shù zài fǎ lǜ de fàn wéi nèi
(object signifier) public behavior restrict in legal scope inside,
从 而 [使] 政 府 能 够 有 效 地 整 合社 会 资 源，
cóng ér shǐ zhèng fǔ néng gòu yǒu xiào de zhěng hé shè huì zī yuán
thus CAUSE government can effectively combine societal resource,
战 胜 危机，将 社 会 损 失 降 低 到 最 低 。
zhàn shèng wēi jī jiāng shè huì sǔn shī jiàng dī dào zuì dī
overcome crisis, (object signifier) societal loss reduce to minimum
Therefore, the government in public crisis must take some powerful actions to enforce the law, and restrict public behaviors of people to such a legal scope as to make the government capable of effectively integrating societal resources, overcoming the crisis, and minimizing the loss.

b. 个 性 化 作 业 不 仅 [使] 学 生 的
gè xìng huà zuò yè bù jǐn shǐ xué shēng de
individualize assignment not only CAUSE student (genitive signifier)
个 性 得 到 张 扬， 能 力得 到 培 养， 而且 学 习 成 绩
gè xìng dé dào zhāng yáng néng lì dé dào péi yǎng ér qiě xué xí chéng jì
personality get display, ability get cultivate, and study score
也 得 到 了 提高 。
yě dé dào le tí gāo
also get (perfect tense signifier) improve

Individualized assignment not only makes students' personality displayed, their ability cultivated, but also makes their academic performance improved.

Example 2.2 (a) has 3 verb phrases as effected predicate – integrate, overcome and minimize. 2.2 (b) is more complicated. It has 3 different causees – personality, ability and performance, and each of them has their own effected predicate, but they are all affected by one causative *shi*. In these situations, our solution is to break the whole sentence up into several sub-clauses to make sure one clause contains only one causee and one VP2. But in order to distinguish the complex pattern of sentence from a simple one, we have it marked as being in a multiple structure rather than a simple one when it comes to the predictor **Structure**. We will come to it again in the following section.

Thus the number of the observations that we have ready for further analysis is 2,818 in total. Table 2.2.3 shows the frequencies.

Causatives	*shi*	*ling*	*rang*	*jiao1*	*jiao2*	*gei*	*yao*
Observations	837	232	1,393	90	14	135	117

Table 2.2.3 Number of frequencies of the 7 causatives

2.3 The variables

The 2,818 observations are annotated for the following variables.

2.3.1 The dependent variable **Causatives**

This study aims to describe the choice of the seven analytic causatives by Chinese language user and discover the underlying conceptual, semantic and grammatical factors that may influence the choice. So **Causatives** is the response variable in the analysis.

Due to the methodological limitation, the value of the response variable with a very low frequency, *jiao2*, may do harm to the model if we keep it as a separate value. So we conflate *jiao2* with the observations with *jiao1*, and make them into one level *jiao* in the present study.

2.3.2 The independent variables

The data is annotated with 24 independent variables based on the assumption that they constitute the context where the analytic causatives are used. The variables can be categorized in three ways: (1) features of different parts of the construction and their relation; (2) structural and grammatical characteristics versus conceptual/semantic and even lexical level; (3) the variables related to the (in)direct causation hypothesis and the others. We present here and in the summary table of Appendix 1 from the first perspective because it contains more sub-categories so that we can easily keep track of what the variables relate to.

Causer-related variables

The variables are features of causer. These include structural, grammatical, semantic and lexical ones. They are coded manually except one logical factor, which is calculated with the help of collocational analysis. We will provide more explanation when we come to that.

Variable 1: **CrExp**

The variable CrExp stands for "explicitness of causer", including two possible states Explicit and Implicit, which distinguish whether the causer is verbally expressed or not. It reflects a structural choice that a specific sentence makes. Example 2.3.2.1 is a case of Implicit.

(2.3.2.1) 干 海 参 用 冷 水 浸 泡 一 天 一 夜，
gān hǎi shēn yòng lěng shuǐ jìn pào yī tiān yī yè
dried sea cucumber use cold water soak one day one night
[让] 海 参 回 软。

ràng hǎi shēn huí ruǎn
CAUSE sea cucumber return soft
(You/We/People) would put dried sea cucumbers in cold water for a whole day and night to make them soft again.

The causer, the subject, is a silent element in this sentence. But it appears in an instructional "how to" text. We can infer that a human being or the plural form should play the role. The causer is there. Only it is not expressed. There is another case that we classify into Implicit as well, shown in Example 2.3.2.2.

(2.3.2.2) 红 色 的 灯 光 看 着 喜 庆，但 却 容 易 [使] 人
hóng sè de dēng guāng kàn zhe xǐ qìng dàn què róng yì shǐ rén
red laplight look festive but easily CAUSE people
血 压 升 高，呼 吸 加 快。
xuè yā shēng gāo hū xī jiā kuài
blood pressure go up breathing speed up
Red lamplight looks festive but (it) is likely to cause high blood pressure and hurried breathing.

In Example 2.3.2.2, red lamplight as the causer is in the context but not in or adjacent to the clause where the causative construction is located. This sort of causer is coded as Implicit. We have 1,796 observations of CrExp=Explicit and 1,022 cases of CrExp=Implicit.

Variable 2: **CrSem**

"Semantic class of causer" CrSem has five values: Anim, when the causer is animate – human beings and collectives like organization, animals and body parts; Inanim, when the causer is inanimate including physical materials, mechanism, abstract entities, etc.; Evt, when the causer is an event, a fact; PhyAct, when the causer is a physical activity performed by somebody; MentAct, when the causer is a mental activity, for instance, a feeling or a thought. In our dataset, the distribution of the observations is 1,834 cases of CrSem=Anim, 761 cases of CrSem=Inanim, 533 cases of CrSem=Evt, 337 cases of CrSem=PhyAct, 114 cases of CrSem=MentAct.

Variable 3: **CrPers**

CrPers stands for "grammatical person of causer". Besides 1Sg (first person singular, 148 cases), 1Pl (first person plural, 45 cases), 2Sg (second person singular, 84 cases), 2Pl (second person plural, 12 cases), 3Sg (third person singular, 1,934 cases), 3Pl (third person plural, 396 cases) to our common knowledge, we also have one value of CrPers=Undef (199 cases) to label the occurrences with implicit causer such as the aforementioned Example 2.3.2.1.

Variable 4: **CrDef**

The variable CrDef means "definiteness of causer" – whether the causer is literal or has a general reference. This semantic factor has two values CrDef=Def (definite, 2,386 cases) and CrDef=Indef (indefinite, 432 cases). An example of the marked CrDef=Indef is given in Example 2.3.2.3.

(2.3.2.3) 如 果 皆 存 在 可 能 的话，试 着 用 适 当 的 搜 索 操 作
rú guǒ jiē cún zài kě néng de huà shì zhe yòng shì dàng de sōu suǒ cāo zuò
if both exist possibility say, try use proper searching operation
来 [使] 你的 搜 索 更 精 炼。
lái shǐ nǐ de sōu suǒ gèng jīng liàn
come CAUSE your searching more refined
If there are two possibilities, (you) should try to search properly to make it more

refined.

In an informative context of skill introduction, the implicit causer *you* in the sentence refers to the google user in general, not you in particular.

Variable 5: **CrIntent**

"The intention of causer" CrIntent is a semantic variable with three values, Intent (intentional, 1,100 cases), Unintent (unintentional, 1,565 cases) and Undef (undefined, 153 cases). It distinguishes the causer who causes the caused event with efforts on purpose from the causer who may just happen to be the cause. The language has many a device to leave us a hint to judge if the causer is intentional, e.g. lexical devices – adverbial *deliberately*, *on purpose* versus *by chance*, and predicate verb phrase *try (one's best) to* versus *happen to*. However, in some cases the sentence per se and its limited context cannot point to Intent or Unintent. We assign CrIntent=Undef to this sort of occurrences in which the causer could be either intentional or unintentional.

Variable 6: **CrCollocSig**

The name of the variable CrCollocSig stands for "lexical collocational significance between CAUSE and causer". It is originally a logical factor with two possible values: TRUE and FALSE. We use it as Speelman and Geeraerts (2009) use the variable sig. lex. col. "Lexical fixation" is the information stored in such factors. With CrCollocSig we want to establish whether there is some degree of lexical fixation at play in the link between the causer and the causal verb in our dataset. In other words, we want to explore whether the occurrence of a specific lexeme as the causer triggers the choice for one of the causatives.

Establishing statistical collocation patterns is done by means of distinctive collexeme analysis developed by Stefan Th. Gries and Anatol Stefanowitsch (2011), one of the three methods of collostructional analysis. The technique of collocational analysis was applied and explicated by Speelman and Geeraerts (2009). For our implementation of CrCollocSig we follow their approach. If we can establish a significant attraction between the causer and the causative verb, CrCollocSig receives the value TRUE, otherwise it receives the value FALSE. In our dataset of 2,818 observations we have 1,066 cases of CrCollocSig=TRUE and 1,752 cases of CrCollocSig=FALSE.

Causee-related variables

These variables are features of causee. Five of them could be considered the corresponding features to those of causer.

Variable 7 – 10, 12: **CeExp**, **CeSem**, **CePers**, **CeDef**, **CeCollocSig**

CeExp, CeSem, CePers and CeDef stand for "explicitness of causee", "semantic class of causee", "grammatical person of causee" and "definiteness of causee" respectively. CeCollocSig means "lexical collocational significance between CAUSE and causee". Table 2.3.2.1 indicates the frequency of every level of these variables in our materials.

No.	Variable	Value	Frequency	No	Variable	Value	Frequency
7	CeExp	Explicit	2,542			2Sg	96
		Implicit	276			2Pl	14
8	CeSem	Anim	2,097			3Sg	1,462
		Inanim	574			3Pl	779
		Evt	17			Undef	19
		PhyAct	66	10	CeDef	Def	1,848

		MentAct	64			Indef	970
9	CePers	1Sg	367	12	CeCollocSig	TRUE	994
		1Pl	81			FALSE	1,824

Table 2.3.2.1 Frequencies of values of Variable 7-10, 12

Variable 11: **CeRole**

The semantic variable CeRole stands for "thematic role that the causee plays in the sub-clause of caused event". We start from the traditional list (Jackendoff 1990) and then boil down to five candidate values in the current study: Agt (627 cases), when the causee is an agent, an actor with volition to perform the caused motion or activity; Ptnt (780 cases), when the causee is a patient passively or even forced sometimes to undergo the caused event; Expcer (793 cases), when the causee is an experiencer, a subject in the effected clause in which the predicate is *see*, *hear*, *witness*, *observe* or *feel*, etc.; Befiry (602 cases), when the causee is a beneficiary that receives the benefits brought about by the causing event. In some cases, a beneficiary is the one that is enabled by the causer to be in a better state or to act out the caused motion successfully; Others (16 cases), when the causer takes other thematic roles, like location, instrument and so on.

Variables concerning causer-causee relationship

We only have one variable in this category in the current study. It is manually coded.

Variable 13: **Coref**

The variable Coref stands for "coreferentiality between causer and causee". Similar to that in Speelman and Geeraerts's study (2009), its has possible values N (no) and Y (yes), which stand for complete absence of coreferentiality versus presence of some type of coreferentiality. Different from theirs, we also have another value Undef (undefined) to explain the cases which are made hard to tell by an implicit causer or causee. In our dataset we have 2,348 cases of Coref=N, 445 cases of Coref=Y and 25 cases of Coref=Undef.

Causing event-related variables

These variables are features of causal verbs and of the matrix sentence. They are also encoded manually.

Variable 14: **Manner**

The semantic variable Manner distinguishes whether only pure causal meaning is used (Manner=N) or the causative conflates causal meaning and manner of causing action (Manner=Y). Example 2.3.2.4 is a case of the latter.

(2.3.2.4) 月 女 反 而 倚 在 阳 台 上 看 排 队 的 兵 走 过，
yuè nǚ fǎn ér yǐ zài yáng tái shàng kàn pái duì de bīng zǒu guo
Yuenv instead lean onto balcony watch queuing soldier go by
还 大惊小怪 [叫] 别 的女 孩 子 都 来 看。
hái dà jīng xiǎo guài jiào bié de nǚ hái zǐ dōu lái kàn
also fuss CAUSE other girls all come watch
Yuenv leaned onto the balcony instead, watching the queuing soldiers pass by, and also made a fuss, calling other girls to come over and watch too.

We also store in this variable the difference between a pure causative or *Cause*-type versus an agentive causative or *Force*-type (Terasawa 1985). We regard the former as Manner=N and the latter as Manner=Y, like in Example 2.3.2.5.

(2.3.2.5) 不 能 再 给 没 问 题的 人 的 心 灵 开 辟
bù néng zài gěi méi wèn tí de rén de xīn líng kāi pì

not can again give not problematic people (genitive signifier) mind open up
另 一 种 “隔 离 区”, [让] 他 们 承 担 不 应 当 承 担
lìng yī zhǒng gé lí qū ràng tā mén chéng dān bù yīng gāi chéng dān
another kind of isolated district CAUSE them bear not should bear
的 “非 典 之 灾 ” 。
de fēi diǎn zhī zāi
(adjective signifier) SARS’s disaster
Do not isolate the mind of the well again that it forces them to endure the SARS disaster that they are not bound to.

We have 2,267 cases of Manner=N and 551 cases of Manner=Y in the dataset.

Variable 15: **CseModality**

CseModality is a semantic variable derived from the structural feature. It stands for “type of Chinese ‘modal verbs’ in front of CAUSE”. We assign five values to this variable: None (2,339 cases), when there is no modal verb in front of the causative in the matrix sentence; Possibility (334 cases), when the modal verb indicates likelihood of the causing event, e.g. *can*; Necessity (21 cases), when the modal verb denotes the need of occurrence of the causing event, e.g. *must*; Inclination (99 cases), when the modal verb shows intention or willingness, e.g. *would rather*; Evaluation (25 cases), when the modal verb assesses degree of value or difficulty, e.g. (be) difficult to. More details of their lexical realization are given in Appendix 2.

You may notice this classification is different from Palmer’s likelihood, ability, permission and obligation (2001), and epistemic, deontic, dynamic (Huddleston & Pullum 2002) and evaluative modalities (Hsieh 2005). The reason is that Chinese language has the category of can-wish verbs, which is a type of auxiliary verbs that is used to indicate modality. They are similar to English modal verbs but not fully mapped to them. In Chinese linguistics, this issue of (dis)similarities between Chinese can-wish verbs and English modal verbs has drawn much attention (Ji 1986, Lai 2006, Liu 2007). Since our target language is Mandarin Chinese, we apply the semantic classification of Chinese can-wish verbs to deal with modality, and this classification is based upon the textbook of modern Chinese (an online course: http://www.yyxx.sdu.edu.cn/chinese/MAIN.htm). This online version is a compilation of a number of publications on modern Chinese, and adopted nowadays by university teachers to teach Chinese grammar.

Example 2.3.2.6 gives us a case of CseModality=Evaluation.

(2.3.2.6) 有 毒 这 两 个 字 很 难 [让] 人 产 生 安 全 感 。
yǒu dú zhè liǎng gè zì hěn nán ràng rén chǎn shēng ān quán gǎn
have poison these two words very hard CAUSE people generate security feeling
“Poisonous” hardly leads to sense of security.

Variable 16: **CseNeg**

The variable CseNeg focuses on the structural feature – “negation in front of CAUSE”. It receives two values: N (absence of negation) and Y (presence of negation). Negation includes usage of the particle *not* and semantically negative adverbs. Double negation counts as CseNeg=N here. In our dataset we have 2,721 cases of no negation and 97 cases with negation.

Caused event-related variables

These variables are features of the effected predicate V2 and of the embedded sentence, which depicts the caused event. Due to the similar nature of Variable 19-21 to the aforementioned

variables, we merely comment on Variable 17 and Variable 18 in detail. Manual annotation is applied to these variables as well.

Variable 17: **CsedCstr**

CsedCstr refers to "grammatical construction of the effected predicate in the embedded clause". We assign five values to this variable: Trans (transitive verb, 1,150 cases), Intrans (intransitive verb, 584 cases), Copula (copular verb and adjective, 809 cases), Idiom (idiom or set phrase, 207 cases), and SVC (serial verb construction, 68 cases). Example 2.3.2.7 illustrates the cases where Chinese idiom (a) and serial verb construction (b) play the role of V2.

(2.3.2.7) a. 她害 怕 的 是 这 场 天 降 的 爱情，[令] 她
tā hài pà de shì zhè chǎng tiān jiàng de ài qíng lìng tā
she fear (genitive signifier) is this heaven fall love CAUSE her
流 连 忘 返。
liú lián wàng fǎn
linger about forget return
She fears that this heavenly love will make her obsessed.

b. 朵颐 命 令 似 地 [叫] 他 过 去 陪 她 聊 天。
duǒyí mìng lìng shì de jiào tā guò qù péi tā liáo tiān
Duoyi imperatively CAUSE him go there accompany her chat
Duoyi ordered him over to chat with her.

Variable 18: **CsedSem**

The variable CsedSem stands for "semantic class of V2 in the embedded clause". Its values are determined in line with the lexical aspect or aktionsart of a verb (Vendler 1957, Comrie 1976, Smith 1991, Lin 2004).

Vendler's classification includes verbs that express activity, accomplishment, achievement and state (1957). Activities and accomplishments are distinguished from achievements and states in that the former allow the use of continuous and progressive aspects. Activities and accomplishments are distinguished from each other by boundedness: activities do not have a terminal point (a point before which the activity cannot be said to have taken place, and after which the activity cannot continue – for example "John drew a circle") whereas accomplishments do. Of achievements and states, achievements are instantaneous whereas states are durative. Achievements and accomplishments are distinguished from one another in that achievements take place immediately (such as in "recognize" or "find") whereas accomplishments approach an endpoint incrementally (as in "paint a picture" or "build a house"). Comrie (1976) added the category semelfactive or punctual events such as "sneeze". His divisions of the categories are as follows: states, activities, and accomplishments are durative, while semelfactives and achievements are punctual. Of the durative verbs, states are unique as they involve no change, and activities are atelic (that is, have no "terminal point") whereas accomplishments are telic. Of the punctual verbs, semelfactives are atelic, and achievements are telic.

Given these, five values are assigned to our variable CsedSem: State (1,034 cases), Activity (256 cases), Accomplishment (224 cases), Achievement (798 cases), and PunctualEvent (506 cases).

Variable 19-21: **CsedModality**, **CsedNeg**, **CsedCollocSig**

The three variables belong to three categories. CsedModality is a semantic variable developed from the grammatical/structural constituent of the embedded sentence. It refers to "type of

Chinese ‘modal verbs’ in front of V2 in the caused event”. CsedNeg is a grammatical/structural variable, with the meaning of “negation in front of the effected predicate V2”. It doesn’t take into account the cases where V2 itself has a negative meaning. CsedCollocSig is a logical variable, which stands for “lexical collocational significance between CAUSE and V2”. It is the original predictor sig. lex. col used in Speelman and Geeraerts’s study (2009). The possible values of the three variables and frequencies of the actual occurrences in our dataset are provided in the following Table 2.3.2.2.

No.	Variable	Value	Frequency	No	Variable	Value	Frequency
19	CsedModality	None	2,704	20	CsedNeg	N	2,674
		Possibility	79			Y	144
		Necessity	14	21	CsedCollocSig	TRUE	712
		Inclination	13			FALSE	2,106
		Evaluation	8				

Table 2.3.2.2 Frequencies of values of Variable 19-21

Variables concerning causing event-caused event relationship

In order to avoid collinearity, we reduce the variables in this section to one in the study – Implict.

Variable 22: **Implict**

This variable means “implicativity”, that is, a sentence is implicative (Imp) when the occurrence of causing event entails the occurrence of caused event; non-implicative (NoImp) when it does not entail the occurrence of the lower clause event (Shibatani 1976). It is also proposed that adding a counterfactual coordinate clause guided by *but* can be a test for these two kinds of sentences. The example is quoted as follows:

A. He begged Coghill to keep the matter to himself, but Coghill told everyone. – non-implicative *beg*

B. *The police got him to confess to the crime, but he didn’t confess. – implicative *get*

Get in Example B ensures an evitable consequence of his confessions to the crime. When combined with a negation of this fact, the whole sentence doesn’t conform to acceptable conventions or standards of English language. This norm is applied to our study as well, as is shown in Example 2.3.2.8.

(2.3.2.8) a. 排 长 [令] 我 班 就 地 还 击， 以 掩护 3 班
pái zhǎng lìng wǒ bān jiù dì huán jī yǐ yǎn hù sān bān
sergeant CAUSE our squad on the spot counterattack to cover number 3 squad
占 领 3 6 0 高 地。（当 我 命 令 在 我 附近 的 2
zhàn lǐng sān liù lìng gāo dì dāng wǒ mìng lìng zài wǒ fù jìn de liǎng
capture number 3 6 0 highland when I command in my vicinity 2
个 战 斗 组 还 击 时， 却 发现 班 长 龙 昌 文
gè zhàn dòu zǔ huán jī shí què fā xiàn bān zhǎng lóng chāng wén
combatant groups counterattack but find squad leader Long Changwen
带 领 的 机 枪 组 没 有 动 静……）
dài lǐng de jī qiāng zǔ méi yǒu dòng jing
heading machine gun group not have movement
The sergeant commanded our squad to fight back on the spot in order to cover Squad 3 to capture the Highland 360. (But when I commanded the 2 groups close to me to counterattack, I found the machine gun group headed by the squad

leader Long Changwen made no movement.)

b. *我 没 有 回 答 她，因 为 极度的 恐 惧已 [令] 我 说 不 出
wǒ méi yǒu huí dá tā yīn weì jí dù de kǒng jù yǐ lìng wǒ shuō bù chū
I not have answer her because extreme fear have CAUSE me say not out
话 来。（但 我 回 答了 她……）
huà lái dàn wǒ huí dá le tā
words up but I answer her
*I didn't answer her because the extreme amount of fear had deprived me of any speech. (But I answered her.)

The context of Example 2.3.2.8 (a) shows the counterattack is not effected. The caused event does not actually occur. The sentence (a) has to itself a non-implicative causative construction. However, (b) would be unacceptable if we add the counterfactual clause. We assign Implict=Imp to sentence (b). In the overall dataset, we have 1,926 implicative occurrences and 892 non-implicative occurrences.

Variables concerning causative construction

These variables are features of the entire causative construction. They are both structural/grammatical, and relatively easier than those semantic factors to code manually.

Variable 23: **SyntFun**

The variable SyntFun stands for "syntactic function of the causative construction in the whole sentence". It has three values: Pred (predicate, 2,229 cases) when causal verb is the main predicate; Inf (infinitive, 400 cases) when the causative construction can be translated into an infinitive phrase or an adverbial clause in English to express the purpose of the causer; Attr (attributive, 189 cases) when the construction functions as an attributive clause modifying a noun/pronoun. Example 2.3.2.3 mentioned above is a case of SyntFun=Inf. The following Example 2.3.2.9 presents a case of SyntFun=Attr, where *causing privacy to disappear* is the attribute of *information technology*.

(2.3.2.9) 那 么 这 种 可 能 [使] 隐私 消 失 的
nà me zhè zhǒng kě néng shǐ yǐn sī xiāo shī de
therefore this kind could CAUSE privacy disappear (adjective signifier)
信 息 技 术 到 底 是 利 大于 弊 还 是
xìn xī jì shù dào dǐ shì lì dà yú bì hái shì
information technology in the end is advantage outweigh disadvantage or
弊 大于 利， 就 值 得 探 讨 一 下 了。
bì dà yú lì jiù zhí dé tàn tǎo yī xià le
disadvantage outweigh advantage then deserve discussion one time
Therefore it deserves another discussion whether the advantages of information technology outweigh the disadvantages or the other way around since such a kind of technology could result in the disappearance of individual privacy.

Variable 24: **Structure**

The variable Structure can be translated into "the number of caused events that CAUSE takes". We assign two possible values to the variable – Single (2,082 cases) and Multiple (736 cases). Example 2.2 has shown us two specific situations of Structure=Multiple.

2.3.3 Summary of the variables related to (in)direct causation and the logic

So far we have gone through the 24 variables, eight of which could be included in the model of

direct/indirect causation to test the hypothesis. The hypothesis has to be redefined here because it is not applied to explain *doen* and *laten* in Dutch or *shi* and *rang* in Chinese. We'd like to see if this conceptual distinction is important for differentiating Chinese analytic causatives, all the seven of them in this study. But we could zoom in on the pair of *shi* and *rang* later since it has been talked about in the literature.

A summary table is presented here of the related variables and their relation to (in)directness.

No.	Variable	Direct causation ----------Indirect causation
2	CrSem	Inanim-Evt-MentAct-PhyAct-Anim
8	CeSem	Inanim-Evt-MentAct-PhyAct-Anim
11	CeRole	Others-Expcer-Ptnt-Befiry-Agt
13	Coref	Y-N
17	CsedCstr	Copula-Idiom-Intrans-Trans-SVC
18	CsedSem	State-PunctualEvent-Achievement-Accomplishment-Activit y
22	Implict	Imp-NoImp
23	SyntFun	Attr-Pred-Inf

Table 2.3.3 Variables related to (in)direct causation

Derived from the core of (in)directness, i.e. mediacy or (no) intervening cause criterion (Comrie 1981, Wolff 2003), the reasoning goes: when the variable gets the value to the left, the sentence is considered to express direct causation; when indirect causation is expressed, we may expect the construction with more values down the line to the right. Note that the variables similar to Variable 2 CrSem, Variable 13 Coref and Variable 17 CsedCstr are retained in Speelman and Geeraerts's analysis of directness/indirectness of *doen* and *laten*.

III. The results and interpretation

In the statistical analysis, we apply both exploratory technique and confirmatory technique. But first of all we will have a look at the raw frequency.

3.1 Frequency and proportion

In Section 2.2, we are acquainted with the token number of the characters in UCLA Chinese Corpus 2 (shown in Table 2.2.1), and the sentence frequency with these characters as analytic causatives (shown in Table 2.2.2). We put them together to find the proportion of their causal use (shown in Table 3.1.1).

Characters	*shi*	*ling*	*rang*	*jiao1*	*jiao2*	*gei*	*yao*
Hits	730	254	1,507	576	161	1,425	3,028
Occurrences	686	205	1,200	77	11	130	100
Proportion (%)	93.9726	80.70866	79.6284	13.36806	6.832298	9.122807	3.30251

Table 3.1.1 Proportion of causal use of the seven characters

And it is pointed out that we have a new conflated category of causative *jiao* for our analysis. So the table changes into Table 3.1.2.

Characters	*shi*	*ling*	*rang*	*jiao*	*gei*	*yao*
Hits	730	254	1,507	737	1,425	3,028
Occurrences	686	205	1,200	88	130	100
Proportion (%)	93.9726	80.70866	79.6284	11.9403	9.122807	3.30251

Table 3.1.2 Causal use proportion of the possible levels of **Causatives**

The information given by this table is semasiological salience (Dirven & Verspoor 2004),

that is, how often these lexemes are used as causal verbs. For the first three *shi*, *ling* and *rang*, causative meaning is their major choice whereas causal use is not the best guess when it comes to *jiao*, *gei* or *yao*. In the dataset, another salience also needs our attention – onomasiological salience (Geeraerts 1993, Geeraerts 2010). It is based on Table 2.2.3, and simply points to the percentage of the observations with different causatives, in other words, how the pie of causatives is divided among them. Therefore we answer the question with a pie chart, Figure 3.1.

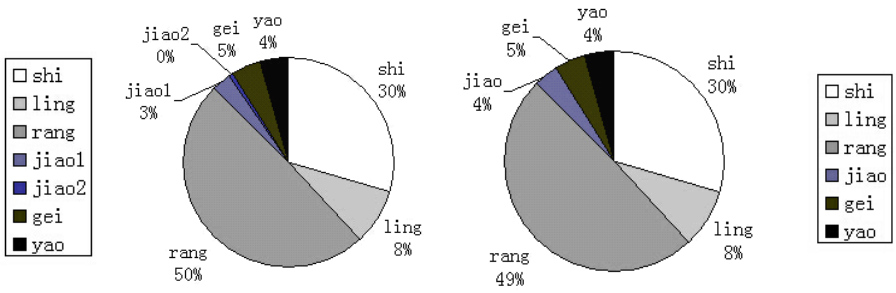

Figure 3.1 Percentage of the observations with the causatives

The pie on the right hand side reflects the percentage after conflation of *jiao1* and *jiao2* on the left hand side, from which we could see that the observations with *jiao2* take up close to 0% of the data. Figure 3.1 shows that *rang* and *shi* are the onomasiologically favored when one wants to express causation with an analytic construction, and *ling*, *gei*, *jiao* and *yao* are chosen after.

3.2 Exploratory analysis

In order to explore the data and find the underlying pattern, we use multiple correspondence analysis, of the initial data matrix with the variables related to the (in)direct causation hypothesis. We first give a brief description of this technique in the following section.

3.2.1 Multiple correspondence analysis

Multiple correspondence analysis (MCA) is an extension of simple correspondence analysis (CA), a multivariate statistical technique proposed by Hirschfeld (1935) and later developed by Jean-Paul Benzécri (1973). It is used in many fields, including social sciences, to represent a set of categorical data as points in low-dimensional geometric space (Hoffman & Leeuw 1992).

MCA is performed by applying the CA algorithm to either an indicator matrix or a Burt table formed from these variables (Greenacre 2007). Associations between variables are uncovered by calculating the chi-square distance between different categories of the variables and between the individuals (or respondents). These associations are then represented graphically as "maps", which eases the interpretation of the structures in the data. Oppositions between rows and columns are then maximized, in order to uncover the underlying dimensions best able to describe the central oppositions in the data. The first axis is the most important dimension, the second axis the second most important, and so on, in terms of the amount of variance accounted for. The number of axes to be retained for analysis is determined by calculating modified eigenvalues.

There are numerous softwares of data analysis offering MCA. The statistical system R is probably the richest free software in this field (Husson & Pagès 2009). Thus there are a number of usage-based studies in linguistics using it (Glynn & Fischer 2010, Glynn 2012a, 2012b, Tummers, Speelman & Geeraerts 2012, etc.). R includes the pacakages {MASS}, {ca} (Nenadic & Greenacre 2007), {languageR}, {anacor}, {homals}, {FactoMineR}, {vegan}, {ade4} and {pamctdp}, and {ExPosition} performing the analysis. The first four are dilated upon in Glynn's chapter (2012b in Glynn & Robinson 2012). In this study we turn to the package {FactoMineR}.

The purpose is to visualize which values of the eight variables (predictive features associated with directness or indirectness) and which response category (specific causatives) tend to co-occur, in other words, to find the patterns that the causatives tend to go with so that their position along the continuum of (in)direct causation can roughly be located, and a sideshow is to learn how similar or dissimilar those causatives are.

3.2.2 MCA solution and interpretation

The MCA solution is displayed in Figure 3.2.2.1. The points are the observations in our dataset. The closer they are to one another, the more features they share.

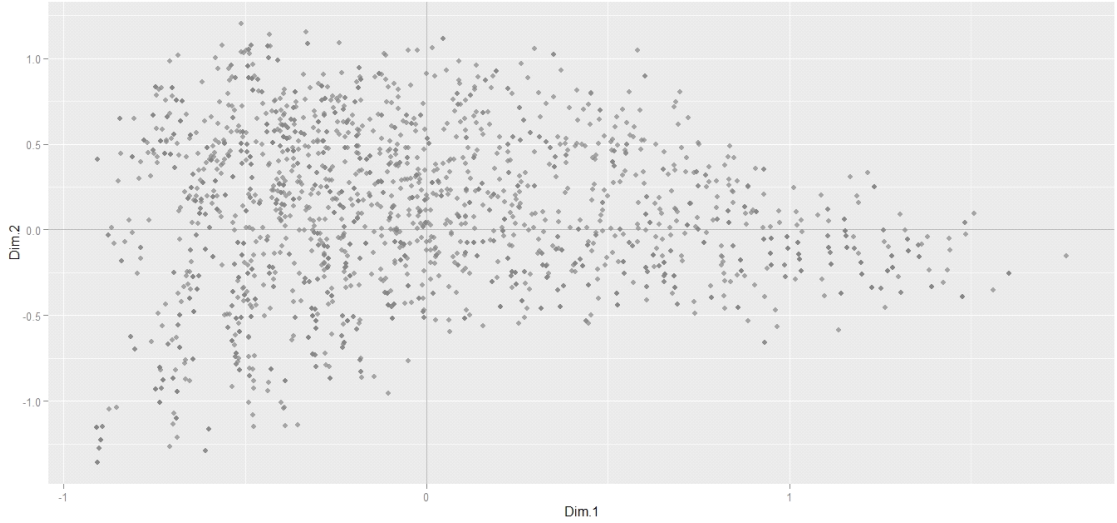

Figure 3.2.2.1 MDS solution of the dataset

Figure 3.2.2.2 maps the variables' levels and labels on the MCA solution. With its help we discern the data points.

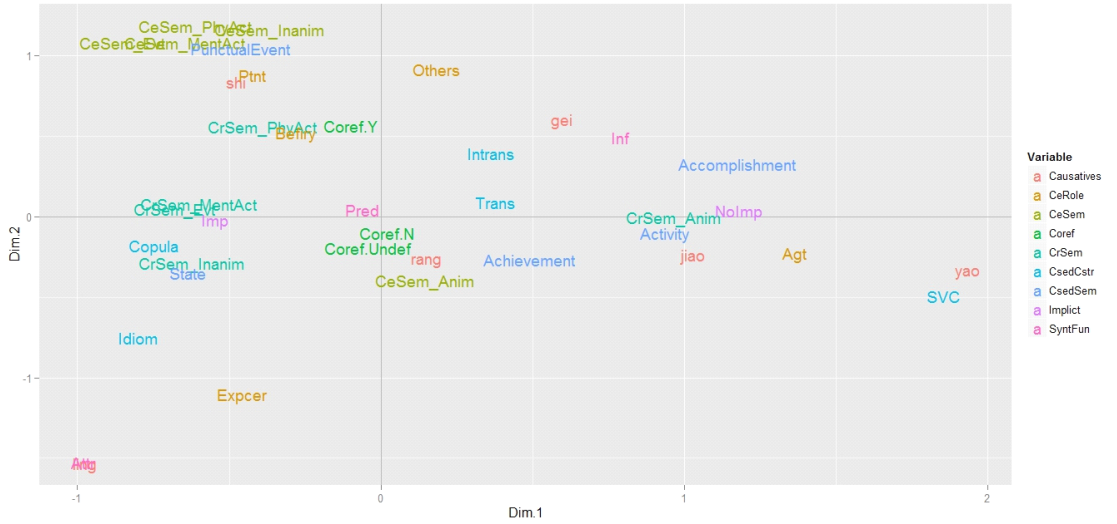

Figure 3.2.2.2 Distribution of the variables in the 2D data space

The major amount of variance is explained by the first dimension. The left side to the centre at the intersection of 0 and 0 is dominated by the levels of the variables predicted to be associated with direct causation. The right hand side has more indirect features down to the right of the continuum in Table 2.3.3. Then the causatives are divided into two groups. *Shi* and *ling* are located at the part of direct causation, and the other four are with the indirect part. Between *shi* and *ling*, the difference lies in that *ling* takes the values at the leftmost end of Table 2.3.3 for most of the

variables except CeSem and Coref, which puts *ling* at the higher scale of direct causation than *shi*. As for the other four causatives, *rang* is a relatively neutral one, which bridges direct and indirect causation, since it is close to the centre. *Jiao* and *yao* share the features farthest to the right of the (in)direct causation continuum, which suggests they are more often with indirect causation than *gei*. The interpretation of the MCA map draws a sketchy conclusion, as shown in Figure 3.2.2.3. We use square brackets to indicate the first layer of distinction – direct vs. indirect causation, and round brackets to suggest the lexemes within are similar to each other or that they share a grey zone that makes it difficult to distinguish between them only based on the MCA plot.

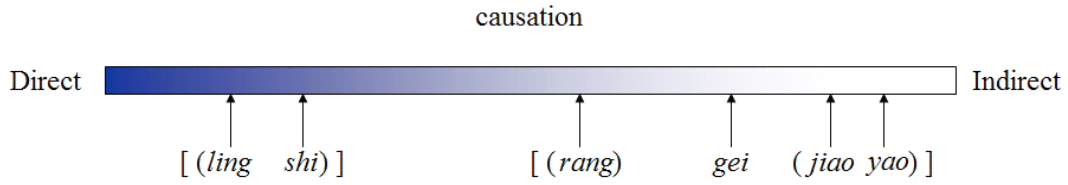

Figure 3.2.2.3 Predicted positions along (in)direct causation continuum based on MCA

We claim that the positions of causatives along the continuum of (in)direct causation are only roughly located because of the number of variables, which makes the plot only a rough guide and one must return to the data to check every correlation visualized (cf. Glynn 2012b). To do that, in a more efficient manner, we resort to logistic regression analysis as a confirmatory test.

3.3 Confirmatory test

In this section, we briefly introduce regression analysis, specifically the subtype suited to our data – multinomial logistic regression analysis. Then the output will be presented and interpreted so that Figure 3.2.2.3 can be tested according to the results. And we try to answer the question whether the (in)direct causation hypothesis is important for Chinese analytic causatives, with the aid of this confirmatory technique as well.

3.3.1 Multinomial logistic regression analysis

Regression analysis (cf. Lindley 1987, Fox 1997, Draper 1998, Sen & Srivastava 2011, etc.) is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand which among the independent variables are related to the dependent variable, and the forms of these relationships, i.e. how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. It finds its way into various fields of application.

And in (corpus) linguistics, logistic regression analysis is the most widely used technique (cf. Speelman 2014). It is so extensively applied that google scholar could provide about 6,660 results if one searches *corpus linguistics* with the phrase *logistic regression*, about 21,800 results if one searches *linguistics* with *logistic regression*. Logistic regression (cf. Balakrishnan 1991, Agresti 2002, Hilbe 2009) is used for predicting the outcome of a categorical dependent variable based on one or more predictor variables. The term often refers specifically to the situation in which the dependent variable is binary, that is, the number of its available categories is two (for example, *doen* vs. *laten* in Speelman & Geeraerts 2009, and Levshina 2011).

In fact, logistic regression can be binomial/binary or multinomial. Binomial logistic regression deals with situations in which the observed outcome for a dependent variable can have only two possible types, as has been mentioned above. Multinomial logistic regression, on the

other hand, deals with situations where the outcome can have three or more possible discrete types (the case of *shi* vs. *ling* vs. *rang* vs. *jiao* vs. *gei* vs. *yao* in our study).

In statistics, multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems (Greene 1993). It builds a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables, which may be real-valued, binary-valued, categorical-valued, etc. The dependent variable in question, which usually comes from a limited set of items, is equivalently categorical or nominal, meaning that it falls into any one of a set of categories which cannot be ordered in any meaningful way. It is agreed that the best values of the predictors for a given problem are usually determined from some training data, e.g. some examples of known words being spoken. Therefore multinomial logistic regression analysis is the type particularly suited for our research.

3.3.2 Regression output and reading

We employ the multinom function from the {nnet} package to estimate the multinomial logistic regression model. Before running it, we choose the reference level for the variables. The baseline values are CrSem=Anim, CeSem=Anim, CeRole=Agt, Coref=N, CsedCstr=Copula, CsedSem=State, Implict=Imp, SyntFun=Inf, and Causatives=rang. In the later analysis, we play with the reference level of Causatives to obtain a more straightforward view of the comparison between a pair of causatives. With this releveling, our multinomial logistic regression now models the odds that the observation occurs with one of the other levels rather than the reference.

We first ask for Anova of our model (Table 3.3.2.1), which reveals that most of the predictors are significant, that is to say, they play a role in telling apart the causal verbs. But there is one exception, the variable Coref, which has no effect on the dependent variable. Figure 3.3.2.1 plots the chi squares so as to get an idea of the importance of these predictors.

Analysis of Deviance Table (Type II tests)				
Response: Causatives				
	LR Chisq	Df	Pr(>Chisq)	
CrSem	127.628	20	<2.2e-16	***
CeSem	112.26	20	7.590e-15	***
CeRole	203.813	20	<2.2e-16	***
Coref	15.849	10	0.104	
CsedCstr	98.052	20	2.806e-12	***
CsedSem	86.914	20	2.555e-10	***
Implict	47.592	5	4.302e-09	***
SyntFun	66.08	10	2.517e-10	***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Table 3.3.2.1 Anova of the model with predictors only related to (in)direct causation

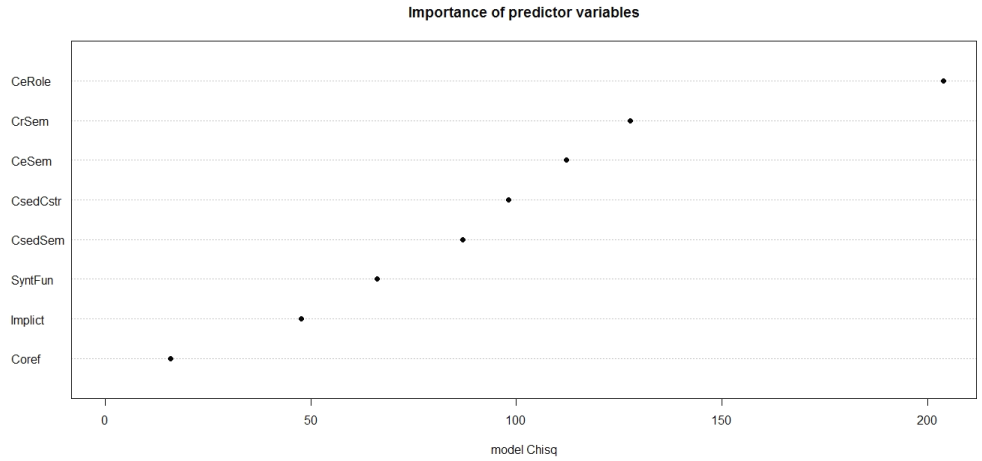

Figure 3.3.2.1 Importance of the predictors derived from (in)direct causation

The figure shows that thematic role of causee is the predictor with the greatest explanatory power. The data proportion explained by the three variables explored in Speelman and Geeraerts's study (2009) almost amounts to that explained by CeRole itself in the Chinese case. Then we ask for a summary of the fitted model and the 95% confidence intervals, going into the specific levels to detect whether a value increases or decreases the probability of having one causative over the other baseline verb. It is then translated into which direction along the (in)direct causation continuum the value points the causatives to. We will discuss in detail the chunk of *shi*'s confidence intervals presented in Table 3.3.2.2. The chunks for the other four levels in this model are given in Appendix 3, and they can be interpreted in the similar way.

„ gei…			
„ jiao…			
„ ling…			
„ shi			
		2.5%	97.5%
	(Intercept)	-1.91449607	-0.84198046
	CrSemEvt	0.56434589	1.24243571
	CrSemInanim	0.29253997	0.92496092
	CrSemMentAct	0.78756895	1.80624952
	CrSemPhyAct	0.85615797	1.5385308
	CeSemEvt	-0.02085188	2.38185537
	CeSemInanim	0.75313888	1.25287471
	CeSemMentAct	-0.43675596	0.7183657
	CeSemPhyAct	0.6232007	1.82825603
	CeRoleBefiry	-0.13334041	0.58978202
	CeRoleExpcer	-0.81291473	-0.05475879
	CeRoleOthers	-0.06797623	2.83600737
	CeRolePtnt	0.51896541	1.22391775
	CorefUndef	-0.82650607	1.05610290
	CorefY	0.04380974	0.54665155
	CsedCstrIdiom	-0.69964618	0.10393146
	CsedCstrIntrans	-0.4899251	0.12495003

	CsedCstrSVC	-1.64940654	0.30007776
	CsedCstrTrans	-0.51194123	0.0103317
	CsedSemAccomplishment	-0.6386589	0.25823966
	CsedSemAchievement	-0.15712192	0.39727756
	CsedSemActivity	-0.21149056	0.62691389
	CsedSemPunctualEvent	-0.07203914	0.49879517
	ImplictNoImp	-0.62610945	0.00196395
	SyntFunAttr	-1.04794149	0.06930546
	SyntFunPred	-0.49751318	0.1178241
,, yao…			

Table 3.3.2.2 Confidence intervals for *shi* of the (in)direct causation related model

We have the upper and lower bound of the 95% confidence interval for each of the predictor estimates. The variable Coref has been ruled out by Anova. For the other predictors, the confidence interval only matters when there is no 0 in between. A predictor is not significant unless the numbers in both 2.5% column and 97.5% column are positive or negative at the same time. Positive estimate indicates that the value in question increases the probability of having *shi* rather than *rang*. Negative estimate indicates that it may favor the reference *rang* and disfavor *shi*.

In the chunk of Table 3.3.2.2, eight values from three variables are significant. The variable CrSem states that *shi* is favored when the causer is inanimate entities, events, mental or physical activities, and *rang* has more animate causers. Likewise, inanimates and physical activities as the causee tend to co-occur with *shi* construction, compared to *rang* construction. More causees of *shi* take the thematic role of patient whereas *rang*'s causee turns out to be more of an agent. These estimates so far signify the same thing that *shi* is used to express direct causation, and *rang* indirect causation, as Ni (2012) concludes. However, another working parameter here is thematic role experiencer. It decreases the probability of the occurrence of *shi*, which functions as pushing force to drive *shi* to indirect causation. But this force is not as strong as the joint one of the other factors, which is implied by the fact that the absolute value of the lower boundary of CeRoleExpcer's confidence interval (0.81291473) is less than the sum of the other seven levels' lower bound values (4.395918). As a result, *shi* is still associated with direct causation while *rang* is with indirect causation. Their positions stay unchanged as in Figure 3.2.2.3.

We examined the intervals closely and did a comparable analysis after releveling the response variable. We will not go further into the details about these analogous steps we performed. But the results retune Figure 3.2.2.3 and turn it into a more accurate portray, Figure 3.3.2.2.

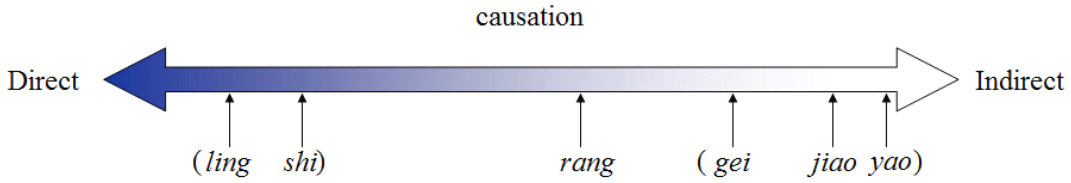

Figure 3.3.2.2 Confirmed positions by multinomial logistic regression

There is no great position shift of the causatives, which shows MCA, as an exploratory technique, is informative and reliable. But notice that the square brackets are dismissed, and that the closed continuum is transformed into an open-ended one, which points to two orientations of directness and indirectness with arrows in Figure 3.3.2.2. This is because the direct or indirect

causation that we associate the causatives with is not in an absolute term, but rather from a comparative point of view. For example, *rang* is used more in a direct causal situation in comparison with *gei*, but it clearly does not mean *rang* construction expresses direct causation. As in Figure 3.2.2.3, *shi* and *ling* are nearer to directness but not distinct from one another. We do not move *ling* to make it paralleled to *shi* but keep it in the original place instead because the estimates of the submodel *shi* vs. *rang* denote that there exists force which makes it complicated to determine the relative (in)directness, yet the predictors of the *ling* vs. *rang* submodel point to the same direction so that there is no element which blurs their distinction. *Rang* is still in the middle but the reason changes. In Figure 3.2.2.3 *rang* is put in a round bracket, which means that *rang* is a neutral causative, different from *gei*, *jiao* and *yao*, though it is slightly weighted towards indirect causation. In Figure 3.3.2.2 now, it stands alone since its central position is not decided by its semantic neutrality but is a result of the comparison. The construction with *rang* is more indirect than that with *shi*/*ling*, and more distinctively direct than that with the other three. The multinomial logistic regression model denotes not only *jiao* and *yao* share a grey boundary but also it is impossible to tell one from another under the distinction of (in)direct causation for all the three *gei*, *jiao* and *yao*. It calls for another hypothesis if we want to differentiate them.

This casts doubt upon the validity of the (in)direct causation hypothesis. We have learnt that it is at work behind the choice of analytic causatives in Chinese. But we would like to know now if it works enough.

3.3.3 Assessing the model of (in)direct causation

In order to evaluate the model of (in)direct causation, we start with the pR2 function in the package {pscl} to get the pseudo R squared measures, as listed in Table 3.3.3.1.

llh	llhNull	G2	McFadden	r2ML	r2CU
-2885.1254872	-3702.4442530	1634.6375315	0.2207511	0.4401409	0.4744143

Table 3.3.3.1 Pseudo R squared measures for the model of (in)direct causation

Low pseudo R squared measures, McFadden (McFadden 1973), r2ML and r2CU, indicate that a model does not explain much variation in the data (cf. Hu, Shao & Palta 2006). The highest one for this multinomial logistic regression model is r2CU, which is not up to 0.5 yet. There is more than 50% variability left out by the (in)direct causation model.

We then check the predictive accuracy of the same model. Compared to the baseline of 1/6 (0.17), 0.5830376 is only what we get. And lastly we use the function somers2 in the {Hmisc} package to have a look at the C measures and Somer's D coefficients (cf. Somers 1962, Göktaş & İşçi 2011). Table 3.3.3.2 lists the measures for each level of Causatives versus the rest.

Causatives	C	Dxy	n	Missing
yao	0.920376	0.840752	2818	0
gei	0.867252	0.734504	2818	0
ling	0.85552	0.71104	2818	0
shi	0.793744	0.587489	2818	0
jiao	0.793648	0.587297	2818	0
rang	0.665788	0.331576	2818	0

Table 3.3.3.2 Some's Ds of the (in)direct causation model (one CAUSE vs. the rest)

The table above has already sorted the causatives in accordance with C measures, which is equal to the probability that one causative is chosen instead of the others, assuming the variable features fit in with the verb. It tells us if the model prediction tallies with the observed. You may

notice that the model is fairly good at predicting *yao*, but as for *rang* or *shi*, the emphasized pair by Ni's study on the effect of the (in)direct causation hypothesis on Chinese analytic causatives, the model is not good enough. A more refined theory, which can model the language above or at least around 70%, is in need than the conceptual difference only (the cut-off set to 70% cf. Klaven 2014).

3.4 Multivariate framework of causation

In order to go ahead towards an overarching theory about the causes for choosing any one of the causatives, we consult Speelman and Geeraerts's (2009) and Levshina's (2011) findings. Speelman and Geeraerts's results show most of the predictions that they derive from the (in)direct causation hypothesis are falsified so they instead suggest a different basic hypothesis: *doen* is an obsolescent form with a tendency towards semantic and lexical specialization. Levshina extends their approach to her comprehensive study, and finds the distinctive exemplars of *doen* and *laten*, re-emphasizes the importance of (in)direct causation dimension, points out another importance distinction between mental and non-mental caused events, and also confirms Speelman and Geeraerts's hypothesis with the findings of lexical effects on the choice and lectal differences between Belgian Dutch and the Netherlandic Dutch, and between different registers (spontaneous face-to-face conversations, online postings or newspaper articles). Their results reveal the architecture of linguistic system, in which structural, grammatical, semantic, discursive and variational factors simultaneously determine the presence or absence of linguistic variables. The question needs to be answered: if Chinese confirm this multivariate conception of the grammar.

Hence we fit another model of multinomial logistic regression, with the variables in the aforementioned (in)direct causation only model and the rest discussed in the previous section 2.3, and summarized in Appendix 1. We keep the baseline values of the variables related to the (in)direct causation hypothesis, and set the other references as CrExp/CeExp=Explicit, CrPers/CePers=3Sg, CrDef/CeDef=Def, CseModality/CsedModality=None, CseNeg/CsedNeg=N, CrCollocSig/CeCollocSig/CsedCollocSig=FALSE, CrIntent=Unintent, Manner=N, Structure=Single. We will first evaluate the multivariate model this time. Its pseudo R squared (a), C along with Somer's D measures (b) are given in Table 3.4.

llh	-2056.87
llhNull	-3702.44
G2	3291.151
McFadden	0.444457
r2ML	0.688982
r2CU	0.742632

Causatives	C	Dxy	n	Missing
gei	0.993599	0.987198	2818	0
yao	0.98673	0.97346	2818	0
jiao	0.916606	0.833212	2818	0
ling	0.894012	0.788025	2818	0
shi	0.854976	0.709952	2818	0
rang	0.809334	0.618669	2818	0

Table 3.4 Pseudo R^2 (on the left, a) & Somer's D (on the right, b) of the multivariate model

The measures of the new model are much better than the previous one. Its explanatory power is strengthened that around 74% of the variation is captured by the model. As for each specific causative, the model's performance is improved too. *Gei* and *yao* are still up as the top two, and *jiao* is ranked the third, which interestingly suggests the multivariate model is unexpectedly excellent at solving the cloud which cannot be untangled by the (in)direct causation hypothesis. *Rang* is the most improved one, from 0.665788 to 0.809334, although there is still about 20% variability not covered by the model. *Ling* and *shi* have the relatively small improvement (0.038492 and 0.061232 respectively). They are the two associated more with direct causation but

still can not be told apart by it. Our multivariate model does not go much deeper into their discrepancy. Although the pair of *shi* and *rang* remains at the bottom in the Somer's D table, the multivariate model of causation has the prediction accuracy of 0.7044003, which shows it is a robust model for explaining Chinese analytic causatives. The Chinese case does confirm the multivariate model of linguistic construal of causality.

This paper will not give detailed account of the confidence intervals of the new proposed model but leave it till later because up to now our study has already achieved the goal of hypothesis testing. However, we will still take a moment, looking at the importance of predictors in the new model, plotted in Figure 3.4 with the assistance of the model's Anova. The most important one is the semantic factor Manner. Then the logical and grammatical factors are ranked so high that they take up five positions out of the top seven. The logical factors are a cue for lexical preference at play. The grammatical factors, CrPers/CePers, are an implication that social identity may be one of the hidden triggers, which leads to a follow-up analysis with language-external variables. But for the time being, we draw our conclusion of the present study.

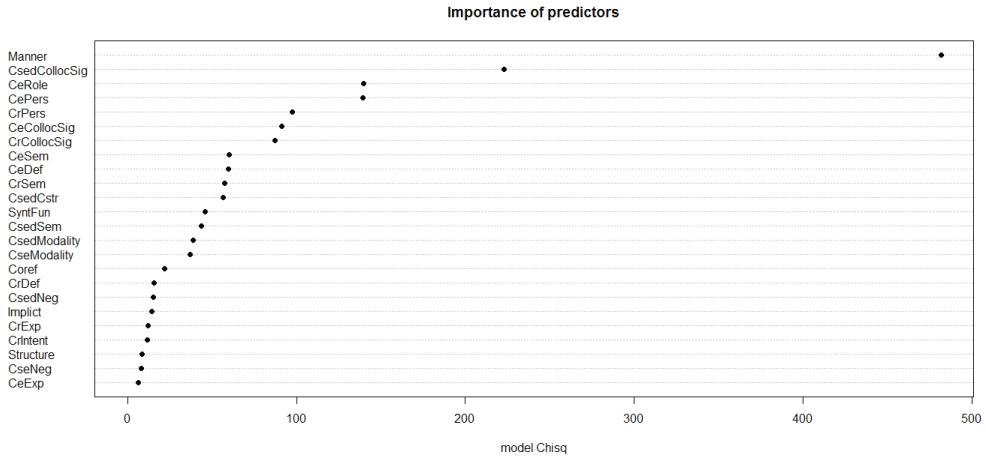

Figure 3.4 Importance of the predictors in the multivariate model of causation

IV. Conclusion

Starting from a set of 2,818 cases of seven analytic causatives extracted from UCLA2, we explored the underlying pattern with multiple correspondence analysis, and performed multinomial logistic regression analysis to model the data, first incorporating a series of factors which on the basis of the (in)direct causation hypothesis were predicted to affect the choice between the target causatives (as Model 1: direct/indirect causation only model), and then bringing in more structural/grammatical and conceptual/semantic factors (as Model 2: multivariate model of causation). The results show that the (in)direct causation hypothesis can roughly classify the causatives into three groups but it is not qualified as the major distinction for partitioning Chinese analytic causatives. A more comprehensive framework, rather than the conceptual difference only, needs taking into account to understand the construction and its lexical realization, which should include language habit like lexical idiomatization and social identity perception or understanding and so on and so forth.

To make a fair conclusion, we attempt to find the rationale behind the results of the models.

5.1 Negative story and likely cause

The (in)direct causation only model is tested to be of limited success, which proves the conceptual distinction is not that important for analytic causatives in Chinese, and even ineffective if the

choice is between *gei*, *jiao* or *yao*, for example. This negative result may originate not just in the distinction itself but in the big picture of language as well.

Wolff (1996, 2003) investigates direct vs. indirect causation with a series of psycholinguistic experiments. The task for one of them is for 15 undergraduates to describe 12 animations (6 direct causal events, 6 indirect events). The results, as in Figure 5.1.1, show that lexical causatives are usually the selected expression when one describes direct causal scene, and that analytic causatives (periphrastics) per se are highly correlated with indirect causation (cf. Wolff In press). Given this division of labor between lexical and analytic causatives, the previous Figure 3.3.2.2 does not render the whole picture of the (in)direct causation continuum. Figure 5.1.2 is more in conformity with Wolff's findings.

The continuum is just an attempt to tackle the issue of linguistic presentation of causation. It is far from complete because there exist in language varieties of ways of describing causation. Besides analytic and lexical causatives, our devices include resultatives, prepositions like *from*, subordinating conjunctions like *because*, coordinating conjunctions like *so*, conjunctive adverbs like *therefore*, lexical cue phrase like *that's why*, and morphological causatives for some languages, etc. Their positions along the continuum could also be an intriguing topic for discussion in further research. But for now, when we zoom in on the space of analytic causatives, a likely cause of the frustration the (in)direct causation hypothesis confronts is that they belong to a category that language users choose within to express indirect causation already. The impact of this distinction within the category is much lessened. The choice between them made by language users tends to be triggered by other reasons.

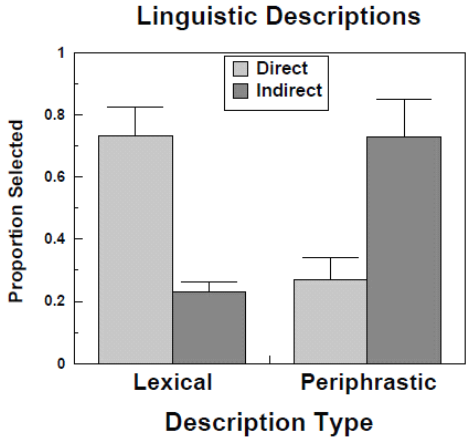

Figure 5.1.1 Results of the description task (Wolff In press)

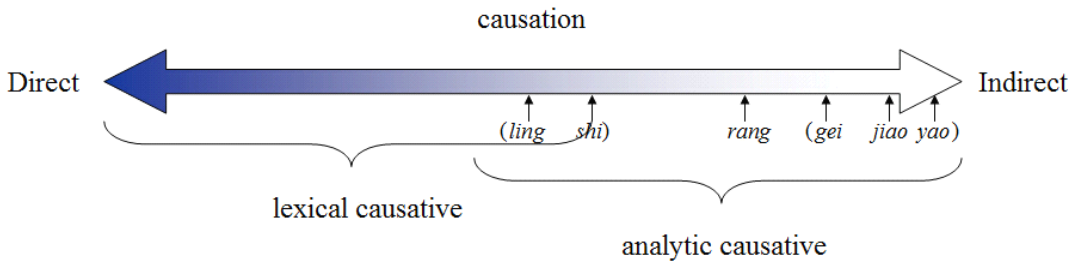

Figure 5.1.2 (In)direct causation continuum from a wider view

5.2 Positive conjecture and further perspectives

Then what are the other reasons? We also take one step towards the explanation with the multivariate model of causation. The model agrees with Geeraerts's multifactorial framework of

the grammar, that is, structural, grammatical, semantic, discursive and lectal factors simultaneously determine the configuration of the final linguistic output.

This conjecture meanwhile accords with Wolff's idea of plurality (In press). He claims that causal expression or causal thought is never singular but plural. There is no one overarching theory that covers it all. Some phenomena are problematic for specific Theory A, but not for Theory B. Other phenomena are problematic for Theory B, but not for Theory A. Thus he proposes causal pluralism for the purpose of understanding causation. Though his hypothesis focuses on structural pluralism, it gives a hint as to what the architecture of language should be.

Our multivariate model of causation is not exhaustive nevertheless. So far it cannot do better for *shi* and *rang*, which have causal salience in both the semasiological and onomasiological terms, than for the other lexemes. However, it is understandable that the two may vary to a larger extent that makes it more difficult to capture their intracategorical variation, since they are used extensively. Given this, the model is kept open source in the sense that more language-internal and -external variables can have a shot, to bring to light whatever is hidden underneath the linguistic semblance.

5.3 Limitations

Before we bring the paper to the end, we will have a few words about the limitations of the current study.

First, there might be a problem concerning the treatment of conflating *jiao1* and *jiao2*. The two causatives are similar to a certain degree from the conceptual point of view. But they may differ due to some unexpected motive, which we overlooked in this study. The reason why we still merge them is that we do not expect much divergence between the two in modern written Mandarin Chinese. But as the investigation continues in the future, they are supposed to be kept apart when dimensions like modes of language (written vs. spoken), diachronic strata (modern vs. archaic) and lectal varieties (standard vs. dialect) are bought in.

Second, the two models we fitted and compared in the study deal with main effects only for the sake of simplicity. But as Speelman and Geeraerts's study (2009) reveals, such models may unjustly oversimplifies the patterns in the data. To further scrutinize the data, we need to take the interactions into consideration.

Third, our models may be susceptible to overfitting. They may infringe the rule of thumb about the maximum number of predictors in one statistical model given the size of the data. Despite this, they are explicit enough to help us embark on the right track and move in the right direction. And this problem will be solved by applying some dimensionality reduction technique or simply enlarging the dataset.

5.4 A closing remark about methodology

The study we presented in this paper is a case study in Chinese of using quantification techniques for hypothesis testing. It is not definitive, but hopefully it will inspire Chinese linguists to draw more support from statistics or computer science to approach our subjects more objectively, systematically and scientifically.

Notes

1. The authors thank the members of the QLVL research unit at KU Leuven for valuable suggestions concerning the study.
2. For those who are interested in the dataset we use and other details, please contact the authors

at: yanan.hu@student.kuleuven.be

3. All remaining errors are solely the authors'.

Appendix 1: Independent variables

No	Variables	Values	Notes
Causer			
1	CrExp: Explicitness of causer	-Explicit	expressed
		-Implicit	not expressed
2	CrSem: Semantic class of causer	-Anim	human, organization, animal, body part
		-Inanim	physical materials, mechanism, abstract entity
		-Evt	event
		-PhyAct	physical activity
		-MentAct	mental activity
3	CrPers: Grammatical person of causer	-1Sg	first person singular
		-1Pl	first person plural
		-2Sg	second person singular
		-2Pl	second person plural
		-3Sg	third person singular
		-3Pl	third person plural
		-Undef	undefined
4	CrDef: Definiteness of causer	-Def	definite
		-Indef	indefinite
5	CrIntent: Intention of causer	-Intent	intentional
		-Unintent	unintentional
		-Undef	undefined
6	CrCollocSig: Collocational significance between CAUSE and causer	-TRUE	significant
		-FALSE	not significant
Causee			
7	CeExp: Explicitness of causee	-Explicit	expressed
		-Implicit	not expressed
8	CeSem: Semantic class of causee	-Anim	human, organization, animal, body part
		-Inanim	physical materials, mechanism, abstract entity
		-Evt	event
		-PhyAct	physical activity
		-MentAct	mental activity

9	CePers: Grammatical person of causee	-1Sg	first person singular
		-1Pl	first person plural
		-2Sg	second person singular
		-2Pl	second person plural
		-3Sg	third person singular
		-3Pl	third person plural
		-Undef	undefined
10	CeDef: Definiteness of causee	-Def	definite
		-Indef	indefinite
11	CeRole: Thematic role of causee	-Agt	agent
		-Ptnt	patient
		-Expcer	experiencer
		-Befiry	beneficiary
		-Others	other role, e.g. location
12	CeCollocSig: Collocational significance between CAUSE and causee	-TRUE	significant
		-FALSE	not significant
Relationship between Causer and Causee			
13	Coref: Coreference of causer and causee	-N	no
		-Y	yes
		-Undef	undefined
Causing Event			
14	Manner: Manner of CAUSE	-N	no, pure causative
		-Y	yes
15	CseModality: Type of modal verbs in front of CAUSE	-None	no modal verb
		-Possibility	can, could, will, would, may, might
		-Necessity	should, must, ought to, need
		-Inclination	shall
		-Evaluation	
16	CseNeg: Negation in front of CAUSE	-N	no negation
		-Y	negation
Caused Event			
17	CsedCstr: Grammatical construction of effected predicate	-Trans	transitive verb
		-Intrans	intransitive verb
		-Copula	copula
		-Idiom	idiom
		-SVC	serial verb construction
18	CsedSem: Semantic class of effected predicate	-State	emotion
		-Activity	
		-Accomplishment	

		-Achievement	perception
		-PunctualEvent	
19	CsedModality: Type of modal verbs in front of effected predicate	-None	no modal verb
		-Possibility	can, could, will, would, may, might
		-Necessity	should, must, ought to, need
		-Inclination	shall
		-Evaluation	
20	CsedNeg: Negation in front of effected predicate	-N	no negation
		-Y	negation
21	CsedCollocSig: Collocational significance between CAUSE and effected predicate	-TRUE	significant
		-FALSE	not significant
Relationship between causing event and caused event			
22	Implict: Implicativity	-Imp	Causing event entails caused event
		-NoImp	possible to add a counterfactual coordinate clause
Features of causative construction			
23	SyntFun: Syntactic function of causative construction in the whole sentence	-Pred	main predicate
		-Inf	infinitive as adverbial clause of purpose
		-Attr	attributive clause in front of noun
24	Structure: The number of caused events that CAUSE takes	-Single	one effected predicate
		-Multiple	several effected predicates

Appendix 2: Chinese “modal verbs” and the English translation

-Possibility:能 can/could,能够 can,会 will/would,可 may,可能 may/might,可以 can,得以 can;
-Necessity:应 should,应该 must,应当 ought to,得(dei) should,该 should,当 should,须得 need,犯得着 deserve,犯不着 do not deserve,理当 should;
-Inclination:愿意 would (like),乐意 would (like),情愿 would rather,肯 would,要 shall,愿 be willing to,想要 would like,要想 want,敢 dare,敢于 dare,乐于 be willing to;
-Evaluation:值得 be worth,便于 be easy to,难于 be difficult to,难以 be difficult to,易于 be easy to.

Appendix 3: Confidence intervals of the (in)direct causation only model

Ref=rang, CrSemAnim, CeSemAnim, CeRoleAgt, CorefN, CsedCstrCopula, CsedSemState, ImplictImp, SyntFunInf			
,, gei			
		2.5%	97.5%
	(Intercept)	-10.4775392	-5.6611075
	CrSemEvt	-0.6895739	0.6308054
	CrSemInanim	-1.4961785	-0.136115
	CrSemMentAct	-2.2830051	0.3852996
	CrSemPhyAct	-1.9081535	-0.2371204
	CeSemEvt	-0.4463941	3.5845133
	CeSemInanim	-0.1018505	0.9414731
	CeSemMentAct	-1.5080279	1.5870867
	CeSemPhyAct	-2.1982105	2.0062776
	CeRoleBefiry	0.895108	2.0539436
	CeRoleExpcer	-0.5386636	0.7709246
	CeRoleOthers	1.472818	4.8302662
	CeRolePtnt	-0.3301313	1.0619648
	CorefUndef	-11.9701284	-11.9701033
	CorefY	-0.4745116	0.5622793
	CsedCstrIdiom	-1.4518818	4.165053
	CsedCstrIntrans	0.7886408	4.8362442
	CsedCstrSVC	-159.6498488	146.1342132
	CsedCstrTrans	1.5378527	5.516731
	CsedSemAccomplishment	0.7458001	3.4068475
	CsedSemAchievement	1.4279108	3.8311381
	CsedSemActivity	0.8909632	3.5255443
	CsedSemPunctualEvent	2.0002562	4.4303415
	ImplictNoImp	-0.3020889	0.8633887
	SyntFunAttr	-3.4849605	0.6948984
	SyntFunPred	-0.6788282	0.3465838
,, jiao			
		2.5%	97.5%
	(Intercept)	-3.3146438	-1.1003327
	CrSemEvt	-0.4476336	1.0346739
	CrSemInanim	-2.3899249	-0.4402826
	CrSemMentAct	-0.8867091	1.5111317
	CrSemPhyAct	-1.9207156	0.2771167
	CeSemEvt	-0.9372454	3.606013
	CeSemInanim	-2.5109897	-0.3970405
	CeSemMentAct	-12.2331568	-12.2330817
	CeSemPhyAct	-0.7120964	1.9000297
	CeRoleBefiry	-0.8438195	0.5285356

	CeRoleExpcer	-2.1194099	-0.6336282
	CeRoleOthers	-280.8191718	264.5665947
	CeRolePtnt	-0.6385701	0.7867689
	CorefUndef	-12.0954353	-12.0953675
	CorefY	-0.8086109	0.4501222
	CsedCstrIdiom	-0.7560186	1.2066965
	CsedCstrIntrans	-0.9039681	0.6235931
	CsedCstrSVC	-0.2713719	1.7726644
	CsedCstrTrans	-0.8976605	0.5035037
	CsedSemAccomplishment	-0.9941268	0.6829589
	CsedSemAchievement	-0.3051945	0.9546079
	CsedSemActivity	-0.7262697	0.8422944
	CsedSemPunctualEvent	-1.5306504	0.308059
	ImplictNoImp	-0.2142276	1.0978744
	SyntFunAttr	-0.811824	1.5893231
	SyntFunPred	-0.4756243	0.6666594
,, ling			
		2.5%	97.5%
	(Intercept)	-6.64148452	-2.3039168
	CrSemEvt	0.52082689	1.68411681
	CrSemInanim	0.13269792	1.24752845
	CrSemMentAct	0.15548471	1.86212789
	CrSemPhyAct	-0.16124813	1.28728927
	CeSemEvt	-464.6161444	444.7185252
	CeSemInanim	-1.09600483	0.38455601
	CeSemMentAct	-0.23427175	1.74970537
	CeSemPhyAct	-2.5717912	1.55770123
	CeRoleBefiry	-1.24771863	0.3110035
	CeRoleExpcer	-0.11032766	1.21328769
	CeRoleOthers	-344.1971151	326.4566275
	CeRolePtnt	-1.05484376	0.45153735
	CorefUndef	-1.38144434	1.69868674
	CorefY	-0.97010809	0.20533279
	CsedCstrIdiom	-0.03130047	0.87160445
	CsedCstrIntrans	-0.33116196	0.7133289
	CsedCstrSVC	-1.20591821	1.97371704
	CsedCstrTrans	-1.02339031	-0.06240997
	CsedSemAccomplishment	-0.97707387	0.8815133
	CsedSemAchievement	-0.89469809	0.01396623
	CsedSemActivity	-1.34603048	0.20671615
	CsedSemPunctualEvent	-1.046219	0.12460687
	ImplictNoImp	-1.83758842	-0.2207495
	SyntFunAttr	1.44858951	5.51502414

	SyntFunPred	0.24538954	4.24957721
,, shi			
		2.5%	97.5%
	(Intercept)	-1.91449607	-0.84198046
	CrSemEvt	0.56434589	1.24243571
	CrSemInanim	0.29253997	0.92496092
	CrSemMentAct	0.78756895	1.80624952
	CrSemPhyAct	0.85615797	1.5385308
	CeSemEvt	-0.02085188	2.38185537
	CeSemInanim	0.75313888	1.25287471
	CeSemMentAct	-0.43675596	0.7183657
	CeSemPhyAct	0.6232007	1.82825603
	CeRoleBefiry	-0.13334041	0.58978202
	CeRoleExpcer	-0.81291473	-0.05475879
	CeRoleOthers	-0.06797623	2.83600737
	CeRolePtnt	0.51896541	1.22391775
	CorefUndef	-0.82650607	1.05610290
	CorefY	0.04380974	0.54665155
	CsedCstrIdiom	-0.69964618	0.10393146
	CsedCstrIntrans	-0.4899251	0.12495003
	CsedCstrSVC	-1.64940654	0.30007776
	CsedCstrTrans	-0.51194123	0.0103317
	CsedSemAccomplishment	-0.6386589	0.25823966
	CsedSemAchievement	-0.15712192	0.39727756
	CsedSemActivity	-0.21149056	0.62691389
	CsedSemPunctualEvent	-0.07203914	0.49879517
	ImplictNoImp	-0.62610945	0.00196395
	SyntFunAttr	-1.04794149	0.06930546
	SyntFunPred	-0.49751318	0.1178241
,, yao			
		2.5%	97.5%
	(Intercept)	-5.91429047	-2.61979759
	CrSemEvt	-1.36146401	1.53057384
	CrSemInanim	-3.7281054	0.48678121
	CrSemMentAct	-10.82696509	-10.82678933
	CrSemPhyAct	-17.58006217	-17.58006168
	CeSemEvt	-83.2853393	75.83749166
	CeSemInanim	-0.60758929	1.10391015
	CeSemMentAct	-11.04466909	-11.04457183
	CeSemPhyAct	-0.83646794	3.58337131
	CeRoleBefiry	-2.51351117	-0.25625086
	CeRoleExpcer	-1.88644555	-0.45317638
	CeRoleOthers	0.18993284	4.53021927

	CeRolePtnt	-1.98153312	-0.05430621
	CorefUndef	-12.27272820	-12.27270567
	CorefY	-1.09633471	0.4463277
	CsedCstrIdiom	-14.40181612	-14.40180411
	CsedCstrIntrans	-0.79219417	1.14283937
	CsedCstrSVC	-0.09827794	2.12521328
	CsedCstrTrans	-0.86847374	1.00707592
	CsedSemAccomplishment	-1.38492278	0.27380962
	CsedSemAchievement	-0.36963131	1.04883393
	CsedSemActivity	-1.19352944	0.47142694
	CsedSemPunctualEvent	-2.04800843	0.65690426
	ImplictNoImp	1.48796723	4.05005757
	SyntFunAttr	-12.60680721	-12.60678149
	SyntFunPred	-0.08895129	0.99956662

References

Agresti, Alan. 2002. *Categorical Data Analysis*. New York: Wiley-Interscience.

Balakrishnan, N. 1991. *Handbook of the Logistic Distribution*. Marcel Dekker, Inc.

Benzécri, J.-P. 1973. *L'Analyse des Données*. Volume II. *L'Analyse des Correspondances*. Paris, France: Dunod.

Comrie, B. 1976. *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems*. Cambridge: Cambridge University Press.

Comrie, B. 1981. *Language Universals and Linguistic Typology*. Chicago: The University of Chicago Press.

Croft, William. 1991. *Syntactic categories and grammatical relations: The cognitive organization of information*. Chicago/London: University of Chicago Press.

Dirven, René & Verspoor, Marjolyn. 2004. *Cognitive Exploration of Language and Linguisitics*. John Benjamins Publishing.

Draper, N.R. & Smith, H. 1998. *Applied Regression Analysis* (3rd ed.). John Wiley.

Fox, J. 1997. *Applied Regression Analysis, Linear Models and Related Methods*. Sage.

Geeraerts, Dirk. 1993. Generalised Onomasiological Salience. In: Nuyts, Jan & Eric Pederson (eds.), *Perspectives on Language and Conceptualization*. 43-56.

Geeraerts, Dirk. 2010. *Theories of lexical semantics*. Oxford: Oxford University Press.

Glynn, Dylan. 2012a. The many uses of run. Corpus-based methods and socio-cognitive semantics. In D. Glynn & J. Robinson (eds.), *Corpus Methods in Cognitive Semantics*. Amsterdam: John Benjamins.

Glynn, Dylan. 2012b. Correspondence analysis. An exploratory technique for identifying usage patterns. In D. Glynn & J. Robinson (eds.), *Corpus Methods in Cognitive Semantics*. Amsterdam: John Benjamins.

Glynn, D. & Fischer, K. 2010. *Corpus-Driven Cognitive Semantics. Quantitaive approaches*. Berlin: Mouton de Gruyter.

Glynn, D. & Robinson, J. 2012. *Corpus methods in Cognitive Semantics. Studies in synonymy and polysemy*. Amsterdam: John Benjamins.

Göktaş, Atila & Öznur İşçi. 2011. A Comparison of the Most Commonly Used Measures of Association for Doubly Ordered Square Contingency Tables via Simulation. *Metodološki zvezki*, Vol. 8, No. 1, 17-37.

Goldberg, A. E. 1995. Constructions: *A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Greenacre, Michael. 2007. *Correspondence Analysis in Practice*, Second Edition. London: Chapman & Hall/CRC.

Greene, William H. 1993. *Econometric Analysis*, fifth edition, 720-723. Prentice Hall.

Gries, Stefan Th. 2012. Corpus linguistics: quantitative methods. In Carol A. Chapelle (ed.), *The encyclopedia of applied linguistics*, 1380-1385. Oxford: Wiley-Blackwell.

Gries, Stefan Th. & Dagmar S. Divjak. 2010. Quantitative approaches in usage-based cognitive semantics: myths, erroneous assumptions, and a proposal. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative methods in cognitive semantics: corpus-driven approaches*, 333-354. Berlin & New York: Mouton de Gruyter.

Gries, Stefan Th. & Anatol Stefanowitsch. 2011. Extending collostructional analysis: a corpus-based perspective on 'alternations'. In Adele E. Goldberg (ed.), *Cognitive Linguistics, Vol. IV*, 217-246. New York: Routledge, Taylor and Francis.

Hilbe, Joseph M. 2009. *Logistic Regression Models*. Chapman & Hall/CRC Press.

Hirschfeld, H.O. 1935. A connection between correlation and contingency. *Proc. Cambridge Philosophical Society* 31, 520–524.

Hoffman, Donna L. & Jan De Leeuw. 1992. Interpreting Multiple Correspondence Analysis as a Multidimensional Scaling Method. *Marketing Letters* 3:3, 259-272. Netherlands: Kluwer Academic Publishers.

Hosmer, David W. & Lemeshow, Stanley. 2000. *Applied Logistic Regression* (2nd ed.). Wiley.

Hsieh, Chia-Ling. 2005. Modal verbs and modal adverbs in Chinese: An investigation into the semantic source. *UST (University System of Taiwan) Working Papers in Linguistics* 1, 31-58.

Hu, Bo, Jun Shao & Mari Palta. 2006. Pseudo-R^2 in Logistic Regression Model. *Statistica Sinica* 16, 847-860.

Huddleston, Rodney. *A Short Overview of English Syntax*. Section 6.5d.

Huddleston, Rodney & Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Husson F., Lê S. & Pagès J. 2009. *Exploratory Multivariate Analysis by Example Using R*. The R Series. London: Chapman & Hall/CRC.

Jackendoff, Ray. 1990. *Semantic Structures*. Massachusetts, Cambridge: MIT Press.

Ji, Yixin (纪漪馨). 1986. 英语情态助动词与汉语能愿动词的比较. *语言教学与研究* 3.

Kemmer, Suzanne & Arie Verhagen. 1994. The grammar of causatives and the conceptual structure of events. *Cognitive Linguistics* 5, 115-156.

Klaven, Jane. 2014. How good is good? Evaluating the performance of probabilistic statistical classification models for predicting constructional choices. Presentation in 5th UK Cognitive Linguistics Conference (UK-CLC5).

Lai, Peng (赖鹏). 2006. A Study on the Causes of Interlingual Transfer Errors in the Acquisition of Chinese Modal Auxiliaries. *语言教学与研究* 5, 67-74.

Levshina, Natalia. 2011. *Doe wat je niet laten kan: A usage-based analysis of Dutch causative constructions*. Leuven: Catholic Univeristy of Leuven dissertation.

Lin, Jimmy. 2004. Event Structure and the Encoding of Arguments: The Syntax of the Mandarin and English Verb Phrase.

Lindley, D.V. 1987. Regression and correlation analysis. *New Palgrave: A Dictionary of Economics* 4, 120–23.

Liu, Wei (刘微). 2007. 试从情态意义角度对比分析英语情态动词与汉语能愿动词. *语文学刊: 高等教育版*.

McFadden, D. 1973. Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.), *Frontiers in Econometrics*, 105-42. New York: Academic Press.

Nenadic, O. & Greenacre, M. 2007. Correspondence analysis in R, with two- and three-dimensional graphics: the ca package. *Journal of Statistical Software*, 20(3).

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

Palmer, F.R. 2001. *Mood and Modality*. Cambridge University Press.

Sen, A. & M. Srivastava. 2011. *Regression Analysis – Theory, Methods, and Applications*. Berlin: Springer-Verlag.

Shibatani, Masayoshi. 1976. The grammar of causative constructions: a conspectus. In Masayoshi Shibatani (ed.), *Syntax and semantics 6: the grammar of causative constructions*, 1-40. New York: Academic Press.

Smith, Carlota. 1991. *The parameter of aspect*. Kluwer.

Somers, R.H. 1962. A new asymmetric measure of association for ordinal variables. *American Sociological Review*, 27, 799-811.

Song, Jae Jung. 1996. *Causatives and causation: a universal-typological perspective*. London: Longman.

Speelman, Dirk. 2014. Logistic regression: A confirmatory technique for comparisons in corpus linguistics. In Glynn D., Robinson J. (eds.), book series: Human Cognitive Processing, vol 43, *Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy*, 487-533. Amsterdam: John Benjamins.

Speelman, Dirk & Dirk Geeraerts. 2009. Causes for causatives: the case of Dutch 'doen' and 'laten'. In Ted Sanders and Eve Sweetser (eds.), *Causal Categories in Discourse and Cognition*, 173-204. Berlin/New York: Mouton de Gruyter.

Stukker, Ninke. 2005. *Causality marking across levels of language structure*. PhD dissertation, University of Utrecht.

Tao, Hongyin & Xiao, Richard. 2012. *The UCLA Chinese Corpus* (2st edition). UCREL, Lancaster.

Terasawa, Jun. 1985. The historical development of the causative use of the verb *make* with an infinitive. *Studia Neophilologica* 57: 133-143.

Tummers, José, Dirk Speelman & Dirk Geeraerts. 2012. Multiple Correspondence Analysis as heuristic tool to unveil confounding variables in corpus linguistics. *Proceedings of the 11th International Conference on Statistical Analysis of Textual Data (JADT 2012)*. 923-936.

Vendler, Z. 1957. Verbs and Times. *The Philosophical Review* 66(2), 143-160.

Verhagen, Arie. 1998. Changes in the use of Dutch *doen* and the nature of semantic knowledge. In Ingrid Tieken-Boon van Ostade, Marijke van der Wal & Arjan van Leuvensteijn (eds.), *DO in English, Dutch and German. History and present-day variation*, 103-119. Amsterdam/Münster: Stichting Neerlandistiek/Nodus Publikationen.

Verhagen, Arie. 2000. Interpreting Usage: Construing the history of Dutch causal verbs. In Michael Barlow & Suzanne Kemmer (eds.), *Usage-Based Models of Language*, 261-286. Stanford, CA: CSLI Publications.

Verhagen, Arie & Suzanne Kemmer. 1992. Interactie en Oorzakelijkheid. *Gramma/TTZ, tijdschrift voor taalkunde*, 1, 1. 1-20.

Verhagen, Arie and Suzanne Kemmer. 1997. Interaction and Causation: Causative Constructions in Modern Standard Dutch. *Journal of Pragmatics* 27, 61-82.

Wolff, Phillip. 1996. What Language Might Tell Us About the Perception of Cause. In Garrison W. Cottrell (ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*.

Wolff, Phillip. 2003. Direct causation in the linguistic coding and individuation of causal events. *Cognition*, 88, 1-48.

Wolff, Phillip. In press. *Ten Lectures on Experimental Cognitive Semantics and the language-thought interface*. Beijing: Foreign Language and Teaching Research Press.

Wolff, Phillip, Grace Song and David Driscoll. 2002. Models of causation and causal verbs. In M. Andronis, C. Ball, H. Elston and S. Neuval (eds.), *Papers from the 37th Meeting of the Chicago Linguistics Society, Main Session*. Vol. 1, 607-622. Chicago: Chicago Linguistics Society.